Opportunities, pitfalls and trade-offs in designing protocols for measuring the neural correlates of speech

Ciaran Cooney a,*, Raffaella Folli b, Damien Coyle a

a Intelligent Systems Research Centre, Ulster University, Derry, UK
b Institute for Research in Social Sciences, Ulster University, Jordanstown, UK

ARTICLE INFO

Keywords:
- Speech decoding
- Brain-computer interfaces
- Imagined speech
- Overt speech
- EEG
- ECoG
- Experimental design
- Protocol
- Speech perception vs production

ABSTRACT

Decoding speech and speech-related processes directly from the human brain has intensified in studies over recent years as such a decoder has the potential to positively impact people with limited communication capacity due to disease or injury. Additionally, it can present entirely new forms of human-computer interaction and human-machine communication in general and facilitate better neuroscientific understanding of speech processes. Here, we synthesise the literature on neural speech decoding pertaining to how speech decoding experiments have been conducted, coalescing around a necessity for thoughtful experimental design aimed at specific research goals, and robust procedures for evaluating speech decoding paradigms. We examine the use of different modalities for presenting stimuli to participants, methods for construction of paradigms including timings and speech rhythms, and possible linguistic considerations. In addition, novel methods for eliciting naturalistic speech and validating imagined speech task performance in experimental settings are presented based on recent research. We also describe the multitude of terms used to instruct participants on how to produce imagined speech during experiments and propose methods for investigating the effect of these terms on imagined speech decoding. We demonstrate that the range of experimental procedures used in neural speech decoding studies can have unintended consequences which can impact upon the efficacy of the knowledge obtained. The review delineates the strengths and weaknesses of present approaches and poses methodological advances which we anticipate will enhance experimental design, and progress toward the optimal design of movement independent direct speech brain-computer interfaces.

1. Introduction

Brain-computer interfaces (BCI) are devices which offer users a channel to interact with a computer, and by extension another person, through brain activity alone (McFarland et al., 1997). Traditional approaches to BCI have focused on decoding brain activity such as modulations associated with motor imagery (MI) (e.g., Wang et al., 2006) or steady state visually-evoked potential (SSVEP) (e.g., Müller-Putz et al., 2005) owing to their well-defined temporal, spatial and spectral attributes (see Ramadan and Vasilakos, (2017) for review). However, there is a growing body of research into the development of a direct-speech BCI (DS-BCI), utilising speech or speech-related processes as a communicative modality (Cooney et al., 2018; Ilijina et al., 2017). Whether the goal is to enable communication for people unable to speak or facilitate covert or direct movement independent communication between humans and machines, one of the aims of DS-BCI research is the development of a system to provide a means of communication through imagined speech, avoiding the requirement for articulation via muscle movement or any audible representation of speech. The use of imagined speech in this scenario offers the possibility of a more naturalistic form of communication (Angrick et al., 2021; Anumanchipalli et al., 2019; Cooney et al., 2021; Makin et al., 2020; Moses et al., 2021, 2019), relying on neural recordings corresponding to units of language rather than brain activity unrelated to speech (e.g., imagining limb movement (Tang et al., 2017; Wang et al., 2006)) or focusing on flickering stimuli (SSVEP) (Müller-Putz et al., 2005; Nakatani et al., 2018). A related goal is to facilitate BCI-based communication through attempted speech (otherwise known as intended speech) (Brumberg et al., 2011; Guenther et al., 2009; Moses et al., 2021). Distinguishing between imagined and attempted speech is important as it is conceivable that attempted speech in patients with vocal tract paralysis may resemble overt speech more closely than imagined speech does. At present there is no direct evidence

* Corresponding author.
E-mail address: cooney-c@ulster.ac.uk (C. Cooney).

https://doi.org/10.1016/j.neubiorev.2022.104783
Received 23 February 2022; Received in revised form 12 July 2022; Accepted 15 July 2022
Available online 27 July 2022
0149-7634/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
against this hypothesis, and it is therefore important to avoid conflating imagined speech with attempted speech when reporting speech decoding. A recent study indicated that overt and "intended speech" may be more closely-related to each other than to imagined speech (Pan et al., 2021), but this is inconclusive as the "intended speech" as defined in this study more closely resembles the "silently mimed" condition in Anumanchipalli et al. (2019) than the attempted speech in Moses et al. (2021). While offering obvious benefits to people with communicative difficulties, high-bandwidth communication of imagined speech between humans and machines can have implications for anyone seeking movement-independent communication or direct interaction between humans and robotics or other artificial intelligence applications (Bell et al., 2008; Coogan and He, 2018; McFarland and Wolpaw, 2010).

Before tackling specific issues with respect to the ways in which speech decoding research is undertaken, it may assist the reader to envision the conceptual design of a DS-BCI (Fig. 1). In this design, a person begins speaking or imagines speaking while their brain activity is recorded using one of several possible acquisition techniques (Fig. 1(a and b)). These neural signals are then processed before features are extracted and classified as different constituents (Fig. 1(c-e)). Finally, a textual representation is generated, and auditory feedback provided to the speaker using text-to-speech (TTS) synthesis (Fig. 1(f and g)). This is only one of several possible designs for such a system. Another method may involve direct synthesis of speech from neural activity in a way that reduces feedback latency by foregoing intermediate text stages (Angrick et al., 2021; Herff et al., 2019). Yet another could decode textual representations which could enable paralysed persons to control personal devices or communicate via the internet (Herff et al., 2015). Ultimately, the aim of these approaches is to provide accurate auditory representation of a person’s imagined speech which not only engages the attention of an interlocutor but also provides essential feedback to the user in a closed-loop system.

Recent promising proof-of-concepts in the speech decoding space have demonstrated its potential using a range of neural recording techniques and experimental procedures (Angrick et al., 2021, 2019a; Anumanchipalli et al., 2019; Cooney et al., 2021; Li et al., 2021; Martin et al., 2016; Moses et al., 2021, 2019, 2018; Nguyen et al., 2017). These include a machine translation system for translating cortical activity to text (Makin et al., 2020) and a bimodal decoding of speech from simultaneously recorded electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) (Cooney et al., 2021). Despite these advances, many challenges persist. Our previous analysis of the implications of neurolinguistics research for DS-BCI applications demonstrated the importance of defining specifically what imagined speech is, its relationship to overt speech, and how its neural correlates can be recorded (Cooney et al., 2018). Other challenges range from difficulties representing neural activity as natural language in experimental settings (Derix et al., 2012, 2014) and validation of participants’ performance of imagined speech (Hwang et al., 2016; Watanabe et al., 2020), to the development of specific machine learning algorithms for decoding speech and speech processes (Angrick et al., 2019a; Cooney et al., 2021; Li et al., 2021; Makin et al., 2020; Nguyen et al., 2017) and the possible confounding influences of different stimuli used to elicit imagined speech (Cooney et al., 2021). Additionally, and just as importantly absent any consensus on the most appropriate methods for researching speech-based communication, researchers have employed a wide range of techniques for designing procedures.

The relationship between overt and imagined speech is of particular importance as, despite its promise for enabling covert communication, research has not always focused directly on imagined speech, often relying on overt speech production (Angrick et al., 2019a; Anumanchipalli et al., 2019; Herff et al., 2019, 2015; Makin et al., 2020; Moses et al., 2019, 2018; Mugler et al., 2014) or perceptive responses to hearing speech (Pasley et al., 2012; Akbari et al., 2019) to investigate the potential development of a DS-BCI.

Considering all the challenges stated and limited acknowledgement in the literature of the pros and cons of different approaches as well as the potential effects they can have on speech decoding, this work evaluates the literature with a focus on the implementation of experimental paradigms used to investigate various aspects of overt and imagined speech decoding. Fig. 2 is a flowchart depicting design considerations researchers can use to construct future speech decoding experiments. Designing paradigms requires interaction of several components and there are potential trade-offs depending on the focus of the study and the resources available. There are also potential opportunities if design considerations are effectively utilised. In our flowchart, a study design must first consider the type of speech to be decoded. Following this is the decision on whether participants should engage in monologic or dialogic speech. This decision influences the selection of stimulus modality used to prompt participants as well as potential investigation of linguistic and prosodic elements of speech. Other decisions, including how to instruct participants to perform consistent speech in experiments or whether to provide feedback must consider all the constituent parts of the experiment to ensure that speech decoding investigations are robustly designed.

With so many challenges, a panoply of different methods have been

---

**Fig. 1.** Conceptual design of a BCI for speech decoding. (a) A BCI for decoding speech begins with a person’s (imagined) speech. (b) Corresponding neural signals are recorded using one of several possible recording techniques. (c) Recorded signals are processed to facilitate maximal information extraction and improvement of signal-to-noise ratio. (d, e) The feature extraction and classification stages compute the most discriminative information in the recorded signals and classify them as parts-of-speech. (f) Decoded signals are represented as text. (g) Text-to-Speech synthesis can be used to convert the text output to enable auditory feedback. In this example, the user actively produces the words “I am thirsty!” with imagined speech. The signals acquired are temporally aligned with each word to facilitate feature extraction and classification. The system produces two outputs: a text printout of the imagined speech words being produced and a synthesized audio output, i.e., “I am thirsty!”.
used in efforts to illuminate some open questions (Cooney et al., 2018; Iljina et al., 2017; Martin et al., 2018, 2015). Here, we demonstrate that experiments vary considerably with respect to stimulus, types of speech, instructions to participants, timings and presence or absence of repetition. We show that different implementations result from a combination of factors, but importantly these are often tied to the specified goals of a study. Section 2 is a summary of the methodology employed to conduct this review. Sections 3 and 4 focus on the different stimuli used to prompt participants and the ways in which experiments are conducted (Fig. 2 – yellow). Section 5 discusses how word selection and linguistic features can influence speech decoding (Fig. 2 – blue). Section 6 reports the ways in which participants are instructed to produce imagined speech during experiments (Fig. 2 – green). The work delineates the various ways in which speech decoding experiments have been carried out to-date and we critically evaluate important components of experimental research such as cuing modality, implementation of experiments and methods for inducing more natural forms of speech. We show the inconsistency in approaches taken to instructing participants in imagined speech and stress the importance of providing consistent directions to an entire cohort and further investigation into the impact of instructing participants. In addition, we highlight the reporting of imagined speech instructions as imperative for enabling thorough understanding of experimental results and their implications.

2. Methods

This review was conducted across research articles that investigated methods for decoding speech and speech processes directly from brain activity. We aimed to synthesize the literature with respect to experimental paradigms, designs and procedures used to elicit, study, and decode speech or speech-related processes. The methodology was based on the PRISMA statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) for undertaking systematic reviews (Moher et al., 2009). We chose to focus this review on several aspects of the experimental procedures used to record and decode speech from brain activity. This includes specific focus on the type of stimuli used to prompt participants, the construction of trial periods and the potential effects and confounds associated with this (even when these are a necessary feature of a design), the possibility of developing novel paradigms to enhance ecological validity and encourage naturalistic speech, and the linguistic characteristics of the phonemes, words and sentences used in experiments.

We considered studies that reported data collection and decoding approaches to overt, imagined and attempted speech. Our previous review highlights important differences between the types of speech under consideration (Cooney et al., 2018). Although overt speech and attempted speech are well-defined and typically referenced by those names, imagined speech is not well-defined and has been identified using several common synonyms in the literature, including inner speech and covert speech. Studies using the terms attempted or intended speech have been classified as such. Others using terms such as covert speech have been evaluated to determine whether they are referring to imagined or attempted speech. For example, studies in which participants have produced overt and covert speech have been classified as imagined speech as participants clearly retain their voluntary motor system (Pei et al., 2011a, 2011b). Studies described in this review were not restricted based on recording technique and include non-invasive neural recording techniques (EEG, fNIRS) and those using implanted electrodes (electrocorticography (ECoG), micro-electrode arrays). Due to the inclusion of recording techniques that do not facilitate high-volume participant data collection, no limits on sample size were set. Given that research in this area often aims to restore communication in patients who have lost the ability to produce speech, no limitations by neurological disorder (e.g., spinal cord injury, anarthria) were applied. Despite our focus on speech decoding, the decoding methodology is not itself part of the inclusion/exclusion criteria. The review aims to examine the experimental methods that may provide opportunities for enhancing speech decoding independent of any specific algorithms.

