UNIVERSITY OF GRANADA

UNIVERSITY OF PADOVA

INTERNATIONAL DOBLE MASTER IN COGNITIVE NEUROSCIENCE AND CLINICAL NEUROPSYCHOLOGY/ NEUROCIENCIA COGNITIVA Y DEL COMPORTAMIENTO

MASTER THESIS

THINKING OUT LOUD: UNVEILING BRAIN OSCILLATIONS IN VOWEL AND SEMANTIC DECODING

Presented by:
Ibon Vales Cortina

Supervisor:
Dr. Marc Ouellet
Dr. José Andrés González
Dr. Mario Bonato

Academic Year 2023 / 2024

INDEX

Abstract

This study explores the neural mechanisms underlying imagined speech and its potential for brain-computer interface (BCI) applications, particularly for individuals with severe communication impairments. Using EEG and a novel deep learning model, we aimed to decode linguistic components from imagined speech, focusing on vowels and semantic categories. Our model, which integrates convolutional and recurrent neural network architectures, was validated through the classification of imagined speech, overt speech, and silence, achieving a significant accuracy of 81.76%. High accuracy was also obtained in vowel classification during imagined speech (92.06%) and in semantic categorization (84.88%), surpassing previous studies. We further investigated the role of different EEG frequency bands—Alpha, Beta, Gamma, and High Gamma—in imagined speech decoding. The results indicate that the Beta and Alpha bands are most effective for decoding, offering reliable neural signal representations. By improving the capacity to classify imagined speech and identifying the most effective EEG frequency bands for decoding, these findings lay the groundwork for more refined and accessible BCI applications in the future.

Keywords: Imagined Speech, BCI, EEG, DL Classification, Frequency Bands

1. Introduction

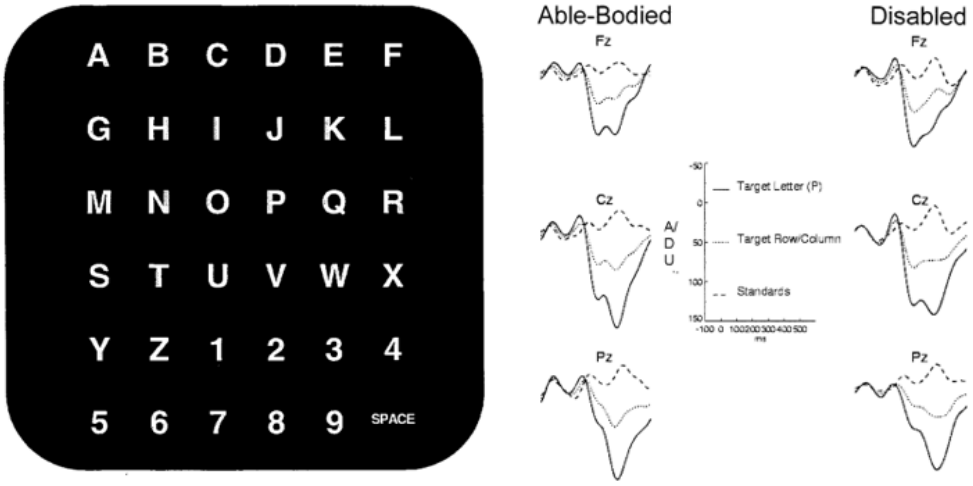
Language is a fundamental tool for communication among individuals (Krishna et al., 2021). However, certain diseases pose significant challenges to communication, rendering it difficult or even impossible. Conditions such as amyotrophic lateral sclerosis (ALS), advanced stages of multiple sclerosis, or cerebrovascular infarctions affecting brainstem regions can disrupt the neural pathways responsible for language production (Coretto et al., 2017). Most of the patients with ALS report that one of their biggest concerns after being diagnosed is the loss of their linguistic ability (Hecht et al., 2002). Some patients can experience a locked-in state, characterized by the inability to voluntarily move specific muscles while maintaining cognitive functions intact. While some residual movements, such as eye or head movements, remain intact for some patients in a locked-in state, these movements are impossible for others (Bauer et al., 1979; Koch-Fager et al., 2019).

Alternative speech systems have been developed to facilitate communication for individuals with limited mobility, commonly relying on residual movements associated with linguistic expression, such as head or eye movements, to enable word or letter selection via a cursor (Koch-Fager et al., 2019). However, these systems face several challenges, such as allowing only a relatively slow production speed, currently averaging around 10 words per minute compared to the natural language production rate of approximately 150 words per minute (Anumanchipalli et al., 2019). One promising avenue for addressing these challenges lies in the synthesis of language through Brain-Computer Interfaces (BCIs). BCIs are devices that can be used to decode non-acoustic biosignals associated with language production, enabling individuals affected by various conditions to communicate through text, text-to-speech (TTS) synthesizers, or cursor control for selection purposes (Wolpaw et al., 2002). Traditional BCIs have focused on motor imagery, event-related potentials (ERPs), and steady-state visually evoked potentials for language synthesis (Farwell and Donchin, 1988; McFarland et al., 2000; Sutter, 1992). Motor imagery BCIs rely on imagining movements with specific body parts, primarily hands, generating discernible patterns in brain waves interpretable by BCIs (Hamedi et al., 2016).

Regarding ERPs, in the study by Donchin et al. (2000), 14 participants (10 healthy and 4 paraplegic) a BCI-based speller was described in which participants were presented with a matrix containing letters, numbers, and the space character. Stimuli were randomly presented to the participants, one after the other, until the one they wanted to select appeared (See Figure 1). The selection was based on the P300 component. Compared to the other stimuli, the flashing of the character of interest elicited a higher positive deflection occurring around 300 ms after stimulus presentation in the central, frontal, and parietal electrodes. However, the time per selection of 5 characters was approximately one minute, making it impractical for clinical applicability. Another strategy involves using steady-state visually evoked potentials.

Figure 1

Stimulus Matrix used for the monitoring. Every 125ms, a row or a column was intensified. Then the ERP P300 was analysed to see if the selected number/letter was in that column (Donchin et al., 2000).



Sutter (1992) demonstrated that steady-state visually evoked potentials could effectively distinguish between different visual stimuli based on occipital EEG patterns, but the technique was limited by its reliance on electroretinographic responses and the complexity of stimulus presentation. These BCI modalities excel in scenarios involving multiple response choices but present several limitations for natural communication. Further innovation is required to provide more intuitive BCIs for natural language communication (Herff et al., 2015). One of the most widely used methodologies to improve these BCIs has been imagined speech decoding (Anumanchipalli, Chartier & Chang, 2019).

1.1. Imagined Speech

Imagined speech, also known as covert speech, involves the silent verbalization of phonemes, words, or sentences without the activation of facial muscles, providing a non-invasive and precise insight into the cognitive processes underlying language production. Some studies have investigated the neural correlates of imagined speech, drawing parallels with the mechanisms of overt speech (LaRocco et al., 2023). Wise et al. (1991) pioneered an investigation into the neuronal activity of key language areas during imagined speech, including Broca's area, Wernicke's area, and the supplementary motor area (SMA). In their Positron Emission Tomography's (PET) study, participants were asked to utter the name of several verbs. Compared to the rest condition, they observed an increase in regional cerebral blood flow (rCBF) in the SMA and Wernicke's areas. According to the authors, the activation of the SMA reflected the motor planning of speech and the activity of the Wernicke's area corresponded to the activation of the speech sounds in order to produce them. Furthermore, Hermes et al. (2015) carried out a study that combined functional magnetic resonance imaging (fMRI) and electrocorticography (ECoG) recordings during a similar silent generation task of verbs. The increased power in gamma activity (65-95 Hz), along with a decrease in theta activity (4-7 Hz) across critical brain regions for speech production, such as the Middle

Temporal Gyrus, the Wernicke's area and the Broca's area, have been interpreted as the use of shared resources to process both types of speech, covert and overt.

However, the relationship between imagined speech and overt speech remains unclear. One of the most explanatory theories regarding the functioning and neuroanatomical correlates of imagined speech is the Efference copy hypothesis within the Internal Forward Model (IFM; Tian & Poeppel, 2010). According to the IFM, the brain can predict the sensory consequences of certain motor actions (i.e., speech production), causing the activation of sensory areas involved in that action. Tian & Poeppel (2013), through five magnetoencephalographic (MEG) studies, found that imagined speech utilizes areas like those of actual speech. During imagined speech, the auditory cortex generates a copy of the anticipated motor signal, defined as an efference copy, activating the auditory cortex in a manner similar to that observed during speech production.

Another notable theory regarding the neural correlates of imagined speech is the Abstraction hypothesis (Cooney et al., 2018). The Abstraction hypothesis proposes that imagined speech can occur without the involvement of an explicit motor plan. This theory suggests that imagined speech operates at a phonemic level, but more flexible views suggest that individuals might also use motor control during the imagination of speech (e.g., Perrone-Bertolotti et al., 2014), depending on the strategy they use. Evidence supporting the Abstraction hypothesis comes from studies demonstrating that, while silently articulated speech exhibits the phonemic similarity effect (i.e., participants make more errors when the phonemes are more similar), this interference effect is not observed in imagined speech (Proix et al., 2022), when the articulatory component is absent. The theory posits that, rather than following the motor neural pathways typically associated with speech production, neural activity during imagined speech is shaped by how each person envisions speech, whether through subarticulation or at a perceptual phonetic level. This highlights the importance of decoding imagined speech not only from motor areas but also from perceptual areas (Proix et al., 2022).

However, recent studies argue that the major difference between overt speech and imagined speech would lie in the degree of brain activation associated with the action, with imagined speech showing reduced cortical activation (Wu et al., 2024). This hypothesis aligns with the findings in the decoding of overt and imagined speech, with overt being much more precise. This characteristic would make the decoding of imagined speech remarkably more complicated through neural signals (Wu et al., 2024). In another study conducted by Lu et al. (2021), it was found that when participants were asked to imagine reciting a poem, the left inferior frontal cortex was activated. According to the authors, this is explained by the brain's preparation for generating the phonological sequence necessary to produce the poem. Similar results have also been observed when asking participants to rhythmically recite numerical counts (Lu et al., 2019). These studies suggest that top-down induction mechanisms help to structure and organize information similarly to real speech. Martinez-Manrique and Vicente (2015) support the idea that the activity of imagining to speak is not a "proper" function of cognition, but would depend on the same cognitive mechanisms involved in overt speech, making the neural networks of both mechanisms the same. However, such studies are often criticized for their ecological validity (Ilina et al., 2017).

In summary, while further research in more ecologically valid conditions is necessary (Ilina et al., 2017), the scientific community largely supports the hypothesis that imagined

speech and overt speech share similar neural mechanisms and areas, particularly within the left hemisphere. However, cortical activation would be reduced during imagined speech compared to overt speech. Studies indicate that key areas such as Broca's area, Wernicke's area, and the SMA are activated during imagined speech, suggesting a shared use of neural resources (Li et al., 2021; Wise et al., 1991). Despite this, the decoding of imagined speech remains more complex, due to its lower precision and probably reduced activation compared to overt speech (Wu et al., 2024). These findings underscore the need for continued exploration into the neural correlates and cognitive processes underlying imagined speech to better understand its relationship with articulated speech.

1.1.1. Procedures and Stimuli used in experiments aimed at decoding Imagined Speech

Research on imagined speech decoding mainly involves the use of the following tasks:

1. Silently reading or repeating phonemes, syllables, words, sentences or pseudowords.
2. Picture naming tasks.
3. Generation of semantically related words (like in the verb generation task, where the participant must generate a verb related to a presented noun or adjective).

The repetition of prompts is a crucial aspect in imagined speech research. Compared to the picture naming tasks or the generation of semantically related words, the repetition of prompts allows the researcher to be more confident regarding the specific response being produced by the participant. Different types of repetition task have been used. For example, Koizumi et al. (2018) and D'Zmura (2009) asked participants to repeat stimuli several times within the same trial, using rhythmic cues to maintain controlled repetition. This method helps in sustaining the imagined speech process (Panachakel & Ramakrishnan, 2021a), but it only improves decoding for the first repetitions, with signal quality deteriorating over time (Panachakel et al., 2020).

Regarding the different ways to present the stimuli to the participants, presenting visual cues on a screen has been the most widely used method. For example, Koizumi et al. (2018) asked participants to imagine themselves saying the names of words displayed on a screen. Nevertheless, the auditory presentation has also been used. Min et al. (2016) used auditory stimuli in a silent repetition task of vowels, comparing the EEG activity between the vowels production condition and a mute condition. Other studies combined the presentation of visual and auditory stimuli (Nguyen et al., 2017), in order to provide greater sensory information and potentially enhance imagery, as what happens in motor imagery tasks (Ikeda et al., 2012). Each presentation method has its advantages and drawbacks. Visual stimuli activate the occipital lobe, an area that is thought to be irrelevant when producing speech. Consequently, in order to exclude the cerebral activity corresponding to the perception of stimuli in their silent speech production tasks, researchers often advocate for removing occipital channels in their analyses (Panachakel & Ramakrishnan, 2021a), resulting in the loss of this information. Auditory stimuli, on the other hand, pose the challenge of separating the cue signature from EEG signals. The combination of both stimuli faces these challenges plus the complexity of the experimental design, but it can yield positive results given that the

common activity between the two types of presentation should correspond to the production of the imagined speech only.

As for the types of prompts used in imagined speech studies, they have evolved over time. Early studies focused on simple word prompts for directions or choices (e.g., "Yes," "No," "Up," "Down"; Sereshkeh et al., 2017), which are crucial for designing BCIs for non-communicative patients. However, recent studies have shifted towards using phonemes and syllables. Suyuncheva et al. (2021) classified Japanese phonemes and syllables with 60% accuracy, although decoding certain syllables and phonemes posed challenges due to the similarity in the way various phonemes are articulated, resulting in similar cerebral activity (Panachakel & Ramakrishnan, 2021a). Panachakel et al. (2021b) achieved 95% accuracy in classifying syllabic categories, while Jahangiri & Sepúlveda (2019) used phonemes to achieve a similar accuracy, identifying key brain areas and emphasizing the importance of gamma waves for classification. LaRocco (2023) even achieved 98% accuracy in classifying English phonemes, demonstrating the potential for developing more intuitive BCIs.

Combining syllables and different types of words could potentially enhance the classification performance of the software used by the BCIs. Nguyen et al. (2017) found improved performance when, in a first step, they used stimulus duration to distinguish between single phonemes vs. short vs. long words. However, further research is required to confirm these findings. To address the challenge of decreasing model accuracy as more words are decoded, Semantic Silent BCIs (SS-BCIs) have been developed. Rekrut et al. (2021) proposed a method to first discriminate the semantic category of a word and then select the specific word within that category, achieving 43.54% accuracy. This approach shows potential for increasing the number of decodable words. However, the use of semantic categories to expand the decodable vocabulary has not been thoroughly explored, primarily due to the difficulty in decoding EEG signals associated with different semantic categories during imagined speech. Thus, one of the primary objectives in this area remains the accurate decoding of EEG signals associated with different semantic categories during imagined speech.

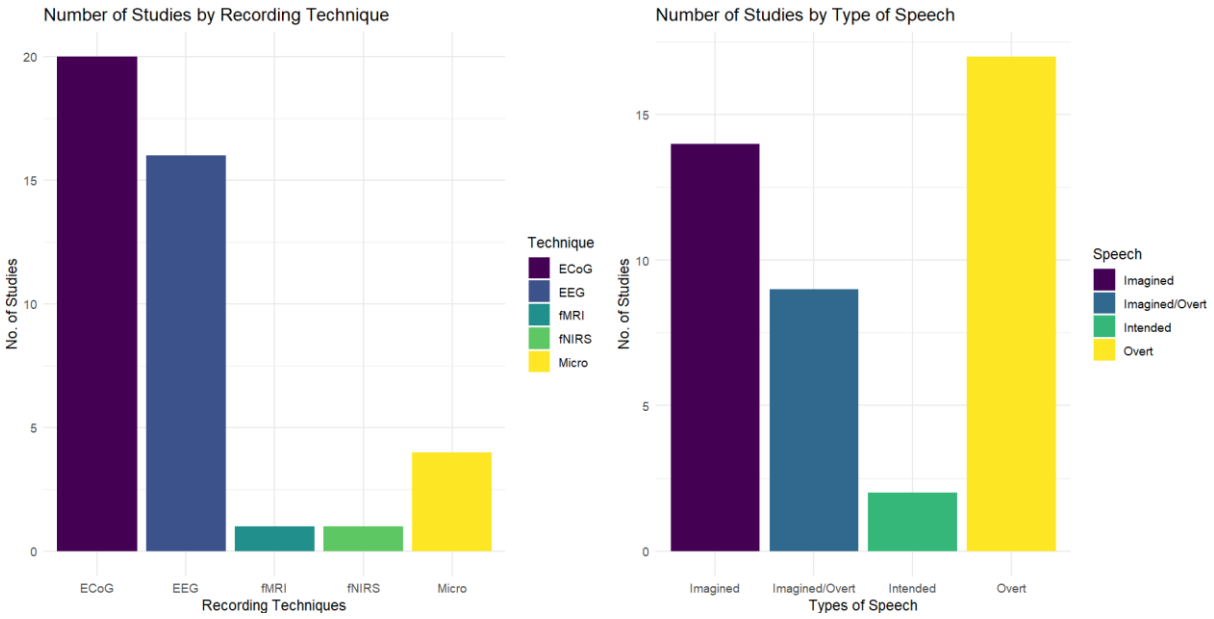
1.2. Electrophysiological Signals

Among the methodologies used for decoding imagined speech, analysis of electrocorticographic (ECoG) signals has emerged as a particularly promising technique (see Figure 2). ECoG entails the invasive placement of electrodes directly onto the brain's surface. This technique offers high spatial and temporal resolution, coupled with a high signal-to-noise ratio, facilitating the examination of various linguistic aspects through recorded brain waves (Gonzalez-Lopez et al., 2020). For instance, Martin et al. (2014) used ECoG recordings from covert and overt speech to decode and reconstruct the spectrotemporal features of speech. When comparing the reconstructed spectrogram from imagined speech data with the original spectrogram (i.e., the spectrotemporal features of the speech recordings for the same items), they observed a marginally significant correlation. Recent attempts have achieved notable improvements, including the decoding of phonemes (Herff et al., 2015), selected words (Chen et al., 2024), and even complete sentences (Anumanchipalli et al., 2019) during overt speech (and articulated speech in the case of Anunchipalli et al., 2019). For example, Anumanchipalli et al. (2019) designed an online decoding BCI for overt and articulated speech. As opposed to offline decoding, this type of BCI decodes and synthesizes

speech while the participant is trying to produce speech. In their study (Anumanchipalli et al., 2019), listeners transcribed the speech output and the transcribed speech was compared to the original sentence produced by the participant. The accuracy of perfectly transcribed sentences in the study ranged from approximately 21% with a word pool of 50 words to 41% with a word pool of 25 words. Willett et al. (2023), with another type of intracranial electrodes implant (microelectrode arrays located on the ventral premotor cortex), demonstrated the feasibility of online decoding of lengthy sentences. Their patient could produce 62 words per minute via the TTS synthetiser and an accuracy of 75% and 89% was achieved with a vocabulary pools of 125,000 and 50 words, respectively, marking a substantial improvement over the previous benchmark.

Figure 2

Distribution of the studies of language decoding by 2020, based in the systematic review of Cooney et al. (2020)

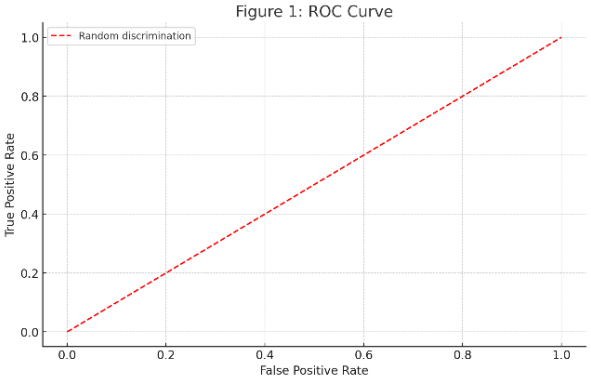


Despite their effectiveness, BCIs based on intracranial electroencephalography remain highly invasive and require extensive training for clinical viability, limiting their practical application (Willett et al., 2023; Dekker et al., 2023). For example, to achieve applicability in terms of accuracy for the BCI developed by Willett et al. (2023), 140 days of daily practice with the neuroprosthesis were needed, and the electrodes implanted surgically also provoked microlesions. Another issue is that, since these studies are conducted primarily for clinical reasons (mainly the resection of areas generating epileptic seizures), the placement of the electrodes is not optimal for speech decoding (Herff et al., 2015). Electroencephalography (EEG), by contrast, offers a less invasive alternative, with electrodes placed on the scalp, typically using a cap, rendering it simpler and more cost-effective to implement (Lopez-Bernal et al., 2022). Nevertheless, EEG-based language decoding has demonstrated limited success. Among the greatest achievements, Suyuncheva et al. (2020) showed that a reduced group of phonemes could be distinguished with EEG signal patterns. Building on this concept, LaRocco et al. (2023) identified the 44 phonemes of the English language by combining amplitude and spatiotemporal features of brain waves, achieving a 98% accuracy rate with machine learning models and an average classification of AUC-ROC (Area Under Curve-Receiver

Operating Characteristic) of 0.68 ± 0.002 across all phonemes, with an AUC of 1 indicating perfect classification and an AUC of 0.5 indicating a random classification (see Figure 3), meaning that the model developed by LaRocco et al. (2023) had good discriminative capacity.

Figure 3

An AUC-ROC representation for a random classification discrimination.



Note. The True Positives rate is placed on the x-axis, while the False Positives rate is placed on the y-axis, and a diagonal line is plotted, representing the discrimination by chance.

Currently, four databases with EEG data are available, two in English (LaRocco et al., 2023; Zhao & Rudzicz, 2015), one in Dutch (DAIS; Dekker et al., 2023), and another in Spanish (Coretto et al., 2017). These databases were created with the aim of establishing more effective decoding methods, by allowing the researchers to train different types of decoding models on the same data sets. The heterogeneous use of collection techniques within the community makes it difficult to estimate the most effective methods (Shah et al., 2022). However, these databases lack internal reliability (Dekker et al., 2023), like the lack of any kind of testing to ensure that imagined speech was used. Furthermore, two of the databases (Coretto et al., 2017; LaRocco et al., 2023) only recorded data in imagined language, rendering the comparison between real and imagined speech conditions impossible. Lastly, although the work by Coretto et al. (2017) is an open database, permission from the authors is required to access it and the authors never responded to our requests for access. Interestingly, the researchers who created these databases also used the same data to test their decoding models (e.g., see the description above of the results of LaRocco et al., 2023). TheThe KARA ONE study, developed by Zhao & Rudzicz (2014), utilized both EEG measurements and facial and audio recordings. Both real speech and imagined speech were measured, but the accuracy of neither procedure was mentioned. Regarding overall accuracy within the EEG modality, it was very poor, correctly classifying between brain electrical activity associated with a vowel or consonant only 18% of the time. Only the combined use of facial stimuli recordings and EEG achieved classification accuracy above chance. In the study by Dekker et al. (2023), two main comparisons were made. The first comparison tested whether their models could classify three conditions: rest, overt, and covert. The accuracy of their models averaged 70.6% (± 4.4%), which is higher than the 33% chance level. They also performed vowel classification in the covert condition, with an accuracy of 19.6% (± 2.1%), which is not above chance (20%). Lastly, the database provided by Coretto et al. (2017) is the only one in Spanish. However, the EEG model used is of low density (18 electrodes), and the results obtained in vowel classification were very poor, with an accuracy of 20%, the same as chance.

1.2.1. EEG Acquisition Protocol

Most imagined speech studies with EEG have utilized 64-electrode systems with a sampling rate of 1 kHz (Panachakel and Ramakrishnan, 2021). However, although

most researchers have employed high-density EEG systems, Wang et al. (2013) argued that an interesting approach could be to solely focus on the Wernicke and Broca areas, since the EEG channels over these areas receive the most significant data for classifying imagined speech. Wang et al. (2014), Nguyen et al. (2017), and Zhao and Rudzicz (2015) followed this reasoning. Another reason to use fewer electrodes stems from practical implications. Commercial EEG devices with fewer channels have shown relatively good decoding results, are more cost-effective, and have shorter setup and maintenance times than high-density EEG systems (LaRocco et al., 2023), making these devices more practical for commercial applications.

However, the utilization of EEGs solely focusing on the Wernicke and Broca areas is not yet excessively robust, which might be due to a loss of information. Neuroimaging studies, such as that conducted by Newman et al. (2010), demonstrated that, aside from the Wernicke and Broca areas, other areas in the temporal lobe are also related to information processing when producing speech. Additionally, the use of high-density EEG devices allows for better Independent Component Analysis (ICA; Klug & Gramann, 2020). ICA permits the decomposition of the original signal components, subsequently enabling the selection of significant features for analysis while discarding noise-related components (Stone, 2002).

1.2.2. EEG Frequency Bands Implicated in Imagined Speech Decoding

EEG signals can be decomposed into 5 main frequency bands: delta, theta, alpha, beta and gamma. These frequencies go from 0.5 Hz till 150 Hz or more. Three frequency bands have been mainly used in the decoding of covert or overt speech: alpha, beta and gamma (Cooney et al., 2020; Hossein et al., 2023).

- Gamma band (30 – 150 Hz): brain waves frequencies in this band, particularly those in the high gamma range (80-150 Hz), are associated with overt speech production, but these high frequencies are also often associated with artifacts in EEG signals, which can complicate their use in decoding imagined speech processes (Koizumi et al., 2018; Lopez-Bernal et al., 2022).
- Beta band (12 – 30 Hz): frequencies in the beta range are linked with demanding cognitive tasks, such as decision-making and problem-solving tasks. This frequency range is also associated with speech production and auditory speech perception tasks (Hossain et al., 2023; Lopez-Bernal et al., 2022).
- Alpha band (8 - 12 Hz): Proix et al. (2024) demonstrated that brain waves frequencies in the alpha range play a role in speech encoding and that they can be useful to decode imagined speech.