Research articles were excluded if: (1) No original data collection was undertaken (Angrick et al., 2019b; Cooney et al., 2019a, 2019b; Herff et al., 2019; Proix et al., 2022); (2) No measures were taken to decode speech (i.e., investigatory neuroscience or neurolinguistics articles) (Leuthardt et al., 2012; Moon et al., 2022); (3) The target of decoding was not speech/speech-related (e.g., imagined handwriting (Willett et al., 2021) or imagination of letter shapes (Ullah and Halim, 2021)); (4) The article was a review paper, a book chapter, presentation, or a preprint; (5) The paper did not have a full article (e.g., abstract only). Key search terms were used to identify records in the following categories: (1) Type of speech – examples include “imagined speech”, “overt speech”; (2) decoding speech – examples include “decoding”, “classification”; (3) decoding concerned with direct brain activity – examples include “neural”, “brain”; (4) linguistic units – examples include “vowels”, “phonemes”.

Following identification of records of interest, all articles were first screened based on title, then on abstract. Remaining articles were afforded a full text review to determine final eligibility. A flowchart summarizing the process followed here is presented in Fig. 3, following the measures outlined in the PRISMA statement (Moher et al., 2009).

3. Stimuli used to cue participants in speech decoding studies

Variations in stimulus type and presentation protocol differentially

---

**Fig. 2.** Design considerations for implementation of experiments for neural speech decoding. These are our recommended steps for researchers to consider when designing experiments. The first consideration is the type of speech to be the target of decoding. Other design decisions are predicated on the type of speech to some degree. Researchers should consider whether they wish participants to engage in monologic or dialogic speech (e.g., response to questions). This decision has implications for the type of stimuli used to prompt participants and potential linguistic aspects that may be investigated. These may be linguistic characteristics or prosodic elements of speech. This in-turn may inform the linguistic units i.e., phonemes, words or sentences participants are asked to speak. The stimuli used and the type of speech being decoded can also determine whether real-time feedback is provided and which modality this is provided in. Linguistic Considerations (blue) are discussed in Section 5. Types of Speech (green) are discussed in Section 4.
affect associated neural activity. This is a potential pitfall should representations of the stimulus feed into a speech decoding window. However, studies can be designed to utilise the differential effects of stimuli to further understand its impact on neural speech decoding. A small number of studies have investigated the impact of different stimuli (Cooney et al., 2021; Pei et al., 2011b), and there has been little discussion regarding the direct effect of the different stimuli on experimental results and how they may influence overt or imagined speech decoding. It is important for accurate decoding that stimulus effects such as stimulus-evoked Event-related Potentials (ERP) are not present, or are accounted for, during the initial phase of a decoding window ensuring the stimulus response mechanism does not influence the decoding...
There are several methods that could conceivably be used by experimenters to prompt participants to produce a certain phoneme, word or sentence in overt, attempted or imagined speech. These include text- and audio-based prompts as well as less-often employed image-based stimuli. In addition, alternative methods may be required for novel investigations, for example, those used to examine spontaneous speech decoding (Derix et al., 2014, 2012). Fig. 4 is a visual depiction of these stimulus categorisations, sub-categorised by the type of speech under investigation (see Table 1). Text and audio are the predominant modalities utilised in experiments, with text alone accounting for 52% of studies. It is also noticeable that, despite common employment in linguistics research (Baldo et al., 2013; Dell et al., 2007; Edwards et al., 2010; Eulitz et al., 2006), images corresponding to words are less often used in speech decoding studies (Fig. 4 – yellow). In fact, even combined text and audio presentation is a more common approach than image presentation.

Different practical and experimental considerations associated with each method partially explain the divergence in usage between text, audio and image stimuli. Below we examine the use of these stimulus types in speech decoding experiments to understand why certain methods are applied.

### 3.1. Text-based paradigms

Text-based paradigms, the most common approach, typically involve letters, syllables, words or sentences and usually involve static presentation i.e., it is presented on a monitor in a fixed position for the duration of the cuing period (Agarwal and Kumar, 2022; Nguyen et al., 2017; Riaz et al., 2015), but have been presented dynamically when longer sequences of words are used (Herff et al., 2015; Lotte et al., 2015). A practical advantage of using text to cue participants in speech decoding/production studies is the ease of implementation associated with this modality in relation to others i.e., it is much easier to present text visually on a computer monitor than to record corresponding audio or acquire and display corresponding images. Not unrelated to this is the ease of access to large text corpora such as MOCHA-TIMIT (Wrench, 2000) or the Modified Rhyme Test (MRT) that make implementation of sentence level experimentation much more efficient than the development of bespoke prompts. In addition, presentation of text stimuli is more reliable and more consistent across participants whereas images used to represent the same words may be more ambiguous or subjective in an experimental setting.

Furthermore, stimulus-evoked ERP amplitudes have been shown to be attenuated in response to text when compared to corresponding responses to images (Lüdtke et al., 2008). Despite the stated advantages to this approach, the use of text to present prompts to participants is not perfect. Text-based designs may consider reading of text as equivalent to speech production despite spontaneous speech differing from reading in important ways (Anumanchipalli et al., 2019; Makin et al., 2020). Of greatest concern in the context of speech decoding, is the fact that direct reading does not engage the participant in the earliest stages of speech production (Cooney et al., 2018). It is conceivable that advanced speech decoding may require access to these early phases, including word retrieval and lemma selection. If this is the case, text paradigms may not be ideal. A reasonable counterpoint to this concern is that experiments can benefit from text-based paradigms as they may enable time locking to intermediate stages of imagined speech production.

As stated above, text stimuli are usually presented on a computer monitor, in a central location. In most instances it is proceeded and succeeded by either a blank screen (Fig. 5(a)) or a fixation cross (Fig. 5(b)), but differences exist as to whether text is removed before, during or after the task production period. The paradigm in Fig. 5(a) presented participants with the text form of Japanese vowels on-screen for 200 ms before they were asked to engage in imagined speech production (Ikeda et al., 2014). A similar approach presented participants with one of

### Table 1

| Study                          | Stimulus | Speech        | Recording Technique |
|--------------------------------|----------|---------------|---------------------|
| (Bouchard and Chang, 2014)     | Text     | Overt         | EcoG                |
| (Mugler et al., 2014)          | Text     | Overt         | Micro-electrode    |
| (Lotte et al., 2015)           | Text     | Overt         | Intracortical Array |
| (Herff et al., 2015)           | Text     | Overt         | EcoG                |
| (Ramsey et al., 2017)          | Text     | Overt         | sEEG                |
| (Mugler et al., 2018)          | Text     | Overt         | fNIRS               |
| (Iyayoshi et al., 2018)        | Text     | Imagined      | fNIRS               |
| (Anumanchipalli et al., 2019)  | Text     | Overt         | sEEG                |
| (Makin et al., 2020)           | Text     | Imagined      | fNIRS               |
| (Blakey et al., 2008)          | Text     | Overt         | Micro-electrode    |
| (Wilson et al., 2020)          | Text     | Overt         | Intracortical Array |
| (Ikeda et al., 2014)           | Text     | Imagined      | fNIRS               |
| (Porbadnijk et al., 2009)      | Text     | Overt         | sEEG                |
| (Torres-Garcia et al., 2013)   | Text     | Imagined      | sEEG                |
| (Wang et al., 2015)            | Text     | Overt         | fNIRS               |
| (Riaz et al., 2015)            | Text     | Overt         | sEEG                |
| (Mohanchander and Saba, 2016)  | Text     | Overt         | fNIRS               |
| (Jabangiri and Sepulveda, 2017)| Text     | Overt         | sEEG                |
| (Razzaadreh Sereshkeh et al., 2017) | Text     | Overt         | fNIRS               |
| (Sereshkeh et al., 2018)       | Text     | Overt         | sEEG                |
| (Leuthardt et al., 2017)       | Text     | Overt         | fNIRS               |
| (Kim et al., 2013)             | Text     | Overt         | sEEG                |
| (Sree and Kavitha, 2017)       | Text     | Overt         | fNIRS               |
| (Verontsova et al., 2021)      | Text     | Imagined      | sEEG                |
| (Kroy et al., 2022)            | Text     | Overt         | fNIRS               |
| (Dash et al., 2020)            | Text     | Overt         | sEEG                |
| (Herff et al., 2012b)          | Text     | Overt         | fNIRS               |
| (Moses et al., 2021)           | Text     | Imagined      | fNIRS               |
| (Zhang et al., 2012)           | Text     | Overt         | sEEG                |
| (Moses et al., 2018)           | Text     | Imagined      | sEEG                |
| (Moses et al., 2010)           | Audio    | Overt         | fNIRS               |
| (Kellis et al., 2010)          | Audio    | Overt         | fNIRS               |
| (Stavisky et al., 2018)        | Audio    | Imagined      | fNIRS               |
| (Bigham and Kumar, 2010)       | Audio    | Imagined      | sEEG                |
| (Deng et al., 2010)            | Audio    | Imagined      | sEEG                |
| (Chi et al., 2011)             | Audio    | Imagined      | sEEG                |
| (Min et al., 2016)             | Audio    | Imagined      | sEEG                |
| (Balaji et al., 2017)          | Audio    | Imagined      | sEEG                |
| (Lee et al., 2020)             | Audio    | Imagined      | sEEG                |
| (Agbari Bejestani et al., 2022)| Audio    | Imagined      | sEEG                |
| (Yoshimura et al., 2016)       | Audio    | Imagined      | sEEG                |
| (Ilwang et al., 2016)          | Audio    | Imagined      | sEEG                |
| (Chaudhary et al., 2017)       | Audio    | Imagined      | sEEG                |
| (Martin et al., 2014)          | Audio    | Imagined      | sEEG                |
| (Martin et al., 2016)          | Audio    | Imagined      | sEEG                |
| (Watanabe et al., 2020)        | Audio    | Imagined      | sEEG                |
| (Meng et al., 2021)            | Audio    | Imagined      | sEEG                |
| (Guenther et al., 2009)        | Audio    | Imagined      | sEEG                |
| (Brumberg et al., 2011)        | Audio    | Imagined      | sEEG                |
| (Fargier et al., 2018)         | Audio    | Imagined      | sEEG                |

(continued on next page)
Table 1 (continued)

| Study                        | Stimulus | Speech          | Recording Technique |
|------------------------------|----------|-----------------|--------------------|
| (Wan et al., 2022)           | Image    | Overt           | Micro-electrode    |
| (DaSalla et al., 2009)       | Image    | Imagined        | ECG                |
| (Li et al., 2021)            | Combined | Overt           | EcoG               |
| (Herff et al., 2016)         | Combined | Imagined        | EEG                |
| (Angrick et al., 2019)       | Combined | Overt/Imagined  | Micro-electrode    |
| (Zhao and Rudzicz, 2015)     | Combined | Imagined        | EEG                |
| (Poi et al., 2011b)          | Combined | Overt/Imagined  | EEG                |
| (Pressel Coretto et al., 2017)| Combined | Overt/Imagined  | EEG                |
| (Cooney et al., 2021)        | Other    | Overt           | EcoG               |
| (Derix et al., 2012)         |          |                 |                    |
| (Derix et al., 2014)         |          |                 |                    |

Fig. 5. Three text-based paradigms. (a) Text presentation of one of three Japanese vowels. A 200 ms text presentation phase is succeeded by a 1100 ms task production phase during which a blank screen is displayed. A fixation cross is presented for a randomly permuted time period between 2 and 3 s. This is followed by a task production period during which 1 of 5 vowels is presented. Finally, a post-task production period of 2-3 s is implemented. (c) Generalization of presentation method for sequences of words. Characters coloured green are those currently visible to the participant. Red characters to the left of the monitor are those already viewed and red characters to the right are those yet to be viewed.

In fifteen Japanese syllables on a monitor for only 300 ms (Ibayashi et al., 2018). In both cases, task production began only after stimuli had been removed. In contrast, another study using text stimuli, allowed the prompt to remain on-screen for the duration of the task production period (Nguyen et al., 2017).

In the paradigm in Fig. 5(b), a relatively lengthy pre-trial period is denoted with a fixation cross before it gives way to text presentation of one of the five English vowels. Many studies use a similar approach to this as it is a relatively easy method for presenting single units of speech (AlSaleh et al., 2018; Angrick et al., 2021, 2019b; Mugler et al., 2018). Other studies have used a selection of monosyllabic words from the MRT for participants to read aloud (Mugler et al., 2014; Wilson et al., 2020).

In Mugler et al. (2014) for each trial, one word was presented on a monitor for 3 s, followed by a blank screen for 1 s. Participants were asked to begin reading aloud as soon as they recognized the stimuli.

Although most text-based paradigms use single units of speech such as words or vowels, several have considered longer sequences and therefore required different implementations for stimulus presentation of text. In order to generate speech synthesis from EEG corresponding to overt speech, Anumanchipalli et al. (2019) instructed participants to read aloud and/or freely speak sentences from well-known texts and the MOCHA-TIMIT database. Importantly, it was noted by the researchers that in most cases participants read aloud. However, one participant read sentences describing several picture scenes and then freely described the scenes aloud, and also spoke sentences in free-response interviews (Anumanchipalli et al., 2019). The same database was used in another recent ECoG study, in which participants were once again instructed to read text aloud (Makin et al., 2020). In both studies, there was no effort to induce the participants to engage in anything like spontaneous speech production as reading of text was accepted as sufficiently analogous to speech production.

Presenting long strings of text can be more challenging than displaying single elements as depicted in Fig. 5a and b. Fig. 5c is a generic depiction of how sequences of text can be presented in speech decoding studies. Martin et al. (2014) used a procedure like this to present participants with short stories which were scrolled through on a computer monitor while they read aloud or covertly. Text excerpts from political speeches or children’s stories were used, and the text was scrolled from left to right at the vertical centre of a monitor at a rate that the user was comfortable with. A similar method has been applied with an eye-tracker used to modulate the rate of text flow as the participant read from the monitor (Lotte et al., 2015). This approach may be more difficult to implement than standard text presentation paradigms, particularly as it must be adapted for each participant. However, the procedure enabled the researchers to investigate sentence level decoding in a dynamic manner. In a study investigating the representation of semantic information in the brain, words from 10 to 15 min long stories were presented one word at a time using a rapid serial visual presentation technique (Deniz et al., 2019). In an interesting design, these words were presented for a period of time aligned to those of the spoken versions of the stories.

The literature shows that there is no uniform methodology for incorporating text as stimuli within an experimental paradigm. There have been studies that have explicitly engaged in direct reading from text, whereas others have taken steps to separate reading from speech by removing the stimuli before task production. It is not yet clear what effect these different approaches have on eliciting imagined speech which is consistent with spontaneous imagined speech and whether decoding performance measured with such paradigms would transfer to spontaneous imagined speech. Additionally, further investigation into differences in neural correlates associated with reading and speech is required to understand the extent to which they can be considered analogous in experiments (see Section 4).

3.2. Audio-based paradigms

Auditory stimuli are also commonly presented in speech production/decoding experiments (Deng et al., 2010; Martin et al., 2016; Meng et al., 2021; Moses et al., 2019, 2018) (Fig. 4). A number of studies have used auditory stimuli to investigate the potential for decoding a person’s response to hearing that stimuli, rather than using it to prompt speech production (Akhari et al., 2019; Moses et al., 2018). However, audio is more often used to cue the words or phrases that participants are being asked to actively inner speak. A significant advantage that audio possesses over other cuing mechanisms is that pronunciation is made
explicit to participants. If not sufficiently coached, participants may be susceptible to mis-pronounce syllables or words presented as text, or misname images used to represent those words. Instruction is key in those cases (see Section 6), but audio can mitigate some degree of ambiguity. Martin et al. (2016) for example, used auditory stimuli to prompt participants to imagine reproduction of the stimuli using the cue wording. Audio presentation also offers the possibility of investigating prosodic elements of speech, such as rhythm (Brigham and Kumar, 2010; Deng et al., 2010; Watanabe et al., 2020) (see Section 4.2).

However, when not being used to investigate specific aspects of the phenomena, there are several drawbacks that may limit the use of audio in speech production/decoding experiments. A potential concern is that completion of audio stimuli may occur too close to the beginning of a task period, causing neural representations of perception to be present during a decoding window. It is possible that this issue may be more pronounced for imagined, rather than overt speech. Given that imagined speech is generally associated with a more distributed network of brain regions, there is potential for areas associated with perception such as the Superior Temporal Gyrus to be more prominent in imagined speech decoding (Shuster and Lemieux, 2005). Protocols can attempt to mitigate this by implementing an interlude between stimuli and task production as in Deng et al. (2010) (see Fig. 16(c)). An additional drawback is the relative difficulty in collecting and presenting audio data in comparison with text presentation as recording audio samples for a specific experimental procedure can be time-consuming. However, this is becoming less problematic with greater availability of high-quality TTS software. There is also an issue with respect to the length of time required to play an audio clip to cue participants during experiments. A number of text-based presentations have displayed their stimuli for less than 300 ms (Ibayashi et al., 2018; Ikeda et al., 2014) (see Section 4.1). However, the minimum length of the cuing period for audio-based experiments is a function of the length of the audio clip. This in-turn may limit the number of trials that can be undertaken during a given session and therefore constrain overall data collection. However, this is not an issue associated with audio alone as in text presentation increasing the number of phonemes, syllables or words per trial necessarily elongates the trial period (Herff et al., 2015; Lotte et al., 2015).

Despite the concerns surrounding audio stimuli, this paradigm has been used in many DS-BCI studies (Fig. 4). Auditory cues have regularly been utilised to induce rhythmic aspects to produced speech (Brigham and Kumar, 2010; Chi et al., 2011; Mohanchandra and Saha, 2016) (see Section 4.2) and to ask questions to which participants respond (Chaudhary et al., 2017; Moses et al., 2019) (see Section 4.3). In one study, researchers instructed an epilepsy patient by verbally articulating the word to be produced multiple times prior to the participant beginning to speak themselves (Kellis et al., 2010). However, a standard approach to audio stimulus presentation is depicted in Fig. 6. In this study, a 1 s cue period was used to play audio of one of five vowels which a participant would then imagine speaking during a 3 s trial period (Min et al., 2016). A similar procedure was used in (Martin et al., 2016), in which an auditory cue corresponding to one of six words was used to prompt the participants’ imagined speech. Hwang et al. (2016) used a novel approach to auditory stimuli by combining two distinct pieces of audio corresponding to an initial phrase and a critical word. This formed part of a question-and-answer paradigm (see Section 4.3) and allowed experimenters to construct multiple audible questions from each separate piece of audio. Questions were asked over a combined period of 5 s and a fixation cross was presented on-screen during this time (Fig. 7).

A recent study used auditory stimuli in a listening task, with the stimuli also used to cue overt and imagined speech production within the same procedure (Watanabe et al., 2020). Each stage of the experiment was communicated to participants using the words LISTEN, SPEAK and IMAGINE (Fig. 8). During the listening phase, speech stimuli were played to participants through headphones. An interesting aspect of this experiment is the instruction to participants to produce overt and imagined speech at the same rate as the auditory stimuli, and the use of a gradually extending progress bar to assist participants in this task (Fig. 8). This is an advantage of using audio stimuli to prompt participants and helps to mitigate variable speech production duration across trials and across subjects, particularly for imagined speech. Interestingly, results indicated that EEG oscillations during imagined speech contained the signature of overt speech of the same task (Watanabe et al., 2020).

The literature reviewed in this section demonstrates several possible approaches for presenting audio stimuli in an experimental setting for both perception and production studies, while also delineating potential challenges associated with this modality. The next section provides a similar analysis to the use of images to prompt speech production.

### 3.3. Image-based paradigms

Picture naming tasks use images corresponding to words to investigate various aspects of language and are extensively used in linguistics experiments for examining aspects of speech production (Baldo et al., 2013; Dell et al., 2007; Edwards et al., 2010; Eulitz et al., 2000). For example, investigating the neural correlates of picture naming, Baldo et al. (2013) used a naming paradigm in which black and white drawings of multiple high (e.g. bed) and low frequency (e.g. abacus) items were presented to participants who were instructed to respond spontaneously. The experimental paradigm presented in Fig. 9 demonstrates the use of images in a linguistic experiment in which the effects of different procedures are examined (Roelofs, 2018). The figure is representative of how images are often used in linguistics experiments. In protocol (a), continuous presentation of images in different ordinal positions within a semantic category is implemented. Protocol (b) uses blocked cycling within semantically homogeneous and heterogeneous cycles and protocol (c) overlays images with semantically related or unrelated distractor words.

Despite common usage in psycholinguistics, image stimuli are generally underutilised in DS-BCI research relative to text and audio (Fig. 4). Reasons for the general omission of image stimuli may be the

![Fig. 6. Auditory presentation of English vowels in an imagined speech experiment (Min et al., 2016) (adapted under CC BY 4.0). The procedure begins with a 1 s period during which an audible ‘beep’ sound alerts participants that a cue is imminent. An audible vowel sound is then presented during a 3 s period. Two beeps were then presented sequentially at an interval of 300 ms. Finally, a 3 s period was provided for imagined speech production.](image-url)

![Fig. 7. Audible presentation of questions using distinct audio clips for initial phrases and critical words (Hwang et al., 2016). This fNIRS study implemented a 10 s preparation period, followed by 2 disjoint auditory periods presenting distinct parts of a question. The second of these was followed by a 10 s task production period. A 5 s validation period allowed participants to indicate their answers and a 5 s rest period lead into the next trial.](image-url)
possible scope for ambiguity associated with the word–image relationship and potential confounding effects resulting from the neurological response (evoked potentials) to image presentation. Ensuring that the image clearly matches the word to be produced is crucial in speech decoding experiments as participant mistakes are costly in terms of data loss or corruption. Ensuring that pairs of images in a stimulus set are appropriately differentiated is critical as visually similar cues could also lead to task production mistakes or blurred difference between neural correlates of associated speech production. As stated in Section 3.1, research has indicated that images produce a greater ERP response in participants than corresponding text representations (Lüdtke et al., 2008). In addition, Cooney et al. (2021) showed significant effects of stimuli on overt and imagined speech decoding, with images resulting in higher decoding accuracies than either text or audio for imagined speech. A recent study used image presentation with the expressed intention of mitigating potential temporal correlation artefacts in imagined speech decoding (Li et al., 2021). Specifically, eight two-character Chinese words were presented to participants as photographic images during a 1 s period. Participants were instructed to judge each picture before pronouncing the corresponding words. However, the extent to which the different stimuli influence speech production/decoding experiments is not currently known.

There are several studies related to speech decoding that have used image stimuli but do not meet the criteria for inclusion in the categorisations presented in Fig. 4. One such ECoG study used 16 pictures from four different semantic categories (clothing, animals, musical instruments, and human dwellings) to investigate the dynamics of word retrieval in speech production (Riès et al., 2017). Participants were shown a picture and had to name it as fast and accurately as possible. The focus on word retrieval enabled using a picture-naming task has potential significance for speech experiments, as this stage of speech production is bypassed when text and audio stimuli are used (see Sections 3.1–2). One of the studies investigating imagined speech with image stimuli looked at imagined speech production during picture-naming in a single-subject undergoing a pre-surgical procedure fMRI scan, reporting general blood oxygen level dependent (BOLD) response overlap with normal speech production networks (Brumberg et al., 2019). An EEG study investigating the phonetic properties of word onset during overt speech used a picture-naming task in which a subset of black and white pictures were extracted from a database (Alario and Ferrand, 1999). These were presented to participants during a 1.5–2 s period and they were directed to produce a word immediately after stimulus onset (Fargier et al., 2018). Recently, Wandelt et al. (2022) used an image paradigm to cue a participant to produce overt speech (alongsie a separate MI task). In one condition, the participant was shown one of five visual images associated with human grasping (same condition used for MI task) and in another one of five squares of different colours were shown. Results showed that during the “grasp” cuing period neuronal activity was very similar during MI and overt speaking tasks and that these both differed to activity during the colour presentation period.

One of the early imagined speech decoding studies introduced pictorial presentation of mouth-shapes used to produce the vowels /a/ and /u/ (DaSalla et al., 2009) (Fig. 10). These prompted participants to
produce the vowel. Participants were asked to respond immediately to stimuli and maintain this for 2 s (Fig. 10). The study reported readiness potential and peak speech-related potentials 200 ms and 350 ms after stimulus presentation. More recently, Anumanchipalli et al. (2019) had one participant freely describe picture scenes. The experimental paradigm presented in Fig. 11 used both auditory and visual cues during experiments but the different stimuli were prescribed to different tasks (Lee et al., 2019). Auditory stimuli were used to prompt imagined speech production (Fig. 11 - top) whereas pictures were used to induce visual imagery (Fig. 11 - bottom). In the example presented in Fig. 11, a picture of a clock was used to prompt visual imagery of a clock rather than the word clock. For both tasks, participants were instructed to perform four iterations during each trial period which potentially led to blocking effects on accuracy as noted by Porbadnigk et al. (2009).