The study of frequency bands associated with the development of brain-computer interfaces (BCIs) is of great importance. These frequency bands seem to play different roles in the processing of covert or overt speech. Understanding the involvement of specific bands in the different speech processes would enable the fine-tuning of BCI algorithms, thereby enhancing their precision and accuracy (Lopez-Bernal et al., 2022).

1.3. Feature Extraction and Classification

1.3.1. Feature Extraction

Feature extraction is the process of transforming raw EEG signals into a set of relevant and useful features that capture the most significant information to enhance the performance of predictive models for speech decoding (Panachakel & Ramakrishnan, 2021). Feature extraction can be conducted in three primary domains: time, space, and frequency (Lopez-Bernal et al., 2022). Early studies in imagined speech decoding, such as Zhao & Rudznick (2015), focused on extracting time-domain characteristics like mean, variance, and skewness. However, current approaches emphasize both the spatial and frequency domains.

In the frequency domain, commonly used methods for extracting critical signal features include Mel Frequency Cepstral Coefficients (MFCC), Fast Fourier Transform (FFT; Figure 4a), and Wavelet Transform (WT) (Lopez-Bernal et al., 2022). Notably, the Wavelet Transform has gained significant attention due to its ability to provide a multi-resolution analysis of signals. This technique is especially valuable for EEG analysis as it allows the decomposition of signals into components of different scales, capturing both time and frequency information simultaneously. The Wavelet Transform's capability to highlight transient features in EEG data makes it particularly suitable for identifying the complex dynamics associated with imagined speech (Cooney et al., 2020; Shah et al., 2022). Given these advantages, our study will employ the Wavelet Transform for feature extraction, aiming to leverage its precision in detecting subtle yet significant variations in the EEG signals related to different speech conditions.

In the spatial domain, Common Spatial Patterns (CSP; Figure 4b) remains a prominent method for analyzing the spatial distribution of neural activity. CSP is particularly effective in identifying areas involved in imagined speech tasks by finding data projections that best separate different classes. The CSP algorithm transforms EEG data to maximize the variance for one class (e.g., silent state) while minimizing it for another (e.g., imagined speech), enhancing the detection of spatial patterns in EEG signals and enabling precise classification. However, a notable limitation of CSP is that, although it can be applied to multiclass classification problems, its binary classification nature restricts its applicability in studies requiring distinctions among more than two classes (Schirrmeister et al., 2017).

Figure 4

a) The Fast Fourier Transform (FFT) algorithm converts the EEG signal from the time domain to the frequency domain. This step is typically performed before applying additional frequency domain extraction techniques. b) The Common Spatial Pattern (CSP) algorithm is applied to maximize the variance of class A and minimize the variance of class B by optimizing the spatial pattern vectors W (Vectors of the spatial patterns).

a) $X[k] = \sum_{n=0}^{N-1} x\,[n] \cdot e^{-j\frac{2\pi}{N}kn}$ b) $W = U^T \Lambda^{-\frac{1}{2}} U^T C_1 U \Lambda^{-\frac{1}{2}} U$

1.3.2. Classification Algorithms

In order to classify features extracted from EEG signals, researchers have employed different types of Machine Learning (ML) algorithms (Shah et al., 2022). These are statistical algorithms designed to recognize patterns within data. In the context of

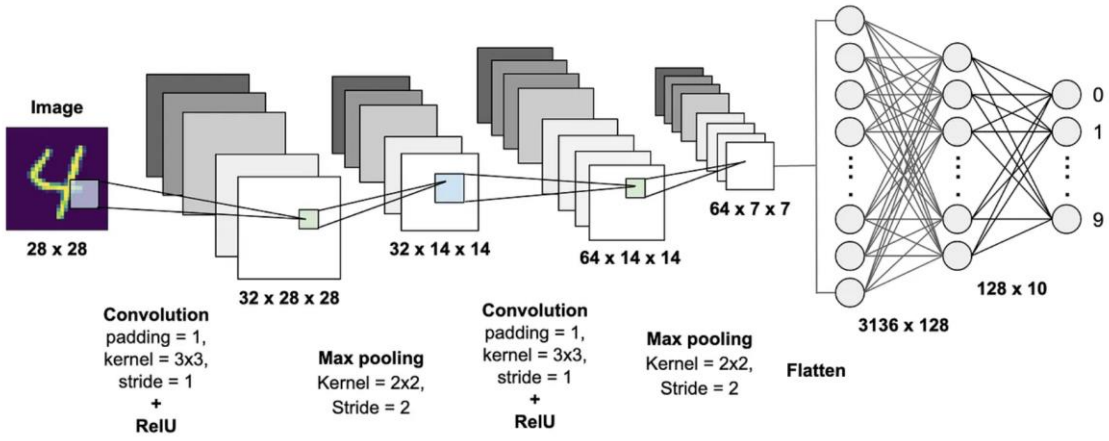
BCIs for speech and language decoding, ML techniques learn to identify specific patterns in the EEG signals associated with different types of speech stimuli (e.g., phones or words). This capability allows them to make accurate predictions and classifications of EEG data (Cooney et al., 2018).

Machine Learning (ML) encompasses a range of techniques for pattern recognition. Common ML models include Support Vector Machines (SVM), Decision Trees (DT), and Linear Discriminant Analysis (LDA) (Lopez-Bernal et al., 2022). These models are trained using data such as epochs of imagined speech for various words (LeCun, 2015). While these techniques have been successful, their reliance on manual feature engineering can limit their performance and scalability. To solve the above problem, Deep Learning (DL) was conceived (LeCun, 2015), which is a subfield of ML employing artificial neural networks as predictive models for regression and classification.

DL algorithms, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), offer several advantages over traditional ML methods. These models are noted for their ability to create hierarchical representations from EEG data. In particular, CNNs are models inspired by the connectivity patterns in the visual cortex that, are highly effective in identifying spatial patterns within EEG signals and hierarchizing them, thanks to their convolutional layers which are systematically able to categorize the importance of the given features (Roy et al., 2019; see Figure 5).

Figure 5

Diagram of the Convolutional Neural Network Architecture*. Input layer receives raw image data and passed through convolutional layers. These layers apply multiple filters to extract features such as edges and textures, producing feature maps. Pooling layers are employed to reduce the dimensionality of the feature maps, retaining the most important information while reducing computational load. Finally, the flattened output combine the extracted features and produce the final classification output.*



These networks excel at detecting subtle variations and spatial configurations, making them particularly useful for classifying imagined speech patterns (Tamm et al., 2020). On the other hand, RNNs are specialized in handling sequential data, which is crucial for capturing temporal dependencies in EEG signals (Rumelhart et al., 1986). RNNs update their internal state with each time step, which allows them to retain information from previous states and adaptively process new data. This capability is particularly valuable

for decoding EEG signals related to imagined speech, where temporal patterns are significant (Roy et al., 2019).

However, RNNs face challenges such as gradient explosion, which occurs when the weights of the network become excessively large, leading to numerical instability. To address this, advanced RNN architectures like Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) have been developed. These models incorporate mechanisms to regulate the magnitude of weights, facilitating effective learning over long sequences (Agarwal & Kumar, 2022).

CNNs are designed to handle spatial patterns through layers of convolution and pooling. The initial layers apply convolution to capture spatial features, which resemble band-pass filtering and CSP. The resulting feature maps highlight distinctive characteristics of brain waves. Subsequent layers, including activation functions and pooling, further refine these features. For instance, while CNNs often use functions like ReLU for non-linearity, some architectures might include operations like squaring to enhance feature discrimination (Oh et al., 2019). Batch normalization and dropout techniques are applied to prevent overfitting, ensuring robust model performance (Lawhern et al., 2018; Cooney et al., 2020).

The integration of CNNs and RNNs presents a promising approach to advancing imagined speech decoding. By leveraging the strengths of both models—CNNs for spatial analysis and RNNs for temporal processing—researchers can develop comprehensive DL architectures that capture both spatial and temporal aspects of EEG signals (Tang et al., 2015; Roy et al., 2018). This integration enhances the depth and complexity of decoding models, potentially leading to improved communication interfaces for individuals with speech impairments (Cooney et al., 2020). The shift towards DL models opens new research avenues, offering improved feature extraction and classification capabilities compared to traditional ML methods. While DL has demonstrated significant promise in enhancing our understanding of the neural mechanisms underlying imagined speech production, further research is needed to identify the most effective DL techniques for EEG signal processing and classification (Cooney et al., 2020).

1.4. Main goals

In this study, we aim to evaluate the efficiency of a novel classification algorithm that leverages modern computational techniques, based on the architecture proposed by Roy et al. (2018). While we started with their original design, we have modified it specifically for the classification of imagined speech, integrating parallel deep learning convolutional networks with advanced recurrent elements like the Gated Recurrent Unit (GRU). Our main objectives are threefold:

1. **Classification of Speech Types**: We will test our model's ability to classify EEG data into three conditions: imagined speech, real speech, and silence, following the methodology of Zhao & Rudznick (2014). If successful, this will demonstrate the model's discriminative capability.
2. **Vowel and Semantic Category Decoding**: Building on previous work, we will extend the model’s application to decode Spanish vowels (/a/, /e/, /i/, /o/,

and /u/) and words from six semantic categories (kitchen utensils, animals, food, clothing, musical instruments, and body parts).

3. **Frequency Band Analysis for Decoding**: We will investigate which frequency bands contribute most to decoding imagined speech by analysing Morlet wavelets across different frequency bands in EEG recordings (1-120 Hz).

1.5. Hypothesis

Three hypotheses were proposed to address the main goals described above.

First hypothesis: Classification of the types of speech production

- **H1:** Our model will be able to classify Rest, Imagined speech and Overt speech above chance level (>33%). Furthermore, fulfilling this hypothesis provides a manipulation check to confirm that the participant was performing imagined speech by observing differences with silent trials (the Rest condition).
    - **H1.1:** Differences between the three conditions will be found across all analysed frequency bands: Alpha, Beta, Gamma, and High Gamma.
    - **H1.0:** Differences between the three conditions will be found only on certain frequency bands.
- **H0:** Our model will not be able to decode the types of speech between the three conditions above chance level.

Second hypothesis: Classification of the Spanish vowels

- **H1:** Our model will be able to classify the 5 vowels of the Spanish language with accuracy above chance level (>20%) in the Imagined speech condition.
    - **H1.1:** Within the frequency bands, vowel classification in the Imagined speech condition will be significantly higher when using the Beta frequency band.
    - **H1.0:** Vowel classification in the Imagined speech condition will not be slightly higher in Beta band in comparison with other bands.
- **H0:** Our model will not be able to significantly classify the 5 vowels of the Spanish language above chance level in the Imagined speech condition.

Third Hypothesis: Classification of words according to their Semantic Category

- **H1:** Our model will be able to decode significantly above chance level (>16.6%) the semantic category of the words in imagined speech among 6 semantic categories: Clothing, Kitchen Utensils, Animals, Body Parts, Instruments, and Food.
    - **H1.1:** In the imagined speech condition, our model will be able to decode significantly above chance level (>16.6%) the semantic category of the words among 6 semantic categories: Clothing, Kitchen Utensils, Animals, Body Parts, Instruments, and Food.

- **H1.0:** Within the frequency bands, semantic category classification in the Imagined speech condition will be significantly higher when using the Beta frequency band.

- **H0:** Our model will not be able to significantly decode the different semantic categories above chance level.

2. Methods

2.1. Participants

A total of 14 participants took part in the experiment (10 females and 4 males; mean age 23.36 years, SD = 3.77). Participant 7 was excluded due to noisy signal (in the preprocessing process, it was impossible to detect the speech onset signals). According to the Edinburgh Inventory (Oldfield, 1971) to assess handedness, all of them were right-handed. All participants reported being native Spanish speakers, having normal or corrected-to-normal vision, and to not have any history of neurological disorder. Informed consent was obtained from all the participants before the experiment began. Participants were compensated for their time, receiving 10€ per hour. At the end of the experiment, each participant was invited to participate in a new session. If the participant accepted, a minimum of 24 hours separated the two sessions. Each participant was informed that she/he could carry out as many sessions as she/he wanted, until a maximum of ten sessions. As a result, two participants carried out three sessions and four participants carried out two sessions (22 sessions in total). In order to have a similar amount of data for each participant, only the data of the first session were used for the analyses. The data of the other sessions will be used in further studies to investigate the inter-sessions variability. The ethical committee of the University of Granada approved the experiment (4210/CEIH/2024).