Evidence presented in this section demonstrates that although not as popular as text or audio protocols, image stimuli have been explored as a potential method for investigating speech decoding. Several likely reasons for this were discussed, including the potential for ambiguity in the word-image relationship and possible stimulus effects associated with images. Another major limitation of using images in speech decoding experiments is the constraint it places on longer sequences such as sentence level decoding. Both audio and text can be adapted just as easily for phonemic or single-syllable presentation, as they can for sentences and paragraphs. Even for overt speech, sequences produced in response to images can be transcribed to extract phonetic or syllabic timings. However, this is not the case with imagined speech trials where produced speech may be limited to words, or a single semantic concept represented by multiple words. This being the case, it is not possible to fully utilise the syntactic dependencies between words in longer sequences when using image stimuli for speech production/decoding experiments.

Image stimuli potentially offers access to information in the speech production process not available with text or audio paradigms e.g., stages such as conceptual preparation and lexical selection. A possible asset of picture-naming in relation to speech decoding is that it can incorporate these earlier stages into the process of producing speech in experiments, with potential benefits including investigation into the dynamics of word retrieval. Ries et al. (2017). A noted concern with picture-naming is that image stimuli may inflate accuracies in speech decoding experiments due to the increased cognitive load associated with the task (Martin et al., 2018). This would be a significant counterweight to the potential benefits of picture-naming if the higher cognitive demand of the task was a primary driver of differences in performance.

3.4. Multi-modal paradigms

Experiments investigating decoding of produced speech have typically employed a single stimulus modality for cuing participants. Several have used more than one stimulus type for performing experiments and related fields have used multimodal experiments to understand how different modes of input corresponding to a common concept are represented in the brain (Price, 2012; Vigneau et al., 2006). Deniz et al. (2019) demonstrated that semantic information coming from text and audio are similarly represented across the cerebral cortex. There have been several multimodal speech decoding experiments, but in general they have not been constructed in ways that enable understanding of important factors such as the effects different stimuli may have on speech production or the extent to which stimulus-evoked responses
may be difficult to disentangle from speech processes. Most of the studies categorised in Fig. 4 as having implemented a combined stimulus approach to experiments used text and audio as the two stimuli. In most cases, multiple modalities are simply used in-tandem during a single trial period and therefore make it difficult to disentangle their respective influence on decoding (Angrick et al., 2019a; Cooney et al., 2020; Herff et al., 2016; Pressel Coretto et al., 2017; Zhao and Rudzicz, 2015).

However, presentation of text and audio during different trials has also been considered (Pei et al., 2011b). A generalised depiction of how studies have implemented dual-stimulus approaches is presented in Fig. 12. Typically following a pre-trial period, a word or syllable is presented on a computer monitor while that same word is simultaneously presented to participants auditorily via a speaker or headphones. Following completion of the audio clip, both stimuli are removed, and the speech production period begins.

The EEG studies of Zhao and Rudzicz (2015) and Pressel Coretto et al. (2017) both used a similar protocol to the one presented in Fig. 12. The first of these utilised a dual-stimulus approach in a combined overt and imagined speech experiment (Zhao and Rudzicz, 2015). In this instance, words or phonemes were presented as text on a monitor and auditorily through loudspeakers. The stimulus period was followed by a 5 s period in which subjects would imagine speaking. This was immediately followed by an overt speech period. In Pressel Coretto et al. (2017), the experiments required Spanish speakers to speak either one of five vowels or one of six words with either overt speech or imagined speech. Following a pre-cue period of 2 s, participants were simultaneously presented with a text stimulus of the vowel or word and an auditory stimulus of the same via headphones.

The ECoG studies of Herff et al. (2016) and Angrick et al. (2019a) differed from the above EEG studies in that combined text and audio stimuli were used to present sentences rather than single words or phonemes. In both studies, sentences were presented to participants as text on a computer monitor for 4 s while audio of that sentence was played through speakers positioned adjacent to the display.

Pei et al. (2011b) used text and audio stimuli independently for different trials, thus facilitating some analysis of potential effects of the stimuli. In this ECoG study, trials consisted of one of 36 possible words being presented visually on a computer monitor or auditorily through headphones. Random presentation of one or the other stimuli was implemented for each trial.

Results demonstrated that overt speech response to auditory stimuli was delayed in comparison with those following text stimuli and that cortical activations varied according to the type of stimuli presented. Although time courses across all electrodes revealed differential peaks in response to text and audio, sparse distribution of electrodes over visual areas made it difficult for the authors to draw concrete conclusions.

Cooney et al. (2021) reported relative decoding performance of trials using text, audio, and image stimuli to prompt participants to perform overt and imagined speaking. Using the paradigm depicted in Fig. 13, participants were asked to perform identical overt and imagined speech tasks with the type of stimuli used to prompt speech changing between trials. Results indicated statistically significant effects of stimulus type on decoding accuracies for both speech types.

Despite several examples of multiple stimuli being used to prompt speech in decoding studies, the fact that the modalities are generally presented together makes it difficult to draw any concrete conclusions regarding the effects of the different stimuli on decoding. Given the complexity of imagined speech and the likelihood that different stimuli may differentially affect speech production, it is crucial that potential differences are further investigated. Additionally, multimodal experiments using text and audio face a potential difficulty in configuring the amount of time the two modalities are presented to a participant. However, it is possible to ensure that text and audio are presented for a precisely equal time period to ensure faithful comparison (Deniz et al., 2019). Studies such as Pei et al. (2011b) can enable useful comparative analysis and provide insight for future methodological advances in the field. This requires thoughtful design of experiments to ensure that the impact of the different stimuli on decoding can be robustly analysed.

4. Implementation of experimental protocols

Each of the experiments referenced in Table 1 used one of the four approaches to stimulus presentation discussed in Section 3 (excluding spontaneous speech studies (Derix et al., 2014, 2012)). However, reviewing the mode of stimulus presentation alone only provides a partial account of how experiments have been conducted and the methods used by researchers to investigate different aspects of speech decoding. Different units of speech, timing in experiments, methods for examining rhythm in speech decoding, and the use of question-and-answer protocols to mimic dialogic speech are aspects of
the experimental procedure discussed in this section.

4.1. Experimental timings

Timing of experiments varies widely across studies. This variance is a feature of the interactions between the many different design considerations highlighted in Fig. 2. An important facet influencing timing in experiments is the data acquisition protocol in use. EEG and ECoG offer temporal precision in the order of milliseconds, whereas the haemodynamic response captured by fNIRS, or fMRI can take several seconds to play out fully. Although experiments can be adapted to take advantage of high (Angrick et al., 2019a; Bouchard and Chang, 2014; Brigham and Kumar, 2010; Deng et al., 2010; Herff et al., 2015; Ikeda et al., 2014; Lotte et al., 2015; Moses et al., 2018; Mugler et al., 2014; Nguyen et al., 2017) and low (Chaudhary et al., 2017; Herff et al., 2012a; Hwang et al., 2016; Sereshkeh et al., 2018) frequency recordings, the data acquisition protocol is not the only important factor, and there is variation within studies using the same protocols as well as between those using different methods. Studies using EEG and ECoG have used task production periods as short as 1 s (Ikeda et al., 2014; Martin et al., 2016), 2 s (Agarwal and Kumar, 2022; Cooney et al., 2021; DaSalla et al., 2009; Lee et al., 2020) and 3 s (Kirov et al., 2022; Min et al., 2016; Riaz et al., 2015).

Short task production periods are often associated with one-time production of a unit of speech i.e., no repetition (Agarwal and Kumar, 2022; Cooney et al., 2021; Ikeda et al., 2014), with studies demonstrating that 1 s is enough for participants to produce single words in overt speech (Martin et al., 2016; Mugler et al., 2014). However, this is not the only reason why short trials are selected. Lee et al. (2020) explicitly stated that motivation for their 2 s task production period was in part to acquire relatively large amounts of data and Riaz et al. (2015) were able to use a 3 s task production period to gather 100 trials per class. Another factor is the use of multiple conditions within a single trial period (e.g., listening, overt and imagined speech) and the necessary trade-off between trial duration and the number of trials (Martin et al., 2016). In this study, timing interacted with selection of stimuli to minimize word length variability to a maximum of 20 ms (Martin et al., 2016). Longer trial periods have been utilised when the subject of study has gone beyond typical production of single units of speech. Studies investigating the impact of repetition of vowels or words throughout a trial have used 4–9.8 s periods (Nguyen et al., 2017; Pressel Coretto et al., 2017), and others investigating the decoding of speech rhythm have used trial periods of 6 s to facilitate rhythmic production of speech (Brigham and Kumar, 2010; Deng et al., 2010). Investigations into continuous speech production can require participants to read texts with several hundred words (Herff et al., 2015; Lotte et al., 2015). This leads to significantly extended trial periods, which in one study ranged between 126 s and 590 s (Herff et al., 2015).

Characteristically, fNIRS studies have implemented longer trial periods, ranging from 8 s (Herff et al., 2012a) to 15 s (Chaudhary et al., 2017), and it might be expected that this extended time course would lead to studies investigating sequences of word production. However, with the exception of Herff et al. (2012a) our analysis indicates that this is not the case. Several fNIRS studies have utilised yes/no responses to questions, continuously repeated over a lengthy trial period (Chaudhary et al., 2017; Hwang et al., 2016; Sereshkeh et al., 2018). Whether to fully exploit the classification potential of each trial period, investigate the prospect of dialogic speech decoding, or design experiments to ensure patient participation, these studies are indicative of the myriad considerations suggested in Fig. 2. The use of parsimonious responses to questions provided to patients may be utilised to facilitate a more manageable experimental procedure (Chaudhary et al., 2017), as factors related to a participant’s ability to engage in a given experimental procedure can also influence the timings researchers can implement. Ikeda et al. (2014) constrained trial periods for patients with intractable epilepsy to 1100 ms on the basis of medical advice and Chaudhary et al. (2017) limited total session time to 9 min for experiments involving four patients with differing degrees of ALS. A patient’s attentiveness and cognitive abilities can impact upon trial construction, with at least one study varying the rate at which words were presented based on the state of individual participants (Lotte et al., 2015).

Another important feature of experimental procedures which influences the timing of a protocol is the relationship between the stimulus presentation and task production periods. There are three common methods for addressing this issue (Fig. 14). The first is to construct a paradigm which requires participants to begin overt or imagined speech immediately upon perceiving a given stimulus (Cooney et al., 2021; DaSalla et al., 2009; Li et al., 2021; Mugler et al., 2014; Nguyen et al., 2017; Pei et al., 2011a, 2011b; Ramsey et al., 2017) (Fig. 14(a)). In this scenario, there is no clear distinction between the stimulus presentation period and the task production period. This can assist researchers in optimising the time required per trial and perhaps enable a greater number of trials per class to be collected in a session, while also avoiding any requirement for participants to hold prompts in memory. However, it might also be associated with difficulty in disentangling stimulus-based neural response from behavior-based neural response.

The second common approach to incorporating stimuli and task production is to have two distinct stages of the protocol, with the stimulus period directly preceding the task production period (Chaudhary et al., 2017; Herff et al., 2016; Hwang et al., 2016; Ikeda et al., 2014; Pressel Coretto et al., 2017; Wang et al., 2013) (Fig. 14(b)). In these studies, stimuli are presented to participants for a pre-determined period. Upon removal of the stimuli, the task production period begins immediately, and participants are required to start speaking. A third common approach is for the stimuli and task production periods to be separated by a defined interlude between the two (Deng et al., 2010; Kellis et al., 2010; Lee et al., 2020; Martin et al., 2016; Zhao and Rudzicz, 2015) (Fig. 14(c)). This gap between stimulus presentation and task production has ranged from several hundred milliseconds (Martin et al., 2016; Min et al., 2016) to multiple seconds (Brigham and Kumar, 2010; Deng et al., 2010; Zhao and Rudzicz, 2015; Sree and Kavitha, 2017) and requires participants to hold the target phoneme, word or phrase in working memory prior to task production.

Fig. 14. Generalization of three methods for implementing stimulus presentation and task production during speech decoding experiments. (a) Stimulus presentation and task production periods fully overlap. Here, participants are expected to begin task production immediately upon perceiving the stimulus. (b) Stimulus presentation is immediately prior to the task production period. Here, participants engage in task production immediately following removal of stimuli. (c) An interlude between stimulus presentation and task production is implemented. Participants must retain the target in memory during this period.
This is an additional cognitive load not associated with spontaneous speech and may impact decoding.