2.2. Stimuli, materials and software for stimuli presentation

The following stimuli were presented to the participants:

Syllables: A total of 95 CV (consonant-vowel) syllables were used, one for each of the 19 Spanish consonants combined with one of each of the five Spanish vowels (see Acknowledgments for the Syllable list). The syllables were centrally presented as a written text in white hue (Arial font, size 16) on a black background or as auditory stimuli via the recordings of a native female voice. Audacity 3.6.0 (with an Audio-Technica AT2020USB-XP microphone) was used to record and edit the auditory stimuli (track(s): mono; coding: 32 bits; frequency: 44100 Hz).

Words: To decode different semantic categories, 60 words from six semantic categories (10 per category) were used. The semantic categories were: 1- Kitchen utensils; 2- Body parts; 3- Food; 4- Musical instruments; 5- Clothes; 6- Animals. In each semantic category, half of the words were high-frequency words (>10/1,000,000) and the other half were low-frequency words (<5/1,000,000) according to the CORPES XXI open database from the Royal Spanish Academy (Corpus del Español del Siglo XXI [CORPES XXI], October 2013). The number of syllables and syllable structure complexity were similar between high- and low-frequency words (see Table 1). The characteristics of the written and auditory presentations were the same as those for the consonant trials. In addition, pictures illustrating each word were presented to refer to the participants. Most of the pictures were sourced from different normalized picture

databases (Adlington et al., 2009; Brodeur et al., 2010; 2014; Moreno-Martínez & Montoro, 2012; Saryazdi et al., 2018). For a few pictures, normalized items were unavailable, and these were downloaded from the Internet under the Creative Commons license. Adobe Photoshop software (version 2017.0.1) was used to adjust the format of the pictures to match those in Brodeur et al.'s (2010; 2014) studies, i.e., a 2000 x 2000 pixels size with a white background (see Acknowledgements for the words list). For presentation, pictures were resized to 1000 x 1000 pixels.

Table 1

The number of syllables according to their structure composing the high- and low-frequency words.

	Word frequency	
Syllable structure	High	Low
CCV	10	8
CCVC	0	2
CV	52	49
CVC	12	9
CVCC	0	1
CVV	3	5
V	2	1
VC	1	5
Total:	80	80

Pseudowords: Additionally, 30 pseudowords were presented both as written text and as auditory stimuli. Pseudowords were created using the same syllables composing the 60 words, maintaining the same syllable order (e.g., the "wi" syllable is in the second place in the word "kiwi" and pseudoword "awi"), but changing the combinations between the syllables. This strategy ensured similar syllable complexities and frequencies between words and pseudowords.

Other stimuli: A plus sign "+" was used as the fixation point. A left and a right brackets ([ ]) with a 22 characters spacing between them were used during the stimulus presentation (except for pictures) and until the end of the response. Three asterisks ("***") indicated the speech onset. The auditory or the written word "silencio", or an iconic picture showing a person shushing where used to indicate to not produce any speech in that trial. The characteristics for the presentation were the same as those for the other stimuli.

Auditory Presentation: Auditory stimuli were presented through two loudspeakers located next to the left and right of the screen.

Software: The PsychoPy program (Peirce et al., 2019) was utilized for programming the experiment and for stimuli presentation.

2.3. Procedure

At the beginning of the experiment, participants were shown the informed consent form. They were informed that they could ask any questions related to the experiment to the experimenter. They were also informed that they could leave the experiment at any moment without any penalty. During this time, the participant's head was measured, and

the EEG cap was fitted. Fitting the cap and applying the conductive gel took between 20 to 40 minutes.

Afterwards, the syllables block began, participants were instructed that they would alternate between an imagined speech task (covert speech) and a speaking aloud task (overt speech) every 20 trials. Specific instructions were provided at the beginning of each set of 20 trials, and participants could take a short pause during this time. Before the experimental trials, two practice blocks of six trials each were presented to the participants (with different syllables), one for the imagined speech task and another for the speaking aloud task. In the experimental blocks, each syllable was presented once as a written stimulus and once as an auditory stimulus in each task: imagined speech or speaking aloud. Additionally, 20 trials were included, 10 in each task, where the participant was instructed to not speak or to not imagine speaking (10 trials with the written and 10 trials with the auditory instructions). The total number of experimental trials in the syllable block was 400.

Following the syllables block, the words-pseudowords block was presented. Again, participants were instructed that the imagined speech and the speaking aloud tasks would alternate every 20 trials. There was also a practice block for the imagined speech task and another for the speaking aloud task, each block comprising eleven trials (with words and pseudowords that differed from the experimental trials). In each task (imagined speech or speaking aloud) of the experimental blocks, each word was presented once pictorially, once as a written text, and once as an auditory stimulus, and each pseudoword was presented once as a written text and once as an auditory stimulus. Additionally, five silent trials were included in each condition (with the word "silencio" – "silence" being presented as written text or auditorily). During the silent trial, participants were asked to remain silent. Additionally, in each task (imagined speech or speaking aloud), 15 trials (5 pictures, 5 written and 5 auditory trials) instructed the participant to not speak. The total number of experimental trials in the words-pseudowords block was 510.

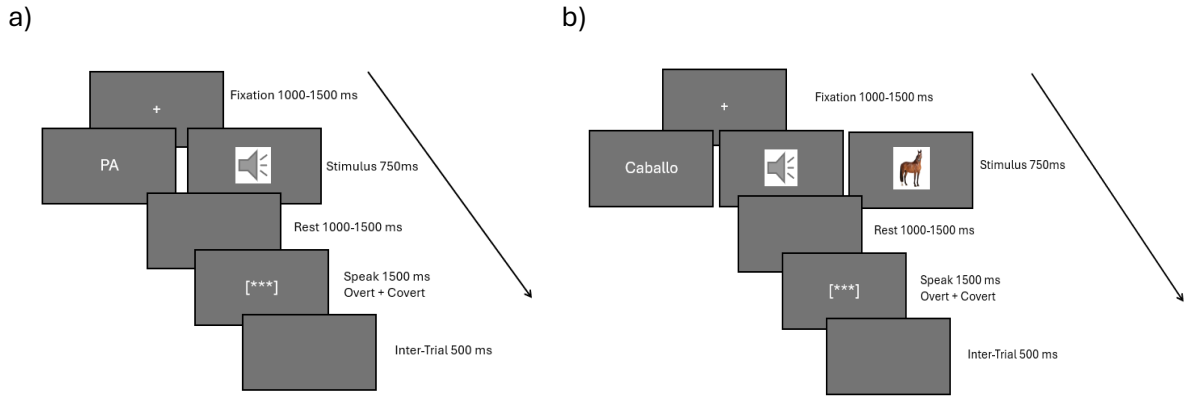
In the syllables and the words-pseudowords blocks, all the stimuli were presented in random order within each task. The practice blocks were presented only during the first session.

The trial structure in the syllables block was as follows (see Figure 6a). First, a fixation cross was presented for a randomly varying time between 1000 and 1500 ms. Then, the brackets and the syllable or the silent word were presented (as a written text or as an auditory stimulus) for 750 ms. It was followed by a screen where only the brackets were presented during a period that varied randomly from 1000 to 1500 ms. Afterwards, the brackets plus the three asterisks indicating the onset of speech were presented for 1500 ms. Finally, an inter trial interval was presented for 500 ms. The trial structure in the words-pseudowords block was similar to that of the syllables (see Figure 6b), with the following exceptions: the word or pseudoword was presented for 1000 ms (instead of 750 ms), the words (but not the pseudowords) were also presented as a picture, and the three asterisks indicating the onset of speech were presented for 2000 ms (instead of 1500 ms).

The entire session lasted around one hour and a half.

Figure 6

a) Order of presentation of the stimuli in the syllables block. b) Order of presentation of the stimuli in the words-pseudowords block



2.4. EEG and Voice Recordings

The EEG signal was recorded using a 64-channel system mounted on a cap (actiCAP snap, Brain Products) and the Brain Vision Recorder Software (version 1.20.0601) was used to calibrate the electrode montage. The EEG signal was amplified thanks to a actiCHamp amplifier (Brain Products GmbH, Munich, Germany). Electrode placement followed the international 10-20 system. Impedances were kept below 10 kΩ in areas of interest, with an attempt to lower impedances in other electrodes, setting a limit of 20 kΩ. Although the supplier recommends reducing impedance to 5 kΩ, time constraints made this unfeasible. The signal was digitized at a sampling rate of 1000 Hz and referenced with electrode FCz. Eye movement activity was monitored using two electrooculogram (EOG) electrodes.

In addition to the EEG, the participant's voice was also recorded using the AudioCapture software from the Lab Streaming Layer (LSL) library (Copyright (c) 2021 Christian Kothe, Tristan Stenner) at 44.1 kHz and 16 bits per sample. This was accomplished using an Audio-Technica AT2020USB-XP microphone placed about 20 cm from the participant’s mouth. This microphone is a cardioid condenser, which allows for sound collection from a single direction. This design helps reduce extraneous noises and reverberations from the room where the task is performed. To reduce popping sounds, a pop filter was placed between the microphone and the participant at 3 cm from the microphone.

EEG, audio recordings and task-specific markers were synchronized using Lab Streaming Layer (Kothe et al., in press) with the LabRecorder (Copyright (c) 2012 Christian Kothe) extension. Additionally, BrainVision LSL viewer (Pfurtscheller & Neuper, 2001) was used to visualize online recordings.

2.5. EEG Analysis and Preprocessing

For the analysis and subsequent preprocessing of the data, the MNE library in Python (Gramfort et al., 2013) was employed, utilizing the Visual Studio Code source code editor. The primary aim of the analysis and preprocessing was to filter the data to

facilitate the model's extraction of the most pertinent features from the waveforms (See Table 2).

Table 2

Number of stimuli used to obtain the Morlet wavelets per participant.

Type of Stimuli	Stimuli	Stimuli Quantity
Production	Real	200
	Imagined	200
	Silence	20
Vowels	a	38
	e	38
	i	38
	o	38
	u	38
Semantic	Body Parts	27
	Kitchen	27
	Food	27
	Animals	27
	Clothes	27
	Instruments	27

The data was divided into epochs of 1500 ms, commencing at the beginning of the speech onset signal. The preprocessing and artifact removal procedure followed five key steps: 1- A notch filter was applied at 50 Hz and its first harmonic (100 Hz); 2- Based on Time-Frequency analysis to identify optimal frequencies for decoding, a band-pass filter ranging from 1 to 120 Hz was implemented; 3- Independent Component Analysis (ICA; See Supplementary Material Figure 1) was conducted to eliminate artifacts associated with eye blinks; 4- Morlet wavelets were extracted by performing a Fast Fourier Transform (FFT) on the frequency bands alpha (8 - 12 Hz), beta (13 - 30 Hz), gamma (30 - 50 Hz), high gamma (50 - 120 Hz) and general band (1 -120 Hz) for the conditions corresponding to the types of speech (Rest, Imagined speech and Overt speech), the vowels (5 vowels plus the silence trials when producing imagined speech speech) and the six semantic categories (when producing imagined speech); 6- After obtaining all the Morlet wavelet analyses, the data were segmented into the aforementioned categories (See Supplementary Material Figure 2).

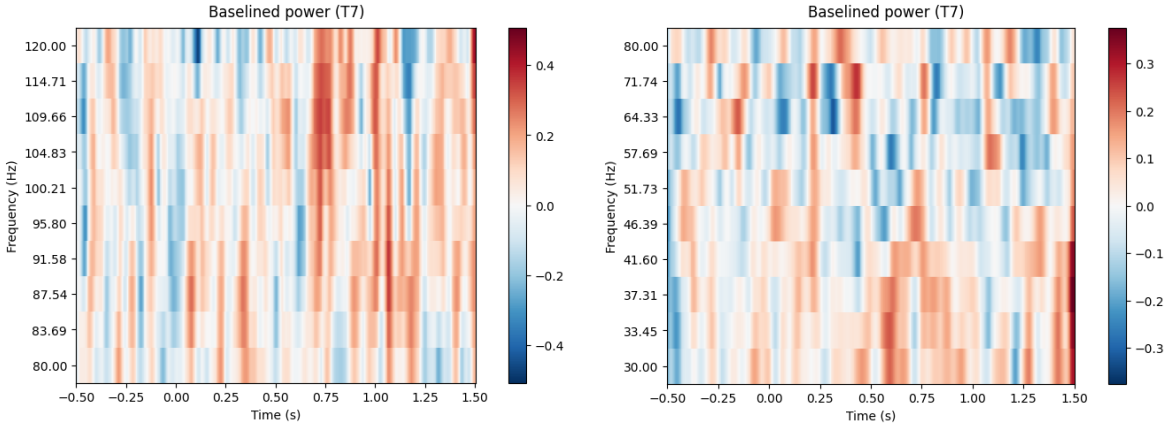
2.5.1. Time-Frecuency Analysis and Morlet Wavelets

Firstly, the frequency bands to be used were established: Alpha, Beta, Gamma, High Gamma, and a general band that encompassed all the bands (1-120 Hz). Subsequently, for each vowel and frequency band, the Morlet wavelet was extracted, with the number of cycles appropriate to the Hz of the frequency band being utilized (see Figure 7). The Morlet wavelet is one of the best time-frequency analyses for decoding and subsequently classifying non-stationary waveforms, where spectral densities change easily (Min et al., 2016). The Morlet wavelet was extracted for each channel,

although it is also possible to perform this extraction through a cross-variance matrix of the EEG channels (Panachakel et al., 2021; Lua et al., 2022).

Figure 7

Distribution of Morlet wavelet power in electrode T7 for the High Gamma (left) and Gamma (right) frequency bands in the pilot participant for the vowel 'a'.



2.6. *Statistical Analyses*

To the hypothesis about the differences brain oscillations in the decoding of imagined speech, we conducted a Wilcoxon signed-rank test to compare the classification accuracy between the following conditions: Rest, Imagined speech and Overt speech across the Alpha, Beta, Gamma, High Gamma and General frequency bands. We applied the same procedure to compare classification accuracy in the vowels classification and the semantic categories classification.

Figure 8

The formula for Min-Max scaling

$$X` = \frac{X - X_{min}}{X_{max} - X_{minx}}$$

Additionally, to assess the model's performance against chance, we used the Wilcoxon signed-rank test to compare the accuracy for the types speech classification, vowels classification, and semantic categories classification relative to chance levels (33%, 20%, and 17%, respectively).

The typical 80-20% data split was applied (LeCun, 2015), where 80% of the data was allocated for model training, while 20% were reserved for assessing the generalization capability, using them as test data.

2.7. *Classification Analysis*

The data was subjected to normalization using min-max scaling (See Figure 8), adjusting the values to the [1, -1] scale, allowing all values to be on a common scale. This enabled our models to converge more quickly and optimization processes to be performed more efficiently. Once the data normalization was completed, a variant of the Convolutional Neural Network (CNN) with characteristics of a Recurrent Neural Network

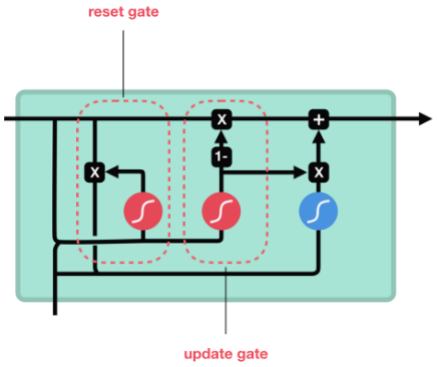
(RNN) was applied. For this model, several layers of Gated Recurrent Units (GRUs) were used.

2.7.1. Convolutional Neural Network (CNN) + Gated Recurrent Unit

The architecture of our model consists of three parallel 1D convolutional (Conv1D) layers, which sequentially apply convolutional operations to capture both spatial and temporal features from the data (Szegedy et al., 2015; Roy et al., 2019). This structure was derived from the deep learning algorithms used by Roy et al. (2019) for the decoding of electroencephalographic signals in patients with schizophrenia and adapted for the decoding of imagined speech. These layers are followed by several densely connected GRU layers (See Figure 9) through feed-forward connections. The use of multiple parallel Conv1D layers allows the model to explore various filter sizes and extract diverse types of features that a single filter type might miss. Additionally, the incorporation of densely connected GRU layers addresses the gradient explosion and degradation issues commonly associated with RNNs. As of this writing, no similar network has been reported in the literature for decoding imagined speech from EEG data.

Figure 9

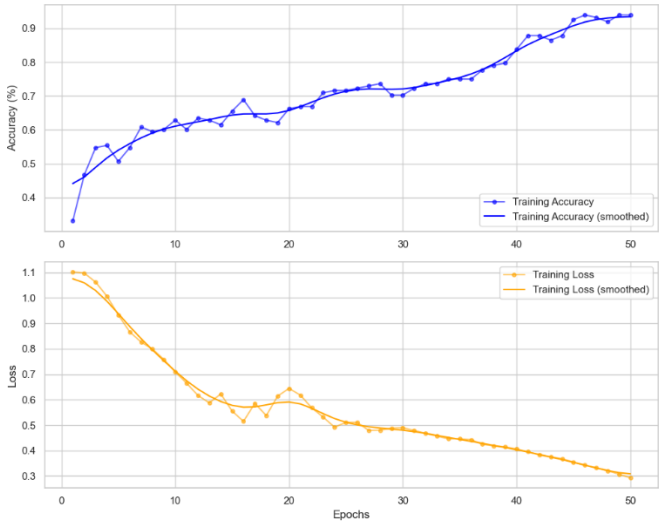
Pictorial representation of a Gated Recurrent Unit



For model training and parameter tuning, a learning rate of 0.001 and the Adam optimizer were employed. These settings were chosen due to their widespread use and effectiveness in training models for decoding Imagined speech from EEG data (Abdulghani et al., 2023; See Figure 10 for a representation of the training accuracy and training loss of the model across different epochs for linguistic production discrimination between Rest, Imagined and Real speech for the participant number 8).

Figure 10

Model training process across different epochs in general band (1-120 Hz) for linguistic production discrimination: Imagined vs. Rest vs. Real in the participant number 8.



3. Results

3.1. Differences in the types of speech production

The initial results of our investigation focused on distinguishing between the three types of linguistic production (Rest, Imagined speech and Overt speech), as this step is crucial for determining whether the model confuses Imagined speech with Rest (Supplementary Material Figure 3). This approach is relatively novel, as pre-decoding manipulation checks for vowels/words/phrases are uncommon in the literature (Lopez-Bernal et al., 2022). For each participant, the EEG data collected corresponded to 200 trials for the Imagined speech condition, 200 trials for the Overt speech condition and 20 trials for the Rest condition (All of them per participant). Due to the significant imbalance in the number of trials between the Overt and Imagined Speech conditions compared to the Silent Speech condition, proportional weighting was applied to ensure balance in the analysis. The model's accuracies for individual participants ranged from 60% to 94.67% using the general band and these classification results were significantly higher than the 33% of chance accuracy level (see Table 2; Figure X). These results provide a basis for believing that our model can effectively classify data within the context of imagined speech. The accuracies observed are consistent with findings reported in the literature, where Convolutional Neural Networks (CNNs) have demonstrated high precision in decoding various datasets (Cooney et al., 2020).

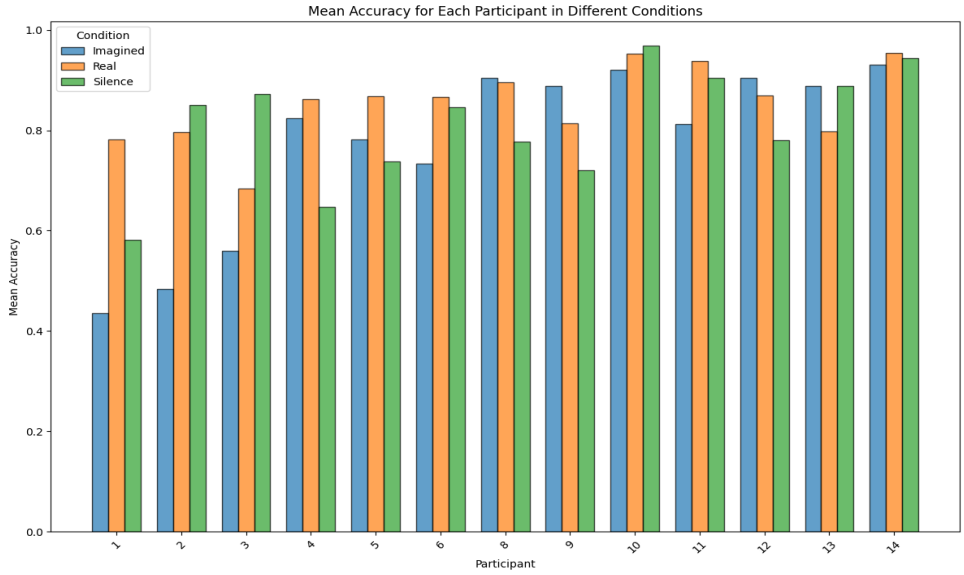
Table 2

Accuracy of the model compared to the chance level in classifying Rest vs. Imagined speech vs. Overt speech on general frequency band using the Wilcoxon signed-rank test

Accuracy Median (%)	*Z*-score	*p*-value	Reference value
91%	26.95	< .001	33%

Figure 11

Mean classification accuracy for each participant across Imagined Speech, Overt Speech, and Rest conditions using the general band (1-120 Hz).



It was also hypothesized that the four frequency bands (Alpha, Beta, Gamma and High Gamma) would provide enough information to significantly discriminate the EEG data corresponding to each type of linguistic production. Accordingly, the classification accuracies were statistically above the chance level for each frequency band (see Table 3; Supplementary Material Figure 4). These findings confirm that our model does not confuse imagined speech with silence. With this validation in place, we can now proceed to evaluate whether our model is capable of decoding vowels and semantic categories within imagined speech.

Table 3

Discrimination accuracy compared to chance level in classifying Rest vs. Imagined speech vs. Overt speech across the different frequency bands using the Wilcoxon signed-rank test

Frequency bands	Accuracy Median (%)	*Z*-scores	*p*-values	**Reference value**
Alpha	90%	32.35	< .001	**33%**
Beta	92%	38.22	< .001	**33%**
Gamma	76%	10.54	< .001	**33%**
High Gamma	76%	10.79	< .001	**33%**

However, when comparing bands between them, significant differences in accuracy were observed between all of them, except for the comparison between the Alpha and the Beta bands, and the comparison between the Gamma and the High Gamma bands (see Table 4). These results underscore the critical role of the Alpha and Beta bands in the overall decoding performance of imagined speech.

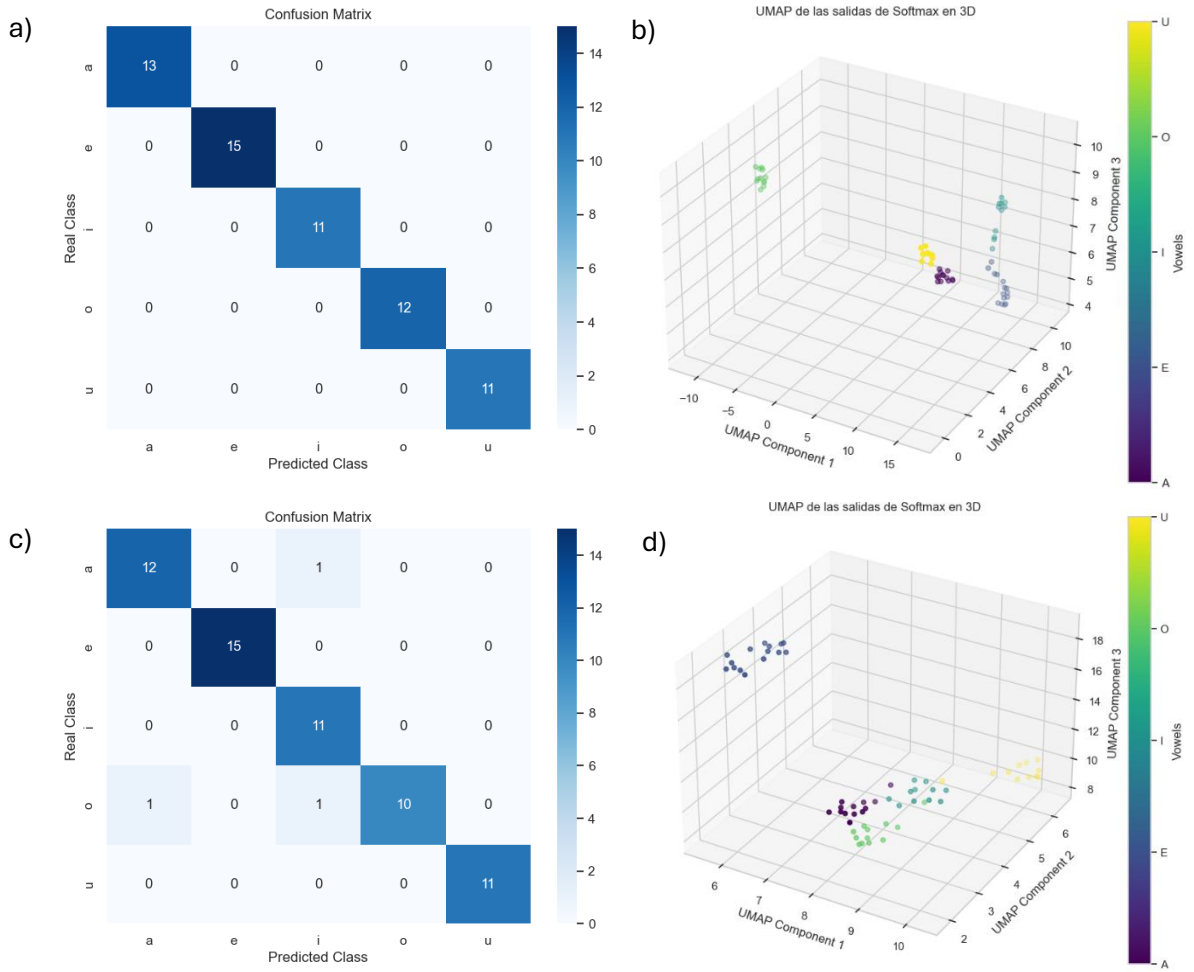
Table 4

Wilcoxon signed-rank tests between all the frequency bands for the classification of Rest vs. Imagined speech vs. Overt speech