The studies of Fargier et al. (2018) and Mugler et al. (2018) both used the first of these paradigms as they presented image and text stimuli to participants. However, this simultaneous presentation and task production procedure has not been used in any of the audio studies presented in Table 1. The imagined speech EEG study of AlSaleh et al. (2018) used a text presentation paradigm similar to the one presented in Fig. 14(b). One of eleven words or phonemes was presented for a period of 2 s, after which a blank screen was used to cue the subject that they should now perform the task. The task production period was 2 s Martin et al. (2016) implemented a 500 ms gap between the end of an auditory stimuli period and the beginning of an imagined speech period. Another study indicated that a 2 s interlude between stimulus presentation and task production was used to allow participants to move their articulators into position before they began pronunciation (Zhao and Rudzicz, 2015). The protocol of Wilson et al. (2020) resembles the example in Fig. 14(b) but with novel implementation. Here, participants were presented a word on a monitor with an accompanying red square for 1.2–1.8 s. This preparatory phase was succeeded by a production period which was prompted by an audible beep, the word text changing to ‘Go’ and the red square turning to green. In this example the participant’s monitor did not go blank but the contents changed to indicate a period of speech. In an equally novel approach, Moses et al. (2021) employed a visual “countdown” to prepare a participant to begin attempted speech. In this case, a word with four period characters (e.g. “…. Hello….”) was displayed on a monitor in white text. Over a 2 s period, the outer period characters on each side would disappear every 500 ms. When the final periods disappeared, the text would turn green, indicating go, and the participant would attempt to produce speech. This paradigm was favored by the participant from a set of potential paradigm choices as it aided consistency in helping to align production attempts with the cue.

4.2. Rhythmic cuing

Rhythmic cuing has been used in several speech decoding studies to investigate the temporal dynamics of patterns of repetition. A number of studies have investigated the dynamics of rhythm during speech production by comparing multiple rhythmic formulations (Brigham and Kumar, 2010; Deng et al., 2010; Watanabe et al., 2020), while several others implemented rhythmic patterns in experiments without providing for any comparative analysis (Nguyen et al., 2017; Pressel Coretto et al., 2017). Furthermore, several imagined speech studies have used continuous repetition of letter or words with no specific rhythmic pattern implemented (Hwang et al., 2016; Min et al., 2016; Pressel Coretto et al., 2017). It is possible that this approach may improve decoding performance, but the extent to which this is the case and the consequent impact on information transfer rate (ITR), a common performance metric in BCI (Obermaier et al., 2001; Wolpaw et al., 1998), is unclear.

The study of Brigham and Kumar (2010) used the experimental procedure presented in Fig. 15 in which an intended rhythm was presented to participants with three audible clicks prior to their beginning imagined speech. During a 6 s task production period, participants were asked to imagine speaking a syllable once every 1.5 s, resulting in three imagined syllables per trial. Depending on different postprocessing regimes, decoding performance was below, or only slightly better than chance. Additionally, the extent to which different rhythms may influence speech decoding is not examined as only one procedure is considered.

Deng et al. (2010) expanded upon the previous paradigm by asking participants to imagine speaking one of two syllables (/ba/ and /ku/) in one of three rhythms (Fig. 16). This research examined the potential for rhythm associated with imagined speech production of syllables to be decoded from EEG. Each trial began with a period during which both the syllable and the rhythm were cued (audible clicks). This was followed by a task-production period during which participants would imagine speaking the cued syllable in the cued rhythm. Two classification tasks were reported with this study. The first was a 6-class problem in which each of the three rhythms for each of the two syllables were considered classes. The second saw syllables combined for each rhythm, resulting in a 3-class problem. Results showed that decoding performance was not significant when the syllables were considered part of the task. However, when only the three rhythms were classified, average performance was 58.05% against a significance threshold of 42%. Results indicated that rhythmic structure could be detected in non-invasive neural recordings and suggest that it may be a stronger feature than the unit of speech itself. The authors suggest that brain activity during imagined speech resembles the temporal structure of the speech itself and therefore can be a target for speech decoding.

This is backed up by another study reporting greater than chance decoding performance between three different speech rhythms (Watanabe et al., 2020). Considering the synchronization between the overt speech envelope and EEG oscillations during imagined speech, the study used rhythmic overt speech production of the syllable /ba/ to cue participants to speak. The waveforms in Fig. 17 depict the three rhythms where participants were asked to listen to, speak aloud and imagine speaking. The paradigm was used to demonstrate that there is a degree of synchronization between imagined speech and the envelope of the corresponding overt speech. Despite above chance performance, results were not highly significant (38.5%; chance: 33.33%), and the question remains open whether the time required to conduct rhythmic speech experiments is worth the potential benefits in terms of decoding performance.

During the experimental protocol presented in Fig. 18, an audible beep was played to participants to assist them in establishing the required rhythm during imagined speech production (Nguyen et al., 2017). The beep was repeated five times with a period, T, which was dependent on the length of the word to be produced. T was 1 s for short words and 1.4 s for long words. Following removal of the beep, participants were instructed to continue imagining speaking in the same rhythm for the remainder of the task production period. Therefore, trials with short and long words had different rhythmic patterns. Nguyen et al. (2017) report reasonable classification accuracies across several tasks. However, the highest mean accuracy was obtained in a short vs long word task with different repetition intervals (80.05%). Given the results reported in previous speech rhythm studies (Deng et al., 2010; Watanabe et al., 2020), it is likely that the repetition strategies may have contributed to classification scores as much as differences in the words being produced.

Another study used separate rhythmic and continuous imagined speech production of words and vowels (Pressel Coretto et al., 2017). In the vowels case there is no control over the number of repetitions per trial, so it is possible that this feature of the experiment varied from person to person and perhaps among the different vowels. This is a scenario where the gradually extending progress bar discussed in Section 3.2 may be applicable (Watanabe et al., 2020). For words, participants were prompted with three audible beeps to produce speech during a trial. However, the results presented in the original paper are not significantly above chance and do not present comparative analysis of
the two repetition strategies. It is therefore unclear what overall effects these approaches may yield in comparison with single unit production studies (Agarwal and Kumar, 2022; Cooney et al., 2021). In protocols similar to the vowels strategy above (Pressel Coretto et al., 2017), other research has used continuous production during a trial period without specifying any particular rhythm or count (DaSalla et al., 2009; Hwang et al., 2016; Min et al., 2016). In general, results reported do not indicate with any certainty that continuous production is a significant boost in decoding performance (DaSalla et al., 2009; Hwang et al., 2016; Min et al., 2016). However, Hwang et al. (2016) suggest that the results of their study in which participants produced continuous imagined yes/no responses to questions over a 10 s trial period may represent a feasible state for a BCI to aid communication for persons with speech pathologies. The reason for this is that despite ITR reduction due to a longer trial period, potential increases in accuracy could sufficiently increase ITR to achieve a net performance gain. There are cases where this trade-off would be acceptable given the communicative potential on offer.

The rhythmic paradigms described above are a novel way of investigating temporal aspects of speech production that may be useful for future decoding applications. However, there are concerns with this approach with respect to its relationship to natural imagined speech production. One issue is that repetition of single units of speech is not typical of natural language and may be boosting decoding performance in comparison with studies using single productions per trial. Another concern with rhythmic cuing is that these manually constructed intervals between repeated phonemes or words themselves become the target of a decoding algorithm. In this respect, it is possible that the content of the speech is being ignored in favour of temporal dynamics associated with rhythm. Potential research to address these concerns could include direct comparison of continuous production strategies (DaSalla et al., 2009; Hwang et al., 2016; Min et al., 2016; Pressel Coretto et al., 2017) with single unit production strategies (Agarwal and Kumar, 2022; Cooney et al., 2021), to provide a clearer picture of potential performance gains and impact on ITR. Additional investigations are required to understand ways in which decoding rhythm complements or detracts from speech decoding. Finally, these repetition strategies do not represent aspects of natural language such as syntactic
Several of the experiments discussed above used these text- and audio-based stimulus types to implement a question-and-answer protocol. These protocols have become more common in recent years as researchers seek to mimic natural dialogic conversation in an experimental setting (Chaudhary et al., 2017; Hwang et al., 2016; Moses et al., 2021, 2019; Sereshtekh et al., 2019). The most popular implementation of this format is the yes/no response. The fNIRS study presented in Fig. 7 asked participants a series of 70 disjunctive questions, which required either a yes or no response (Hwang et al., 2016). Questions were split into two parts, the first consisting of the formulation of a question (e.g., “Are you?”) and the second consisting of a critical word (e.g., “hungry?”). Splitting the different elements of the questions facilitated multiple combinations. Balaji et al. (2017) also used recorded audio to ask participants questions that required unambiguous yes/no answers. However, the study of Sereshtekh et al. (2019) used similar yes/no responses but presented questions as text on a computer screen, rather than orally.

The procedure of Moses et al. (2019) increased the complexity of the question-and-answer paradigm by asking participants questions which had an associated non-binary set of possible answers. For example, the question “How is your room currently?” could be answered with one of five possible responses (“Bright, Dark, Hot, Cold, Fine”). The experimenters reported several question sets which were accompanied with several possible answers per question. The technique first involved decoding participants’ ECoG while they listened to a question being asked. As participants responded aloud, the decoded questions were used as priors to inform an answer decoding model. The method has potential for constraining the decoding space associated with each specific question while still providing BCI users with greater degrees of freedom by enabling the construction of multiple question and answer sets. A recent attempt to verify speech decoding visually prompted a person suffering with anaphoria (the loss of the ability to articulate speech) statements and questions to which they would attempt to answer with words from a predefined set of 50 (Moses et al., 2021).

One of the advantages of the question-and-answer paradigm is that it facilitates a degree of ecological validation with respect to responses given in imagined speech as the question can effectively constrain the number of likely answers. In Section 4.4, another technique proposed for enhancing the ecological validity in which experiments use post-trial procedures to effectively give participants the power to label their own data is discussed.

4.4. Post-trial validation to mitigate stimuli effects

Mitigating stimuli effects is a considerable challenge in DS-BCI studies, particularly in those utilising imagined speech for communication. Self-controlled post-trial validation of participant responses during single trials by participants is one technique designed to minimise the impact of stimuli on speech production and its neural correlates. In the question-and-answer paradigm of Hwang et al. (2016), a post-task production period of 5 s allowed participants to indicate with a button press whether they had answered ‘yes’ or ‘no’ in response to the previous question. With this procedure, responses to questions were not fixed in advance of data acquisition and participants were expected to respond naturally to each new question as if they were taking part in a conversation. This method allowed the use of 70 different questions and therefore no repetition of cues to prompt participants’ production of imagined speech.

The procedure presented in Fig. 8 used a post-trial validation mechanism through which participants could indicate whether or not they had successfully produced imagined speech (Watanabe et al., 2020). This took the form of a positive or negative button press corresponding to success or failure of task production. In behavioural studies, the experimenter has the authority to remove from the data those trials that do not conform to the designed procedure. This option is not available for imagined speech experiments and the virtue of this post-trial approach is that it enlists a method of excluding bad trials that would have otherwise been included in training or test data. Unfortunately, the study does not indicate what impact the removal of bad trials had on imagined speech decoding. At present, it is unclear to what extent this approach can positively affect imagined speech decoding, and further research is required.

A potential criticism of post-trial validation techniques is that they remove control of experimental procedures from the researcher and place it in the hands of the participant. It is reasonable to suggest that these methods are simply replacing one element of uncertainty i.e., around imagined speech production, with a different element of uncertainty i.e., the accuracy of the participant’s post-trial validation responses. Additionally, these types of experiment increase cognitive load and potentially fatigue. These are not trivial concerns. However, the tremendous difficulty in verifying imagined speech in experiments makes these methods potentially valuable tools for experimenters.

Other methods used to enhance the robustness of experiments include the use of microphone recording to verify no overt speech has been produced during imagined speech trials (Martin et al., 2016), silent or idle states to enable comparison with active speech (DaSalla et al., 2009; Min et al., 2016), and a Go/No Go procedure in which participants produced speech (Go) or did nothing (No Go) during trials (Wandelt et al., 2022). In particular, the use of control conditions or control groups is something that researchers can consider when designing experiments and the subsequent analyses.

Together, the literature presented in this section demonstrates high variability in the way speech decoding experiments have been designed. Several factors directly influence the construction of a single trial, including the data acquisition protocol, the presence or absence of repetition during task production and the relationship between the stimulus presentation period and the task production period. None of these component parts of the experiment are fully independent and it is likely that trade-offs must be made when designing effective experiments.