Comparison	*Z*-scores	*p*-values
Alpha vs. Beta	27.0	.345
Alpha vs. Gamma	7.0	.004
Alpha vs. High Gamma	11.5	.017
Beta vs. Gamma	7.0	.004
Beta vs. High gamma	4.0	.009
Gamma vs. High gamma	36.0	.81

Figure 12

*(a) **Confusion Matrix for Vowel Classification in General Band (1-120 Hz) of Participant 10.** This figure displays the confusion matrix for the classification of the five vowels. The values on the main diagonal indicate the number of correct predictions, while the off-diagonal values reflect the misclassifications between classes. The colors in the matrix represent the frequency of classifications, with more intense colors indicating a higher number of predictions. (b) **UMAP Projection of Vowels in General Band (1-120 Hz) of Participant 10**. This figure presents the UMAP projection of the five vowels in a reduced feature space. Each point represents a vowel, with colors indicating the corresponding classes. The figure illustrates how the vowels are grouped and separated in the reduced space, aiding in the evaluation of the model's differentiation capability. (c) and (d) are the same for the participant 14.*



3.2. Vowel Decoding in Imagined Speech

Each participant was asked to produce 38 times each vowel (each vowel combined with each of the 19 consonants, presented once with in the written format and once in the auditory format) in the syllables production task. According to the second main goal of this study, we concentrated on classifying the 5 vowels from the EEG data in the imagined speech task with the goal of achieving fine-level classification accuracy. Our model achieved an average accuracy of 92% for vowel classification using the general band (Supplementary Material Figure 5). This classification result was significantly higher than the chance level of 20% and signifies a notable advancement in the field of imagined speech decoding, particularly in distinguishing among specific vowel classes. Classification results of two participants are represented in Figure 10. Our model demonstrates strong performance, aligning closely with established models (e.g., Hossain et al., 2024) and contributing valuable insights into the decoding of vowels through imagined speech.

When comparing the decoding accuracy with randomized data (20%) on each frequency band, we observed performance levels that were significantly above chance in each of them (see Table 5; Supplementary Material Figure 6). These results underscore that linguistic information can be decoded on all evaluated frequency bands: Alpha, Beta, Gamma, and High Gamma. However, it is important to note that some studies reported higher classification accuracies in higher frequency bands (e.g., Koizumi et al., 2018). Nevertheless, there is considerable debate about the reliability of these results due to

potential muscular artifacts and signal-to-noise ratio issues. Specifically, high gamma activity in EEG signals is often suspected to be influenced by muscle artifacts rather than reflecting true neural activity (Panachakel & Ramakrishnan, 2021). This skepticism is supported by findings suggesting that the gamma band may suffer from a lower signal-to-noise ratio and that its power decreases with increasing frequency, following a 1/f power law (Panachakel & Ramakrishnan, 2021; Koizumi et al., 2018).

Table 5

Discrimination accuracy compared to chance level in classifying the vowels across frequency bands using the Wilcoxon signed-rank test

Frequency bands	Accuracy (means)	*Z*-scores	*p*-values	**Reference value**
Alpha	87.58%	33.38	< .001	**20%**
Beta	88.06%	29.11	< .001	**20%**
Gamma	80.80%	10.54	< .001	**20%**
High Gamma	63.34%	10.12	< .001	**20%**

Thus, in order to evaluate the relative effectiveness in vowel decoding in the Alpha (8-12 Hz), Beta (13-30 Hz), Gamma (31-50 Hz) and High Gamma (51-120 Hz) bands, we conducted Wilcoxon signed-rank tests to compare the decoding differences across these frequency bands (see Table 6). Following the proposed hypothesis, our results highlighted the effectiveness of using the Beta and Alpha bands for decoding imagined speech. While no significant difference was observed between these two bands, all the other, the accuracy levels on all the other bands were significantly lower, being lowest for the high gamma frequency. Figure 11 represents the distribution of the accuracy data across each frequency band compared to a random classification.

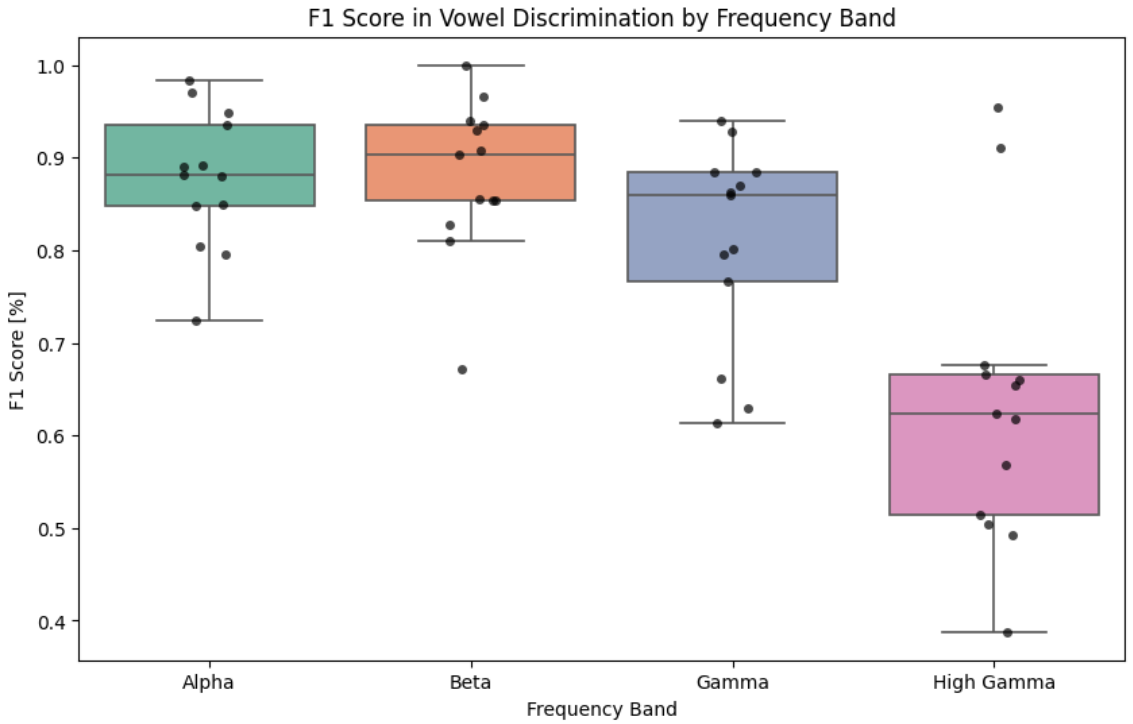
Table 6

Statistical comparisons of frequency bands for vowel decoding using the Wilcoxon signed-rank test

Comparison	*Z*-scores	*p*-values
Alpha vs Beta	39.0	.684
Alpha vs Gamma	14.0	.026
Alpha vs High gamma	1.0	< .001
Beta vs Gamma	14.0	.026
Beta vs High gamma	2.0	< .001
Gamma vs High gamma	3.0	< .001

In order to analyze how well the model avoids false positives and captures true positives in each band, we used a metric frequently used in the fields that use Deep Learning or Machine learning models (Panachakel & Ramakrishnan, 2021), the F1 score. This metric combines two other metrics: Precision and Recall. In our case the Precision value corresponds to the accuracy of the model in predicting the specific vowels (a high precision means that the model makes few false positive errors) and the Recall value corresponds to model's ability to identify all instances of the specific vowels correctly (a high recall means that the model makes few false negative errors). The formula to

Figure 12.
F1 scores with Alpha, Beta, Gamma and High gamma bands data when classifying vowels



performance. As shown in Figure 12, the highest F1 scores were observed with the Alpha and the Beta bands data, followed by the data of the gamma band and the lowest F1 score was observed with the High gamma data (see Figure 12).

These analyses suggests that, in contrast to the High Gamma band, the Beta and Alpha bands offer more reliable information for decoding imagined speech. These results contrast with the inconsistent findings regarding the use of information from the higher frequency bands (e.g., Koizumi et al., 2018) and align with existing literature, which often highlights the superior classification performance with data from the Beta and Alpha bands across various cognitive and speech-related tasks (Hossain et al., 2024).

3.3. Semantic categories decoding in Imagined Speech

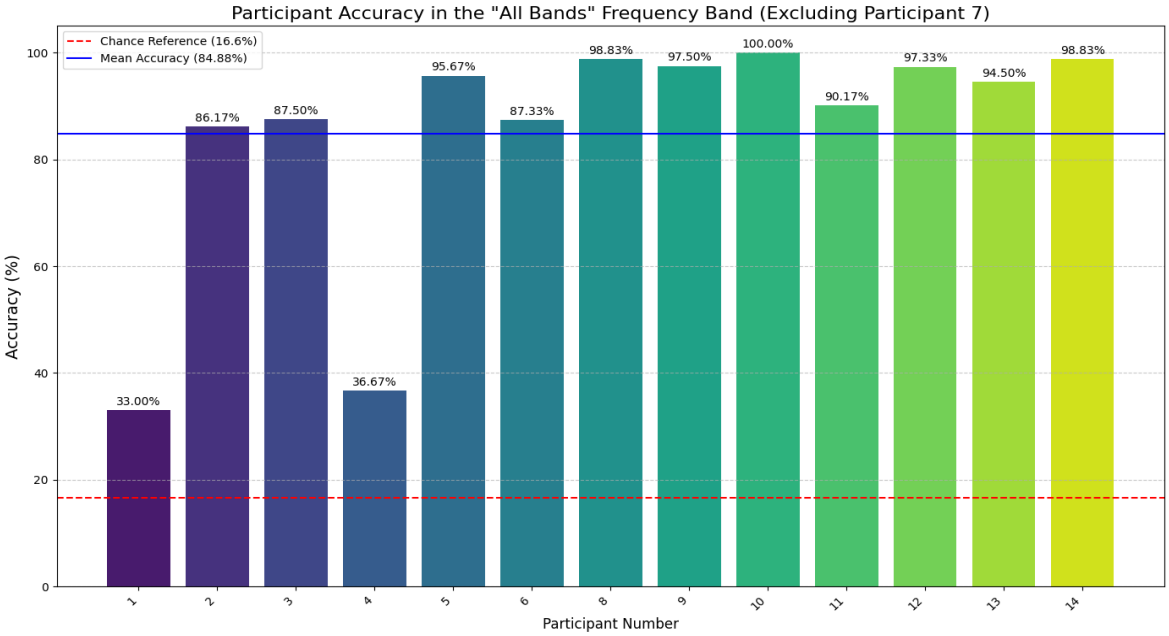
In this section, we present the application of our model using Morlet wavelets for different semantic categories: Clothing, Kitchen Utensils, Animals, Body Parts, Instruments, and Food. With six categories in total, the probability of random guessing resulting in a correct classification was 16.66%. This aspect is particularly significant in the context of Semantic Silent BCIs (SS-BCIs), which aim to address the challenge of decreasing model accuracy as more words are decoded. Despite the promising potential of SS-BCIs, the literature on this topic remains limited. To the best of our knowledge, the study by Rekrut et al. (2021) is the only one that has addressed this methodology in EEG-based imagined speech studies. Rekrut et al. (2021) proposed a method for first discriminating the semantic category of a word and then selecting the specific word within that category, achieving an accuracy of 43.54%. This approach shows potential for increasing the number of decodable words by utilizing semantic categories. However, the application of semantic categories to expand the decodable vocabulary has not been extensively explored, primarily due to the challenges associated with decoding EEG signals linked to various semantic categories during imagined speech. Consequently, accurately

decoding EEG signals associated with different semantic categories during imagined speech remains one of the primary objectives in this field.

In the words-pseudowords block, participants were asked to produce 60 stimuli per semantic category (10 words being presented in the auditory, written or picture format and being produced in imagined or overt speech). The average accuracy across all participants was 84.88% using the general frequency band, ranging from 33% to 100% (see Figure 13; see Supplementary Material Figure 7).

Figure 13

Individual participants' accuracy compared to chance in decoding Morlet wavelets for semantic



To evaluate performance across different EEG frequency bands, we conducted Wilcoxon signed-rank tests on each frequency band. The results showed that the classification accuracy was above chance on each band (see Table 7; See Supplementary Figure 8).

Table 7

Discrimination accuracy compared to chance level in classifying the vowels across frequency bands using the Wilcoxon signed-rank test

Comparison	*Z-scores*	*p-values*
Alpha vs Beta	39.0	0.684
Alpha vs Gamma	14.0	0.026*
Alpha vs High Gamma	1.0	p > 0.001
Beta vs Gamma	14.0	0.026
Beta vs High Gamma	2.0	p > 0.001
Gamma vs High Gamma	3.0	0.001

Additionally, we performed a Wilcoxon signed-rank test to compare the accuracy between the different frequency bands. The results, summarized in Table 8, indicate no

significant difference between the Alpha and Beta bands, but significantly lower scores for the Gamma band data and the lowest scores for the High gamma band data (see Table 8; see Figure 14). These results suggest that while overall performance is high, there are significant differences in accuracy depending on the frequency band used.

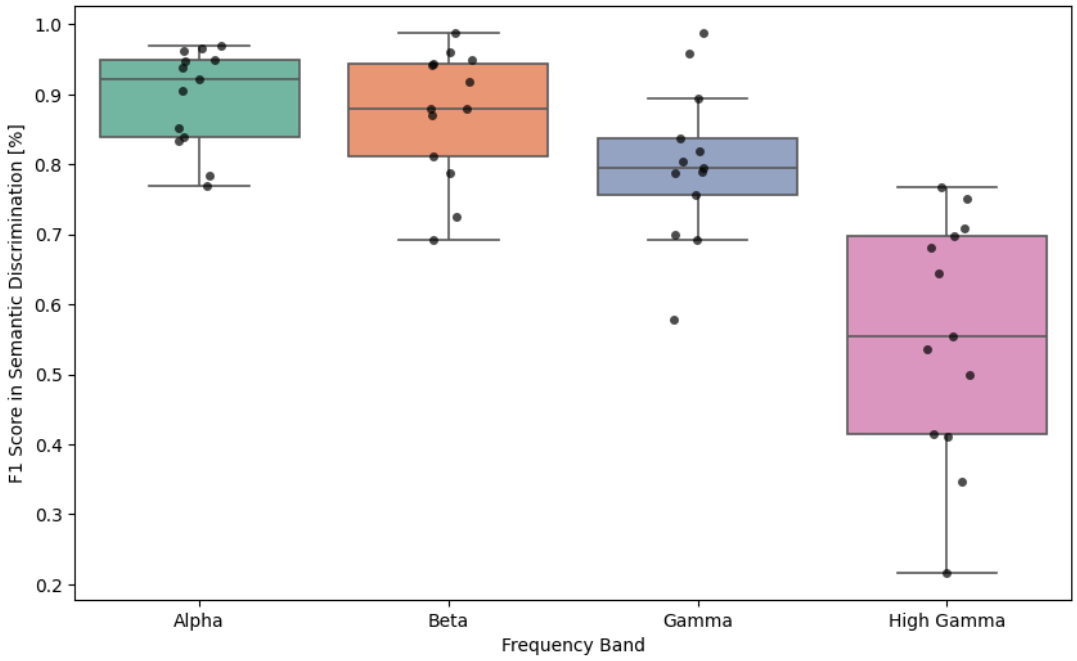
Table 8

Statistical comparison of frequency bands for vowel decoding using the Wilcoxon signed-rank test

Comparison	*Z*-scores	*p*-values
Alpha vs Beta	39.0	.684
Alpha vs Gamma	14.0	.026
Alpha vs High Gamma	1.0	< .001
Beta vs Gamma	14.0	.026
Beta vs High Gamma	2.0	< .001
Gamma vs High Gamma	3.0	.001

Figure 14

F1 scores with Alpha, Beta, Gamma and High gamma bands data when classifying semantic categories.



4. Discussion

This study focused on the decoding of imagined speech using deep learning models, an area of great relevance for the development of brain-computer interfaces (BCIs). The advancement of such technology is particularly crucial for assisting individuals with severe communication impairments, such as those caused by amyotrophic lateral sclerosis (ALS). Over the past decade, the interest on this topic has increased and various methodologies focusing on different neural oscillations have been explored. As a result, the field has become somewhat fragmented, with diverse approaches and models emerging. This study aimed to provide a comprehensive overview of different methodologies relevant to imagined speech decoding.

The use of statistical analysis in our study is particularly noteworthy, as the field typically prioritizes decoding accuracy measures without the use of any statistical test. These analyses allowed us to confirm the statistical significance of the results.

4.1. Types of speech production decoding

Our initial analyses focused on distinguishing between imagined speech, overt speech, and rest. These analyses are instrumental in assessing whether or not our model can effectively differentiate imagined speech from rest and/or overt speech. Our results demonstrated that the model can accurately distinguish between the production speech versus without the involvement of the articulators. They also demonstrated that it is possible to distinguish the production of imagined from the absence of speech production (the Rest condition). The statistical results indicated that our model not only classifies EEG signals associated with imagined speech effectively but also performs well beyond random guessing. Although these results may not represent a novel breakthrough in the field, they validate the operational effectiveness of our model, ensuring that it performs reliably under the tested conditions.

4.2. Vowels Decoding

Decoding vowels from imagined speech is not entirely novel; early studies achieved above-chance performance in vowel decoding as early as 2016 (Min et al., 2016; Coretto et al., 2017). More recently, LaRocco et al. (2023) achieved an impressive 98% accuracy in decoding English phonemes. Our study, however, represents a pioneering effort in the effective decoding of vowels from imagined speech in Spanish. Notably, previous attempts, such as those by Coretto et al. (2017), showed performance levels close to chance and lacked statistical analysis to ascertain the significance of their findings. In contrast, our study achieved an average accuracy of 92.06% in vowel decoding, making it the most accurate study to date for vowel decoding in Spanish.

Research into neural oscillations associated with imagined speech is relatively recent (Proix et al., 2022). This field remains contentious due to the challenges linked with high-frequency bands. Studies utilizing electrocorticography (ECoG) have indicated that the High Gamma band (80-120 Hz) often demonstrates superior performance in decoding imagined speech (Panachakel & Ramakrishnan, 2021). However, the use of high-frequency bands in EEG studies is debated due to their lower signal-to-noise ratios and potential for artifact contamination (Panachakel & Ramakrishnan, 2021). For instance, Synigal et al. (2020) argued that frequency bands above Beta are often contaminated, making frequencies above 30 Hz less useful for decoding purposes.

Our results are consistent with the findings of Hossain et al. (2024), who identified the Alpha (8-12 Hz) and Beta (13-30 Hz) bands as particularly effective for classifying letters and digits. In addition to these bands, we also used analyses that included higher frequency bands, primarily due to the ongoing controversy surrounding the High Gamma band. We found that the Beta and Alpha bands offered a more reliable information for decoding imagined speech compared to the Gamma and the High Gamma bands. Nevertheless, our results indicated that reliable information form the Gamma and High gamma bands could be used to classify the vowels. This finding challenges the hypothesis put forward by Synigal et al. (2020), which suggested that frequencies above Beta are not useful due to contamination. Our results indicate that Gamma and High gamma bands remain quite useful for decoding tasks, despite their known issues with artifact contamination.

4.3. Semantic categories decoding

One of the most significant findings of our research is the remarkably high performance in semantic decoding from EEG signals, an area that has shown limited results in the scientific literature to date. Our model achieved an average accuracy of 84.88% in semantic categorization, with a chance level set at 16.6%. This result significantly surpasses the benchmark previously established by Rekrut et al. (2021), who reported an accuracy of 44.54% in similar tasks, with a chance level of 20%. This advancement represents an important step forward in the field of Brain-Computer Interfaces (BCI), particularly in enhancing the ability to interpret neural signals associated with the semantics of imagined speech.

The study by Rekrut et al. (2021) was pioneering in addressing semantic decoding using EEG, but it faced several limitations that our model has successfully overcome. One of the primary challenges in semantic decoding lies in the diffuse and complex nature of semantic representations in the brain. Rekrut et al. utilized basic machine learning (ML) algorithms, such as Random Forest (RF) and Support Vector Machine (SVM), which may have constrained their classification performance. In contrast, our study leverages deep learning (DL) models that combine features from convolutional and recurrent neural networks, along with advanced optimization algorithms like Adam, which have significantly improved semantic decoding performance.

This work introduces significant contributions to the existing literature, particularly in the development of Semantic Silent Speech BCIs. These systems use a semantic classification stage prior to word classification, enabling a substantial increase in the number of words our algorithm can accurately decode. These advancements hold the potential to enhance the effectiveness and applicability of BCI systems for imagined speech decoding, offering new opportunities for individuals with severe communication disabilities (Rekrut et al., 2021).

Moreover, our statistical analyses highlighted the importance of the Alpha (8-12 Hz) and Beta (13-30 Hz) bands in semantic categories decoding, similar to our findings in vowel decoding. However, it is important to note that somewhat less reliable information could also be used from the Gamma and High gamma bands. To the best of our knowledge, no other semantic classification study has evaluated the effectiveness of the different

frequency bands in semantic categorisation, making our findings a potential contribution to the state-of-the-art techniques aimed at decoding semantic information from EEG signals. These differences underscore the superiority of the Alpha and Beta bands over the Gamma and High Gamma bands when decoding semantic information is of interest.

4.4. Future Directions

4.4.1. Transfer Learning

The field of computational neuroscience is rapidly evolving, and the algorithms developed in this work may be surpassed within a short span of time. To keep pace with the state-of-the-art techniques, particularly in EEG-based decoding of imagined speech, one of the most promising directions for future research is the development of transfer learning algorithms. Specifically, the application of Riemannian manifold-based techniques holds significant potential for advancing in this field, as it is already being applied in other areas of EEG decoding, such as motor decoding, which is also linked to the development of BCIs (Xu et al., 2021).

Transfer learning is crucial for addressing one of the most persistent challenges in EEG research: the high inter-participants variability. EEG signals are notoriously difficult to generalize across different individuals due to the unique neural signatures of each person. Traditional machine learning models often struggle with this variability, requiring extensive retraining for each new user, which is both time-consuming and inefficient. By leveraging transfer learning, it becomes possible to extrapolate and adapt the learned features from one participant to another, thereby reducing the need for individualized training sessions and improving the system's scalability (Cooney et al., 2019).

Riemannian manifolds offer a powerful framework for this purpose. They allow for the representation of EEG data in a way that captures the underlying geometric structure of the brain's activity patterns. This representation can facilitate the transfer of knowledge across participants by aligning the neural data from different individuals in a common space. In this space, the intrinsic properties of the EEG signals are preserved, making it easier for algorithms to generalize across participants with minimal loss of accuracy. The use of Riemannian manifold-based transfer learning could, therefore, significantly enhance the adaptability and robustness of EEG-based imagined speech decoding systems.

4.4.2. Semantic Decoding

Another promising avenue for future research is the development of a hierarchical decoding interface that first identifies the semantic category of the imagined word and then decodes the specific word within that category. This two-stage decoding process could potentially lead to more accurate and efficient decoding. By initially narrowing down the possible words through semantic categorization, the system can focus on a smaller set of candidate words, thereby reducing the complexity of the decoding task and improving overall performance (Rekrut et al., 2021).

4.5. Limitations

One of the primary limitations of this study is the inter-subject variability, a problem closely linked to the use of EEG. Despite ensuring that impedance did not drop

below 20Ω, the variability was particularly noticeable in the average semantic decoding accuracy, where two participants had a precision near 35%. This limitation is especially important when considering future research directions, such as the aforementioned transfer learning. We are now intending to train our DL on a database created with previous participants (Cooney et al., 2018), in order to investigate if it can generalize to new participants. This is the major challenge in developing BCIs for imagined speech decoding. Due to the high inter-participants variability, developing an effective non-invasive neuroprosthesis is extremely difficult.

Another limitation is the low number of participants. While it is common in this field to find works with 10 participants or fewer (Abdulghani et al., 2023), increasing the number of participants could aid in the generalization of imagined speech decoding algorithms. However, it is important to consider that this kind of research is time demanding, each session lasting around three hours and a half plus the preparation time and the cleaning of the electrodes and the cap after each session.

5. Conclusion

In this study, we demonstrated the efficacy of deep learning models in decoding imagined speech from EEG signals. Our research addressed several key issues in this field, including differentiating between imagined speech, overt speech, and rest as a manipulation check in decoding (Zhao and Rudznick, 2014), as well as classifying vowels and semantic categories. The results provide valuable insights into the relative effectiveness of using different EEG frequency bands and introduce a new decoding methodology. One of the most notable findings is that the Alpha (8-12 Hz) and Beta (13-30 Hz) bands are more informative than the Gamma and High Gamma bands for decoding both vowels and semantic categories. Our model showed to be similarly efficient compared to other state-of-the-art models in classifying vowels (Hossain et al., 2024) and surpassed recent benchmarks in semantic categories classification (Rekrut et al., 2021). These results opens promising avenues for the development of new Brain-Computer Interfaces (BCIs) for individuals with severe communication impairments, like allowing a reduction of the vocabulary search by the BCI thanks to the identification of the semantic categories.