5. Word selection and linguistic implications

Careful selection of phonemes or words for prompting participants in overt and imagined speech studies has not tended to be the subject of focused attention for researchers. Given the complexity of other aspects of neural speech decoding that require attention it is understandable that several studies in the literature have not paid close attention to the fine-grained linguistic properties of test items. However, there are opportunities for researchers to uncover important facets of our understanding of the potential of neural speech decoding by utilising linguistic expertise in experimental designs. Although there are instances of experiments being designed to examine possible linguistic implications for speech decoding (Chi et al., 2011; Zhao and Rudzicz, 2015), these are in a minority of studies. Many studies have attempted to decode overt, imagined or attempted speech by asking participants to produce phonemes (Brumberg et al., 2011; Jahangiri and Sepulveda, 2019; Zhao and Rudzicz, 2015), syllables (Brigham and Kumar, 2010; Deng et al., 2010) or individual letters (Agarwal and Kumar, 2022; Ikeda et al., 2014; Iqbal et al., 2015; Yoshimura et al., 2016) during experiments. Studies have indicated that the phonological properties of different phonemes (e.g., presence of nasal or bilabial syllables) can influence decoding (Jahangiri and Sepulveda, 2019; Zhao and Rudzicz, 2015) and another used phonemes to include jaw, lip, tongue and velum movement as well as two fricatives (Chi et al., 2011). However, a thorough examination was not provided in either case. A virtue of attempting to decode phonemes or
syllables is that they can be used to construct words and sentences. Being able to decode such small components offers the prospect of utilising methods from automatic speech recognition and TTS synthesis that facilitate combining phonemes or syllables into words or phrases for communication. This could enable the generation of a larger inventory from a relatively small number of classes (60–80 phonemes in spoken English (Vaseghi, 2007)). Transcription codes such as ARPABET have been used in the Carnegie Mellon University Pronouncing Dictionary (Weide, 1998) and can facilitate extremely large vocabularies.

However, such approaches to decoding limit access to neural representations of semantic and syntactic information in relation to words and phrases, and to dependencies between words in sentences. Despite there being several studies which have considered overt or imagined speech decoding of words (Kim et al., 2013; Lee et al., 2019; Martin et al., 2016; Pei et al., 2011b; Zhao and Rudzicz, 2015), they have not always contemplated the potential effects of semantic or phonological relationships between the words. For example, Porbadnigk et al. (2009) chose the first five words of the international radiotelephony spelling alphabet (‘alpha’, ‘bravo’, ‘charlie’, ‘delta’, ‘echo’) without reporting the implications of the phonological or semantic aspects of the words. The study of Martin et al. (2016) selected words (‘spoon’, ‘cowboy’, ‘battlefield’, ‘swimming’, ‘python’, ‘telephone’) to maximise variance with respect to number of syllables, semantic categories and acoustic features. However, the extent to which any of these characteristics impacted decoding performance is not discussed in the results presented. Studies attempting to decode phonological properties of speech could look to build on techniques used in linguistics and neurolinguistics studies. For example, studies into phonological influences on word recognition often utilise homophony to investigate effects relating to words sharing phonological features while having different semantics (Newman and Connolly, 2004). Phonological activation has been shown to be important for accessing meaning when reading text (Coltheart et al., 1988), and it may be possible to use homophones (e.g. “rose”/“rows”; “steal”/“steel”) and pseudohomophones (e.g. “fog”/“phog”; “brain”/“brane”) to further investigate the interaction between phonology and semantics on neural speech decoding.

Phonological similarity has been considered for imagined speech decoding with the words ‘pat’, ‘pot’, ‘knew’, and ‘gnaw’ forming phonologically similar pairs (Zhao and Rudzicz, 2015) but no analysis of the impact that the degree of similarity exhibited was reported. It is likely that careful selection of words can enable semantic discrimination of overt and imagined speech. Eight Korean monosyllabic words categorised as part of the face (‘cheek’, ‘nose’, ‘eye’, ‘mouth’) or as a number (‘three’, ‘five’, ‘nine’, ‘ten’) were used for this purpose (Kim et al., 2013). Although statistically significant differences in brain activity associated with the two semantic conditions (face and number) were reported, the overall effect of semantics on speech decoding was not. There may be an opportunity to investigate the relative decoding potential of semantic and phonological features of speech by studying words that share meaning but have different phonology, and vice versa. Using the word “powerless” as an example, imagine two binary classification examples: 1. semantically similar words (“powerless” vs “weak”), 2. phonologically similar words (“powerless” vs “powerless”). Examples such as this could be used to ask the question: In some conceptual imagined speech decoding space is “powerless” more closely related to “powerless” or to “weak”? Another potentially interesting paradigm could be to compare sentences under different conditions. Sentences with and without negation, for example, “John is not going to the beach”, and sentences in different tenses like “Jane is reading her book” and “Jane was reading her book” could be used to investigate the extent to which these two syntactic categories evoke differential neural responses and the consequent effect this might have on decoding. Studies utilising images to prompt participants must use visually expressive words. Li et al. (2021) used pictures of types of fruit (e.g., ‘watermelon’) and animals (e.g., ‘rhinoceros’), while Cooney et al. (2021) presented images representing motor-related action words (e.g. ‘kick’) and two-word pairs (e.g. ‘red ball’).

Words with significant functional usage such as care-related words (Lee et al., 2020; Mohanchandra and Saha, 2016) or useful commands (Kirov et al., 2022; Vorontsova et al., 2021) have been decoded. Lee et al. (2020) used words selected from a communication board commonly used in hospitals for paralysed and aphasic patients (Patak et al., 2006). The twelve words/phrases were, ‘ambulance’, ‘clock’, ‘hello’, ‘help me’, ‘light’, ‘pain’, ‘stop’, ‘thank you’, ‘toilet’, ‘TV’, ‘water’, and ‘yes’. Another study used the words ‘water’, ‘help’, ‘thanks’, ‘food’ and ‘stop’ (Mohanchandra and Saha, 2016). These commands and responses have obvious utility for patients with limited communication and support important interactions between caregivers and patients. However, from a purely experimental perspective they have not been selected in a manner that facilitates examination of how semantics, phonology or syntax may influence decoding performance. A hypothetical improvement to single word or short phrase commands for these patients may be to utilise words with high imageability or concreteness ratings. Although imageability of words i.e., words that are easily and readily imagined, have been shown to differentially effect brain activity (Klaver et al., 2005), it would be interesting to determine whether this translates to differences in discriminating words from recorded neural activity. Abstract nouns such as “love” or “alternative” could be compared with concrete words such as “leg” or “ambulance”. If positive decoding effects of high-imageability were determined, it potentially provide significant benefit to patients with communication difficulties. Alsaleh et al. (2018) attempted to decode words corresponding to directions (‘Left’, ‘Right’, ‘Up’, ‘Down’), responses (‘Yes’, ‘No’) and emotions (‘Happy’, ‘Sad’, ‘Help’) without reporting how the different categories may have affected decoding. The emotional valence of words is another area of potential future study. Sources such as The Affective Norms for English Words (ANEW) database rate the positive/negative (e.g., friendly/betray) valence or high/low arousal level (e.g., pressure/kettle) of words (Bradley and Lang, 1999). It is worth considering whether these emotional and arousal factors have any impact upon decoding beyond the words, and whether it is possible that any effects would be common across a cohort for the same words. One study has examined the effect that length of word has on imagined speech decoding accuracy, finding that accuracy improved significantly when longer words are compared with shorter words (Nguyen et al., 2017). However, it is possible that this is nothing more than direct decoding of the temporal difference in time taken to pronounce longer versus shorter words.

In other studies, stimuli included sequences for sentence-level speech decoding. In a pilot ECoG-to-Speech synthesis study, Herff et al. (2016) presented sentences to participants visually or audibly for 4 s. Participants were then expected to recite the prompted sentence aloud from memory. In total, 50 sentences were recited from the Harvard sentence corpus (Rothenauer, 1969). Another study presented participants with short stories which were scrolled through on a computer monitor while they read aloud or covertly (Martin et al., 2014). Text excerpts from political speeches or children’s stories were used and the text was scrolled from left to right at the vertical centre of a monitor at a rate that the user was comfortable with. Variability associated with speaking rate, pronunciation and speech errors are cited as possible issues with sentence level decoding of speech. However longer sequences can facilitate a greater number of features of language. Moses et al. (2018) demonstrated this by decoding fricatives, nasals, and diphthongs among other phoneme types from participant neural responses to perceiving spoken input sentences.

The MOCHA-TIMIT database (Wrench, 2000) has been used recently to prompt participants to read sentences aloud for speech decoding applications (Anumanchipalli et al., 2019; Makin et al., 2020). Use of this type of corpus enables the use of language decoding models that take advantage of information on the relationships between elements i. e., words, in a sequence. However, this method is currently more applicable to studies using implanted electrodes with overt speech as the noisier signals associated with non-invasive data acquisition, and the
difficulty with time-locking experiments with imagined speech makes sequence modelling extremely difficult. Sentence level decoding is more common in overt speech studies than with imagined speech. This is unsurprising given the difficulty in time-locking imagined speech production (Cooney et al., 2018). The experimental setup proposed in Fig. 19(b) is one way that researchers could potentially investigate both sentence-level and single word decoding. In this example, the paradigm consists of several important elements. The first of these is a set of short sentences, with each sentence sharing several words in common with several alternative sentences. Each sentence communicates a distinct idea. Second, during trials, a sentence is randomly presented as text on a screen and the participant must retain it in memory when stimuli are removed and a pretrial interlude (during which the monitor is blank) is implemented. Third, during the task production period, a number of progress bars corresponding to the number of words in the sentence are presented on-screen in an approach similar to that of Watanabe et al. (2020), or the “countdown” approach of Moses et al. (2021). Here, each progress bar would be activated sequentially, so that each word can be produced in imagined speech in a time-locked manner.

This section indicates that while many different approaches to the selection of phonemes, words or sentences for decoding applications have been studied, it is not obvious how these different methods affect decoding performance from a linguistics point of view. In addition, several potential methods for incorporating complimentary aspects of linguistics and neural speech decoding research within studies have been suggested. The use of phonemes, syllables and vowels may offer the possibility of developing a more expansive lexicon by combining these constituent parts into new words and sentences, but they do not facilitate direct analysis of implications associated with semantic and phonological aspects of words and syntactic features of sentences. Despite there being several studies utilising a range of different words for speech decoding, and with some even considering semantic or phonological characteristics, it is nevertheless still unclear whether these features have any significant impact on increasing separability between the associated neural correlates and thus on decoding performance. The studies presented here are limited in their analysis of possible effects of word selection and potential influence of linguistic characteristics. Further analysis is required to ascertain the impact of different linguistic characteristics on speech decoding tasks.

6. Imagined speech instructions are inconsistent and often inscrutable

It is well known that subjective perception of imagined speech varies from person to person (Alderson-Day and Fernyhough, 2015). In addition, analysis of the literature presented in Cooney et al. (2018) makes it clear that the phenomenology of imagined speech is not yet concretely understood and that it is implicated in multiple functional networks in the brain (Alderson-Day and Fernyhough, 2015; Iljina et al., 2017). Furthermore, the label imagined speech is used to refer to several different processes, and its relationship to overt speech remains an open question. Therefore, providing participants with clear and consistent instructions on what is meant by the term imagined speech and how they are expected to produce it in experiments is complex and yet essential. Several, although not all, studies have reported the detailed instructions they have provided participants prior to their engagement in imagined speech tasks. These can range from asking participants to “read in mind” (Wang et al., 2013) to instructing them to “internally answer” in response to questions (Hwang et al., 2016). It is important to note the differences in instructions provided to participants and to consider the possible effects they have on task performance, decoding performance and the conclusions drawn.

Experiments invoking imagined speech as the mode of communication are reasonably uniform with respect to instructions provided to participants on what not to do during task performance. This is typically expressed as directions to avoid overt vocalization as well as movement of any articulatory muscles such as the tongue or lips (Cooney et al., 2021; Ikeda et al., 2014; Kim et al., 2013; Lee et al., 2020; Li et al., 2021; Nguyen et al., 2017; Porbadnikg, 2009; Sereshkeh et al., 2018; Zhao and Rudzicz, 2015). Participants have been explicitly instructed to avoid making any sounds during imagined speech trials. Nguyen et al. (2017) instructed participants to avoid any overt vocalization while Lee et al. (2020) asked participants to imagine speech without “making the sound” associated with overt speech. At their least explicit, instructions can direct participants to imagine speaking “without moving” (Zhao and Rudzicz, 2015). However, in general they specify that it is movement of the articulatory muscles associated with speech production i.e., lips, tongue, jaw which must be suppressed (Ikeda et al., 2014; Porbadnikg et al., 2009; Sereshkeh et al., 2018).

Even in the presence of instructions to avoid muscle movement or overt articulation, there is still significant room for variation in the imagined speech produced by participants. Instructions are often reported as simply asking participants to imagine speaking a given phoneme, word or sentence (Brigham and Kumar, 2010; Min et al., 2016; Porbadnikg et al., 2009; Sree and Kavitha, 2017; Zhao and Rudzicz, 2015). Similar terminology has been used to ask participants to “covertly” articulate or speak in response to stimuli (Ikeda et al., 2014; Jahangiri and Sepulveda, 2017). However, there are several examples of studies in which imagined speech has been defined differently or in which instructions given to participants do not use common terminology (Balaji et al., 2017; Chaudhary et al., 2017; Kim et al., 2013; Sereshkeh et al., 2018, 2017; Wang et al., 2013; Watanabe et al., 2020). Among the terms used to direct participants to produce imagined speech are the words internally (Hwang et al., 2016; Nguyen et al., 2017) and think (Chaudhary et al., 2017; Sereshkeh et al., 2018). Nguyen et al. (2017) asked participants to pronounce words internally in their minds while Chaudhary et al. (2017) instructed participants to “think yes” or “no” answers in response to questions. Although these instructions do provide reasonable direction to participants, they omit any specification of prosodic elements (e.g., pitch, tone) or whether the participant should

![Fig. 19. Experimental proposals for incorporating linguistics in imagine speech decoding experiments. (a) The proposal is to investigate the relative decoding potential of semantic and phonological features of speech by studying words that share meaning but have different phonology, and vice versa. Here, “powerless” semantically similar to “weak” but phonologically similar to “powerless”, but what effect does this have on decoding? (b) Proposed experimental paradigm in which decoding of sequences and isolated words can be investigated. Several short sentences communicating different meaning but containing several common words. Sentences are presented on-screen as text and participants must retain this prompt in memory. Progress bars are used to control the timing of each element in the sequence during imagined speech production.](image-url)
perceive their own voice or that of another during imagined speech and may therefore limit the constraints on variance in production among participants.

More ambiguous terms used to instruct participants to produce imagined speech include directions such as “read in mind” (Wang et al., 2013) or “continuously dwell upon” (Balaji et al., 2017) the prompts provided during trials, or to “mentally rehearse” yes or no responses to questions (Rezazadeh Sereshtehk et al., 2017; Sereshkeh et al., 2017). Another study reported that the participant “subvocalized the word in his mind” (Mohanchandra and Saha, 2016). It is not clear how these instructions affect a participant’s performance of imagined speech in comparison with the basic instruction to imagine speaking, nor is it clear whether it is intended to. Asking participants to imagine repeating a prompt (Pei et al., 2011b), particularly those presented auditorily, likely leads to a different voicing of imagined speech than instructing participants to imagine speaking as if performing overt speech (Lee et al., 2020). Indeed, one study instructed participants to “imagine hearing” in response to auditory stimuli as they were interested in the auditory perceptual representation of imagined speech (Martin et al., 2016), demonstrating the importance of clear instructions as well as clear reporting of instructions. Several studies have emphasized the link between overt and imagined speech articulation in their instructions. In two studies in which both overt and imagined speech were investigated, participants were instructed to imagine performing the overt articulation task without actually vocalizing the word (Kim et al., 2013; Watanabe et al., 2020). Other studies have asked participants to perform imagined “vocal articulation” (Chi et al., 2011) and “imagined vocalization” (DaSalla et al., 2009).

Most of the studies referenced above have been concerned with phoneme or word decoding and have tended to avoid the use of reading in experiments. Studies investigating the potential of sentence-level decoding of imagined speech have directed participants to “imagine reading” sentences aloud (Herff et al., 2012b) or to “read [text] silently” (Martin et al., 2014). These instructions resemble recent overt speech decoding studies in which participants are asked to read directly from sources of text (Anumanchipalli et al., 2019; Makin et al., 2020). Similar studies involving imagined speech would pose challenges in time locking speech to the text being read and thus labelling the neural data. The study of Anumanchipalli et al. (2019) had an additional condition in which participants “silently mimed” speech i.e., they performed the same articulatory movements as in an overt speech trial without producing any sound. The precise relationship between overt and imagined speech, and mimed speech is not clear and requires further research, but the study demonstrated decoding of spectral features of speech during miming.

Some recent studies have reported more concrete instructions provided to participants. Kirov et al. (2022) asked participants to “say the words at their usual pace and in a normal voice”, thus removing some of the ambiguity that has been present in many of the studies reported. Cooney et al. (2021) ensured that all participants were presented with an identical set of instructions by formalising them in text as part of their experimental design (Table 2). The study provides specific information on how participants should attempt to produce imagined speech: “imagine speaking in your natural voice/ accent and at a normal tempo”; while also stating that participants should avoid replicating the sound of audio prompts and visualization of image prompts. Detailed instructions offer several benefits. They help ensure that participants are comfortable with what they are being asked to do and that ambiguity surrounding the task is kept to a minimum. This in turn can be expected to reduce the variance within a cohort with respect to how imagined speech is produced. Prosodic elements of speech can be important aspects of the research aims of a study, making detailed and consistent instruction essential. Another important factor impacting instruction, is the type of stimuli used in experiments and the potential interaction between that and speech production. For example, it may be required in some studies that a participant repeat auditory prompts precisely as they are presented, or they may alternatively be asked to respond to such prompts in their natural voice or accent.

In this section we demonstrated the variety of ways in which participants can be instructed to perform imagined speech during experiments. However, at present it is very difficult to quantify any differences that are borne out by the use of different terms such as “think”, “mentally rehearse”, “covertly articulate” or “internally in their minds” (Chaudhary et al., 2017; Ikeda et al., 2014; Nguyen et al., 2017; Rezazadeh Sereshkeh et al., 2017; Sereshkeh et al., 2018, 2017). In fact, it is difficult to discern the extent to which any terminology beyond a parsimonious instruction to “imagine speaking” impacts decoding performance or neural representations of speech. That being the case, there is a need for studies that are designed in ways that can help answer these questions. What impact the presence or absence of comprehensive and consistent instructions to produce imagined speech like those reported in Cooney et al. (2021) have on participants’ ability to communicate through a BCI is one of these questions. It is possible to construct an experiment in which one group of participants are provided with detailed instructions while another group are provided with a minimum instruction such as “covertly articulate”. The goal of this type of research would be to ascertain whether the uniform, detailed instructions have a consistent effect on imagine speech production and decoding across a cohort. A possible addition to this protocol could be the inclusion of a control group where participants are provided with identical stimuli but asked not to produce any speech. This addition would allow researchers to study the impact of stimuli in speech decoding experiments.

Another open question is the extent to which the specific content of detailed instructions may differentially affect imagined speech production and decoding. This could be investigated by providing two groups with identical audio prompts of sequences of speech, but two different instructions to imagine speaking. Differing instructions like “imagine speaking in your natural voice/ accent and at a normal tempo” and “imagine speaking the words presented with the same voice as in the audio” could help researchers begin to understand the effect of attempting to engage participants in imagined speech with different prosodic elements. Other studies may look to investigate whether there are any important differences resulting from instructions to “think”, “mentally rehearse” or “read in mind” and whether or not they actually lead to different communication strategies being employed by participants. The interaction between stimuli, instruction and the strategy implemented by the participant is another area open to investigation. Where a text-based paradigm may ask participants to “read [text] silently” (Martin et al., 2014), an image-based one may instruct participants to pronounce an associated word in imagined speech. It is likely that these different approaches would yield different outcomes, but the specific impact remains unclear.

Clearly, experiments such as these require more careful construction than we can provide here, but the proposals may lead to important future research in this field. In addition, we believe it would be beneficial, where possible, for researchers to be clear about what is meant by imagined speech, or any alternative terminology, when designing

| Study | Instruction |
|-------|-------------|
| 1. | Internally pronounce the words or phrases presented without emitting any sounds or making any facial movements. |
| 2. | The imagined speech should be produced in much the same way as overt speech, without the motor-based articulation. Thus, imagined speech should be produced ‘as if’ the person was going to produce overt speech. |
| 3. | The imagined speech should represent your typical production of internal speech i.e., try to imagine speaking in your natural voice/ accent and at a normal tempo. |
| 4. | Try to avoid replication of the sound of the audio prompt or visualisation of the object in picture prompts, when producing imagined speech. |

Table 2: Imagined speech instructions provided to participants in Cooney, Folli and Coyle (2021).
experimental procedures and analysis. Although there are not currently any definitive correlations between instructions and experimental outcomes, it is unlikely that instructions would be detrimental to experiments while potentially providing some of the benefits suggested here. In addition, it is imperative that clear and consistent directions are provided to participants, unless the objective of the study contradicts this, to ensure that there is uniformity within a cohort. Some studies have provided instructions in the form of a script to ensure that all participants received identical information prior to undertaking experiments (Alsaeleh et al., 2018; Cooney et al., 2021). Finally, it is important for the development of the field that directions provided to participants are precisely reported so that comparative analysis and replication can be accurately and consistently performed.

7. Conclusion

If the full potential of a BCI system for neural decoding of speech is to be realised, design and implementation of targeted and robust experimental procedures is fundamental consideration. In this review we have demonstrated that a battery of different protocols has been applied in speech decoding experiments, with the type of stimulus, timing of procedures and the instructions given to participants on how to produce imagined speech all being components that vary among experiments. Targeted design of experimental paradigms enables researchers to study specific aspects of speech decoding. However, despite the enormous variety in the design of experimental protocols and the selection of stimulus methods, it is difficult to accurately gauge the relative effects of the different approaches. This requires further comparative study into the impact of different stimuli as well as consistent and comprehensive reporting of all procedure and protocol employed.

There are numerous challenges associated with the investigation of speech decoding from neural signals. These include difficulties in representing natural language in experimental settings, in effectively eliciting speech production through stimulus presentation, and uncertainty regarding participants’ performance of imagined speech. With respect to imagined speech, the phenomena itself is still relatively unknown and poorly understood. Substantial experimentation, neural decoding and neuroimaging at multiple scales is required to better understand ways in which it can be accurately decoded and constrained. Innovations aimed at improving the robustness of experimental procedures for the investigation of imagined speech decoding have been reviewed and differentiated. These include the use of post-trial validation, question-and-answer paradigms, and a progress bar to assist participants in pacing their imagined speech.

Despite there being no clear consensus on the most appropriate methods for investigating overt and imagined speech decoding, text and audio stimulus presentation paradigms were predominant among the literature reviewed. Text is a useful modality due to its ease of implementation and familiarity to study participants. Audio disambiguates any potential pronunciation errors and facilitates examination of temporal aspects of speech, such as rhythm. Picture-naming paradigms are relatively under-represented in BCI research, a symptom that may reflect experimenters’ concerns around potential ERP effects and the potential for direct visualization of objects rather than imagined speech of the corresponding words. All paradigms have trade-offs which must be managed. Text stimuli lends itself to direct reading of the prompt and thus cannot represent spontaneous speech production. Auditory stimuli on the other hand could lead to direct decoding of the stimuli, as has been demonstrated previously (Akbari et al., 2019).

The construction of experimental procedures has also been shown to vary significantly, with the relationship between the stimulus presentation period and the task production period, and the total length of the task production period, being important features of experimental protocols, which researchers must address given the objectives and constraints of their experiments. Similarly, attempts to harness specific linguistic aspects of speech, including semantics, syntax and phonology, are present in the literature but have not yet been fully utilised to enable thorough analysis of the effects these features may have on speech decoding. Several proposals have been made here for potential experimental paradigms. We anticipate that this review will aid researchers in designing and reporting on informative experiments aimed at advancing the current state-of-the-art in neural speech decoding.

Funding sources

This work was supported by: the Tier 2 High Performance Computing resources provided by the Northern Ireland High Performance Computing (NI-HPC) facility, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant No. EP/T022175/1. Damien Coyle is funded by UKRI Turing AI Fellowship 2021–2025, funded by the EPSRC, Grant No, EP/V025724/1. Ciaran Cooney was funded by a PhD studentship provided by the Department for the Economy, Northern Ireland.

Competing interests

The authors declare no competing interests. Prof Damien Coyle if Founder and CEO of neurotechnology company, NeuroCONCISE Ltd.

References

Aagarwal, P., Kumar, S., 2022. Electroencephalography based imagined alphabets classification using spatial and time-domain features. Int. J. Imaging. Syst. Technol. 32, 111–122. https://doi.org/10.1002/ima.22655.

Akbari, H., Khalighinejad, B., Herrero, J.L., Mehta, A.D., Megzarani, N., 2019. Towards reconstructing intelligible speech from the human auditory cortex. Sci. Rep. 9, 1–12. https://doi.org/10.1038/s41598-018-37599-z.

Alarri, P., Ferrand, L., 1999. A set of 400 pictures standardized for French: Norms for name agreement, image agreement, … Behav. Res. Methods, Instruments, Computers. 31, 531–552.

Alderston-Day, B., Fernyhough, C., 2015. Inner speech: development, cognitive functions, phenomenology, and neurobiology. Psychol. Bull. 141, 931–965. https://doi.org/10.1037/bul0000621.

Alsaeleh, M., Moore, R., Christensen, H., Arvaneth, M., 2018. Discriminating between Imagined Speech and Non-Speech Tasks using EEG, in: Annual International Conference of the IEEE Engineering in Medicine and Biology-Proceedings.

Angrick, M., Herff, C., Johnson, G., Shih, J., Kruisienki, D., Schultz, T., 2019a. Interpretation of convolutional neural networks for speech spectrogram regression from intracranial recordings. Neurocomputing 342, 145–151. https://doi.org/10.1016/j.neucom.2018.10.080.

Angrick, M., Herff, C., Mugler, E., Tate, M.C., Sutczky, M.W., Kruisienki, D.J., Schultz, T., 2019b. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. J. Neural Eng. 16 https://doi.org/10.10117/jneuroeng.2019.1022.

Angrick, M., Ottenhoff, M.C., Diener, L., Ivucic, D., Ivucic, G., Goulis, S., Saal, J., Colon, A.J., Wagner, L., Kruisienki, D.J., Ruben, P.L., Schultz, T., Herff, C., 2021. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. Commun. Biol. 4, 1–10. https://doi.org/10.1038/s42003-021-02578-0.

Anumanchipalli, G.K., Chartier, J., Chang, E.F., 2019. Speech synthesis from neural decoding of spoken sentences. Nature 568, 493–498. https://doi.org/10.1038/s41586-019-1119-1.

Asghari Bejestani, M.R., Mohammad Khani, G.R., Nafisi, V.R., Darakeh, F., 2022. EEG-based multword imagined speech classification for persian words. Biomed. Res. Int. 2022. https://doi.org/10.1155/2022/8333084.

Balaji, A., Haldar, A., Patil, K., Ruthvik, T.S., Ca, V., Jartarkar, M., Barts, V., 2017. EEG-based Classification of Bilingual Unspoken Speech using ANN. 2017 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 1022–1025. https://doi.org/10.1109/EMBEC.2017.8307006.

Baldo, J.V., Arevalo, A., Patterson, J.P., Drnecf, N.F., 2013. Grey and white matter correlates of picture naming: evidence from a voxel-based lesion analysis of the Boston Naming Test. Cortex 49, 658–667. https://doi.org/10.1016/j.cortex.2012.03.001.

Bell, C.J., Shenoy, P., Chalodhorn, R., Rao, R.P.N., 2008. Control of a humanoid robot by a noninvasive brain-computer interface in humans. J. Neural Eng. 5, 214–220. https://doi.org/10.1088/1741-2550/5/2/012.

Blakely, T., Miller, K.J., Rao, R.P.N., Holmes, M.D., Ojemann, J.G., 2008. Localization of primary motor cortex.2012.03.001.

Blakely, T., Miller, K., Rao, R.P., Holmes, M.D., Ojemann, J.G., 2008. Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids. 2008 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 31, 4964–4967. https://doi.org/10.1109/IEMBS.2008.4605238.

Bouchard, K.E., Chang, E.F., 2014. Neural decoding of spoken vowels from human sensory-motor cortex with high-density electrocorticography. Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf. 2014, 6782–6785. https://doi.org/10.1109/EMBC.2014.6945185.
Herf, C., Heger, D., de Pesters, A., Telaar, D., Schalk, G., Schultz, T., 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. Front. Neurosci. 9, 1–11. https://doi.org/10.3389/fnins.2015.00127.

Herf, C., Heger, D., Putze, F., Guan, C., Schultz, T., 2012a. Cross-subject classification of speaking modes using sNIRS. In: International Conference on Neural Information Processing. Springer Berlin Heidelberg, pp. 417–424. https://doi.org/10.1007/978-3-642-34817-5_51.

Herf, C., Johnson, G., Diener, L., Shih, J., Krusienski, D., Schultz, T., 2016. Towards direct speech synthesis from ECoG: A pilot study. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS 2016-Octob, 1540–1543. https://doi.org/10.1109/EMBC.2016.7751904.

Herf, C., Putze, F., Heger, D., Guan, C., Schultz, T., 2012b. Speech mode recognition from functional Near Infrared Spectroscopy, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. pp. 3546–3549. https://doi.org/10.1109/EMBC.2012.6346972.

Hwang, H.-J., Choi, H., Kim, J.-Y., Chang, W.-D., Kim, D.-W., Kim, K., Jo, S., Im, C.-H., 2016. Toward more intuitive brain–computer interfacing: classification of binary intentions using functional near-infrared spectroscopy. J. Biomed. Opt. 21, 091303 https://doi.org/10.1117/1.JBO.21.9.091303.

Ibayashi, K., Kuni, M., Matsuo, T., Ishihita, Y., Shimada, S., Kawai, K., Saito, N., 2018. Decoding speech with integrated hybrid signals recorded from the human ventral motor cortex. Front. Neurosci. 12, 1000221.

Ieda, S., Shibata, T., Nakano, N., Tsuyuguchi, N., Ieda, K., Kato, A., 2014. Neural decoding of single vowels during covert articulation using electrocorticography. Front. Hum. Neurosci. 8, 1–8. https://doi.org/10.3389/fnhum.2014.00607.

Ijino, O., Derix, J., Schirrmacher, R.T., Schulze-Bonhage, A., Aufer, P., Aertsen, A., Ball, T., 2017. Neuroimaging and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication. Brain Comput. Interfaces 4, 186–209. https://doi.org/10.1017/bci.2017.25.

Iqbal, S., Khan, Y.U., Farooq, O., 2015. EEG based classification of imagined vowel sounds. 2015 Int. Conf. Front. Sustain. Glob. Dev. 1591–1594.

Jahangiri, A., Sepulveda, F., 2019. The relative contribution of high-gamma linguistic processing stages of word production, and motor imagery of articulation in class separability of covert speech tasks in EEG. Data J. Med. Syst. 43 https://doi.org/10.1109/19.13797.

Jahangiri, A., Sepulveda, F., 2017. The contribution of different frequency bands in class separability of covert speech tasks for BCIs, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. pp. 2093–2096. https://doi.org/10.1109/EMBC.2017.807266.

Kells, S., Miller, K., Thomson, K., Brown, R., House, P., Greger, B., 2010. Decoding spoken words using local field potentials recorded from the cortical surface. J. Neural Eng. 7, 056007 https://doi.org/10.1088/1741-2560/7/5/056007.

Kim, T., Lee, J., Choi, H., Lee, H., Kim, I.Y., Jung, D.P., 2013. Meaning based covert speech classification for brain–computer interface based on electroencephalography. IEEE/EMBS Conf. Front. Syst. Neur. Eng. 53–56. https://doi.org/10.1109/NER.2013.695869.

Kirov, V.Y., Balbint, O.M., Krivko, E.M., Lazurenko, D.M., Aslanyan, E.V., Shaposhnikov, D.G., Shcherban, I.V., 2022. Spoken and inner speech-related EEG connectivity in different spatial direction. Biomed. Signal Process. Control 71, 70–77. https://doi.org/10.1016/j.bspc.2021.103224.

Klaver, P., Fell, J., Dietl, T., Schier, S., Schaller, C., Elger, C.E., Fernandez, G., 2005. Word substitution task: the flexibility affects the semantic retrieval. Memory 13, 704–712. https://doi.org/10.1080/09658210500203385.

Klaver, P., Fell, J., Dietl, T., Schier, S., Schaller, C., Elger, C.E., Fernandez, G., 2006. Word substitution task: the flexibility affects the hippocampal memory. Hippocampus 16, 213–223. https://doi.org/10.1002/hipo.20222.

Kloos, A., Julier, S., Scaramuzza, D., Siegwart, R., 2011. Feature-based visual servoing with feature tracking and feature mapping. J. Field Robot. 28, 1–23. https://doi.org/10.1002/rob.20458.

Koeck, C., Folli, R., Coyle, D., 2020. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. Sensors 20, 4629. https://doi.org/10.3390/s20164629.

Koeck, C., Korkik, A., Folli, R., Coyle, D., 2020. Classification of Imagined Spoken Word-pairs using Convolutional Neural Networks, The 8th Graz BCI Conference, 2019. https://doi.org/10.3217/978-3-8125-6826-2/2.

Koeck, C., Folli, R., Coyle, D., 2018. Neurolinguistics research advancing development of a direct-speech brain–computer interface. Science 8, 103–125. https://doi.org/10.1126/science.aat6383.

Koeck, C., Folli, R., Hauk, O., 2000. Electroencephalographic activity over temporal lobe during non-experimental, real-life social interaction. Front. Hum. Neurosci. 6, 10840. https://doi.org/10.3389/fnhum.2012.00099.

Koeck, C., Korkik, A., Folli, R., Coyle, D., 2020. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. Sensors 20, 4629. https://doi.org/10.3390/s20164629.

Koeck, C., Korkik, A., Folli, R., Coyle, D., 2018. Neurolinguistics research advancing development of a direct-speech brain–computer interface. Science 8, 103–125. https://doi.org/10.1126/science.aat6383.

Koeck, C., Korkik, A., Folli, R., Coyle, D., 2020. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. Sensors 20, 4629. https://doi.org/10.3390/s20164629.

Koeck, C., Korkik, A., Folli, R., Coyle, D., 2018. Neurolinguistics research advancing development of a direct-speech brain–computer interface. Science 8, 103–125. https://doi.org/10.1126/science.aat6383.

Koeck, C., Korkik, A., Folli, R., Coyle, D., 2020. Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. Sensors 20, 4629. https://doi.org/10.3390/s20164629.
Wang, Y., Gao, S., Gao, X., 2006. Common Spatial Pattern Method for Channel Selection In Motor Imagery Based Brain-computer Interface. IEEE, pp. 5392–5395. https://doi.org/10.1109/IEMBS.2005.1615701 in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference.

Watanabe, H., Tanaka, H., Sakti, S., Nakamura, S., 2020. Synchronization between overt speech envelope and EEG oscillations during imagined speech. Neurosci. Res. 153, 48–55. https://doi.org/10.1016/j.neures.2019.04.004.

Weide, R.L., 1998. Carnegie Mellon Pronouncing Dictionary” release 0.6.

Willett, F.R., Avansino, D.T., Hochberg, L.R., Henderson, J.M., Shenoy, K.V., 2021. High-performance brain-to-text communication via handwriting. Nature 593, 249–254. https://doi.org/10.1038/s41586-021-03506-2.

Wilson, G.H., Stavisky, S.D., Willett, F.R., Avansino, D.T., Kelemen, J.N., Hochberg, L.R., Henderson, J.M., Druckmann, S., Shenoy, K.V., 2020. Decoding spoken english from intracortical electrode arrays in dorsal precentral gyrus. J. Neural Eng. 17 https://doi.org/10.1088/1741-2552/abbfed.

Wolpaw, J.R., Ramoser, H., McFarland, D.J., Pfurtscheller, G., 1998. EEG-based communication: improved accuracy by response verification. IEEE Trans. Rehabil. Eng. 6, 326–333. https://doi.org/10.1109/86.712221.

Wrench, A.A., 2000. A Multichannel Articulatory Database and its Application for Automatic Speech Recognition, in: In Proceedings 5 Th Seminar of Speech Production. pp. 305–308.

Yoshimura, N., Nishimoto, A., Belkacem, A.N., Shin, D., Kambara, H., Hanakawa, T., Koike, Y., 2016. Decoding of covert vowel articulation using electroencephalography cortical currents. Front. Neurosci. 10, 1–15. https://doi.org/10.3389/fnins.2016.00175.

Zhang, D., Gong, E., Wu, W., Lin, J., Zhou, W., Hong, B., 2012. Spoken sentences decoding based on intracranial high gamma response using dynamic time warping. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS 3292–3295. (https://doi.org/10.1109/EMBC.2012.6346649).

Zhao, S., Rudzicz, F., 2015. Classifying phonological categories in imagined and articulated speech. 2015, IEEE Int. Conf. Acoust. Speech Signal Process. 992–996. (https://doi.org/10.1109/ICASSP.2015.7178118).