6. Acknowledges

To ensure that the analyses conducted, and the structure of the deep learning model are as transparent as possible, a GitHub repository with all the information is available upon request. https://github.com/ibonvales/Decoding-of-imagined-speech-in-EEG-by-Deep-Learning.git

References

Adlington, R. L., Laws, K. R., & Gale, T. M. (2009). The Hatfield Image Test (HIT): A new picture test and norms for experimental and clinical use. *Journal of clinical and experimental neuropsychology*, *31*(6), 731-753. https://doi.org/10.1080/13803390802488103

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568*(7753), 493-498. https://doi.org/10.1038/s41586-019-1119-1

Bauer, G., Gerstenbrand, F., Rumpl, E. (1979). Varieties of the locked-in syndrome. *Journal of Neurology, 221*(2): 77–91. https://doi.org/10.1007/BF00313105

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one*, *5*(5), e10773. https://doi.org/10.1371/journal.pone.0010773

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. PloS one, 9(9), e106953. https://doi.org/10.1371/journal.pone.0106953

Chen, X., Wang, R., Khalilian-Gourtani, A., Yu, L., Dugan, P., Friedman, D., Doyle, W., Devinsky, O., Wang, Y., & Flinker, A. (2024). A neural speech decoding framework leveraging deep learning and speech synthesis. *Nature Machine Intelligence, 6*, 467–480. https://doi.org/10.1038/s42256-024-00824-8

Cooney, C., Folli, R., & Coyle, D. (2018). Neurolinguistics Research Advancing Development of a Direct-Speech Brain-Computer Interface. *iScience*, *8*, 103–125. https://doi.org/10.1016/j.isci.2018.09.016

Cooney, C., Folli, R., & Coyle, D. (2019, October). Optimizing layers improves CNN generalization and transfer learning for imagined speech decoding from EEG. In *2019 IEEE international conference on systems, man and cybernetics (SMC)* (pp. 1311-1316). IEEE. https://doi.org/10.1109/SMC.2019.8914246

Cooney, C., Korik, A., Folli, R., & Coyle, D. (2020). Evaluation of Hyperparameter Optimization in Machine and Deep Learning Methods for Decoding Imagined Speech EEG. *Sensors (Basel, Switzerland)*, *20*(16), 4629. https://doi.org/10.3390/s20164629

Coretto, G. A. P., Gareis, I. E., & Rufiner, H. L. (2017, January). Open access database of EEG signals recorded during imagined speech. In *12th International Symposium on Medical Information Processing and Analysis* (Vol. 10160, p. 1016002). SPIE. https://doi.org/10.1117/12.2255697

Corpus del Español del Siglo XXI. (octubre de 2013). https://www.rae.es/banco-de-datos/corpes-xxi

D'Zmura, M., Deng, S., Lappas, T., Thorpe, S. & Srinivasan, R. (2009). Toward EEG sensing of imagined speech. Jacko, J.A. (Ed.), *Human-Computer Interaction,*

Part I, HCII 2009, LNCS 5610 (Berlin:Springer) 40-48. https://doi.org/10.1007/978-3-642-02574-7_5

Dekker, B., Schouten, A., & Scharenborg, O. (2023). DAIS: The Delft Database of EEG Recordings of Dutch Articulated and Imagined Speech. *In Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings; Vol. 2023-June). IEEE. https://doi.org/10.1109/ICASSP49357.2023.10096145

Donchin, E., Spencer, K. M., & Wijesinghe, R. (2000). The mental prosthesis: assessing the speed of a P300-based brain-computer interface. *IEEE transactions on rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, *8*(2), 174–179. https://doi.org/10.1109/86.847808

Farwell, L. A., & Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical neurophysiology*, *70*(6), 510–523. https://doi.org/10.1016/0013-4694(88)90149-6

García, A. A. T., García, C. A. R., & Pineda, L. V. (2012, February). Toward a silent speech interface based on unspoken speech. In *International Conference on Bio-inspired Systems and Signal Processing* (Vol. 2, pp. 370-373). SciTePress. https://doi.org/10.5220/0003769603700373

Gonzalez-Lopez, J. A., Gomez-Alanis, A., Donas, J. M. M., Perez-Cordoba, J. L., & Gomez, A. M. (2020). Silent Speech Interfaces for Speech Restoration: A Review. IEEE Access, 8, 177995-178021. https://doi.org/10.1109/ACCESS.2020.3026579

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L. & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in neuroscience, 7*, 70133. https://doi.org/10.3389/fnins.2013.00267

Grandchamp, R., & Delorme, A. (2011). Single-trial normalization for event-related spectral decomposition reduces sensitivity to noisy trials. *Frontiers in psychology*, *2*, 10583. https://doi.org/10.3389/fpsyg.2011.00236

Hamedi, M., Salleh, S.hH., & Noor, A. M. (2016). Electroencephalographic Motor Imagery Brain Connectivity Analysis for BCI: A Review. *Neural computation, 28*(6), 999–1041. https://doi.org/10.1162/NECO_a_00838

Hecht, M., Hillemacher, T., Gräsel, E., Tigges, S., Winterholler, M., Heuss, D., Hilz, M. J., & Neundörfer, B. (2002). Subjective experience and coping in ALS. *Amyotrophic lateral sclerosis and other motor neuron disorders, 3*(4), 225–231. https://doi.org/10.1080/146608202760839009

Herff, C., Heger, D., De Pesters, A., Telaar, D., Brunner, P., Schalk, G., & Schultz, T. (2015). Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, *8*, 141498. https://doi.org/10.3389/fnins.2015.00217

Herff, C., Krusienski, D. J., & Kubben, P. (2020). The potential of stereotactic-EEG for brain-computer interfaces: current progress and future directions. *Frontiers in neuroscience*, *14*, 483258. https://doi.org/10.3389/fnins.2020.00123

Hermes D, Miller KJ, Vansteensel MJ, Edwards E, Ferrier CH, Bleichner MG, van Rijen PC, Aarnoutse EJ, Ramsey NF (2014). Cortical theta wanes for language. *Neuroimage*. *85*(2), 738-748. https://doi.org/10.1016/j.neuroimage.2013.07.029

Hossain, A., Khan, P., & Kader, M. F. (2024). Imagined speech classification exploiting EEG power spectrum features. *Medical & biological engineering & computing*, 10.1007/s11517-024-03083-2. Advance online publication. https://doi.org/10.1007/s11517-024-03083-2

Ikeda, K., Higashi, T., Sugawara, K., Tomori, K., Kinoshita, H., & Kasai, T. (2012). The effect of visual and auditory enhancements on excitability of the primary motor cortex during motor imagery: a pilot study. *International Journal of Rehabilitation Research*, *35*(1), 82-84. https://doi.org/10.1097/MRR.0b013e32834d2032

Iljina, O., Derix, J., Schirrmeister, R.T., Schulze Bonhage, A., Auer, P., Aertsen, A., and Ball, T. (2017). Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication. *Brain Computer Interfaces 4*, 186–199. https://doi.org/10.1080/2326263X.2017.1330611

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. *Journal of medical systems*, *43*(2), 20. https://doi.org/10.1007/s10916-018-1137-9

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

Klug, M., & Gramann, K. (2021). Identifying key factors for improving ICA-based decomposition of EEG data in mobile and stationary experiments. *European Journal of Neuroscience*, *54*(12), 8406-8420. https://doi.org/10.1111/ejn.14992

Koch-Fager, S., Fried-Oken, M., Jakobs, T., & Beukelman, D. R. (2019). New and emerging access technologies for adults with complex communication needs and severe motor impairments: State of the science. *Augmentative and Alternative Communication*, *35*(1), 13-25. https://doi.org/10.1080/07434618.2018.1556730

Koizumi, K., Ueda, K., & Nakao, M. (2018, July). Development of a cognitive brain-machine interface based on a visual imagery method. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 1062-1065). IEEE. https://doi.org/10.1109/EMBC.2018.8512520

Kothe, C., Yahya Shirazi, S., Stenner, T., Medine, D., Boulay, C., Grivich, M. I., Mullen, T., Delorme, A., & Makeig, S. (in press). The Lab Streaming Layer for Synchronized Multimodal Recording. *bioRxiv.*

Krishna, G., Tran, C., Carnahan, M., & Tewfik, A. H. (2021, May). Advancing speech synthesis using EEG. *In 2021 10th International IEEE/EMBS Conference on*

Neural Engineering (NER) (pp. 199-204). IEEE. https://doi.org/10.1109/NER49283.2021.9441306

LaRocco, J., Tahmina, Q., Lecian, S., Moore, J., Helbig, C., & Gupta, S. (2023). Evaluation of an English language phoneme-based imagined speech brain computer interface with low-cost electroencephalography. *Frontiers in neuroinformatics*, *17*, 1306277. https://doi.org/10.3389/fninf.2023.1306277

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of neural engineering*, *15*(5), 056013. https://doi.org/10.1088/1741-2552/aace8c

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541-551. https://doi.org/10.1162/neco.1989.1.4.541

Lee, S. H., Lee, M., & Lee, S. W. (2020). Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *28*(12), 2647-2659. https://doi.org/10.1109/TNSRE.2020.3040289

Li, F., Chao, W., Li, Y., Fu, B., Ji, Y., Wu, H., & Shi, G. (2021). Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network. *Journal of neural engineering*, *18*(4), 0460c4. https://doi.org/10.1088/1741-2552/ac13c0

Lopez-Bernal, D., Balderas, D., Ponce, P., & Molina, A. (2022). A state-of-the-art review of EEG-based imagined speech decoding. *Frontiers in human neuroscience*, *16*, 867281. https://doi.org/10.3389/fnhum.2022.867281

Lu, L., Sheng, J., Liu, Z., & Gao, J. H. (2021). Neural representations of imagined speech revealed by frequency-tagged magnetoencephalography responses. *NeuroImage*, *229*, 117724. https://doi.org/10.1016/j.neuroimage.2021.117724

Lu, L., Wang, Q., Sheng, J., Liu, Z., Qin, L., Li, L., & Gao, J. H. (2019). Neural tracking of speech mental imagery during rhythmic inner counting. *Elife*, *8*, e48971. https://doi.org/10.7554/eLife.48971

Lua, Z., Lib, W., Niec, L., & Zhaod, K. An Easy-to-Follow Handbook for EEG Data Analysis based on Python. 10.1002/brx2.64

Ma, Z., Wang, K., Xu, M., Yi, W., Xu, F., & Ming, D. (2023). Transformed common spatial pattern for motor imagery-based brain-computer interfaces. *Frontiers in Neuroscience*, *17*, 1116721. https://doi.org/10.3389/fnins.2023.1116721

Martin, S., Brunner, P., Holdgraf, C., Heinze, H. J., Crone, N. E., Rieger, J., Schalk, G., Knight, T. R. & Pasley, B. N. (2014). Decoding spectrotemporal features of overt

and covert speech from the human cortex. *Frontiers in neuroengineering*, *7*, 14. https://doi.org/10.3389/fneng.2014.00014

Martínez-Manrique, F., & Vicente, A. (2015). The activity view of inner speech. *Frontiers in psychology, 6,* 232. https://doi.org/10.3389/fpsyg.2015.00232

Maturana, D., & Scherer, S. (2015, September). Voxnet: A 3d convolutional neural network for real-time object recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 922-928). IEEE. https://doi.org/10.1109/IROS.2015.7353481

McFarland, D. J., Miner, L. A., Vaughan, T. M., & Wolpaw, J. R. (2000). Mu and beta rhythm topographies during motor imagery and actual movements. *Brain topography*, *12*, 177-186. https://doi.org/10.1023/a:1023437823106

Min, B., Kim, J., Park, H. J., & Lee, B. (2016). Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram. *BioMed research international*, *2016*. https://doi.org/10.1155/2016/2618265

Moreno-Martínez, F. J., & Montoro, P. R. (2012). An ecological alternative to Snodgrass & Vanderwart: 360 high quality colour images with norms for seven psycholinguistic variables. *PloS one*, *7*(5), e37527. https://doi.org/10.1371/journal.pone.0037527

Müller-Putz, G., Scherer, R., Brunner, C., Leeb, R., & Pfurtscheller, G. (2008). Better than random: a closer look on BCI results. *International journal of bioelectromagnetism*, *10*(1), 52-55.

Newman, A. J., Supalla, T., Hauser, P., Newport, E. L., & Bavelier, D. (2010). Dissociating neural subsystems for grammar by contrasting word order and inflection. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(16), 7539–7544. https://doi.org/10.1073/pnas.1003174107

Nguyen, C. H., Karavas, G. K., & Artemiadis, P. (2018). Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of neural engineering*, *15*(1), 016002. https://doi.org/10.1088/1741-2552/aa8235

Oh, S. L., Vicnesh, J., Ciaccio, E. J., Yuvaraj, R., & Acharya, U. R. (2019). Deep convolutional neural network model for automated diagnosis of schizophrenia using EEG signals. *Applied Sciences*, *9*(14), 2870. https://doi.org/10.3390/app9142870

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113. https://doi.org/10.1016/0028-3932(71)90067-4

Panachakel, J. T., & Ramakrishnan, A. G. (2021a). Decoding covert speech from EEG-a comprehensive review. *Frontiers in Neuroscience*, *15*, 642251. https://doi.org/10.3389/fnins.2021.642251

Panachakel, J. T., & G, R. A. (2021). Classification of Phonological Categories in Imagined Speech using Phase Synchronization Measure. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE*

Engineering in Medicine and Biology Society. Annual International Conference, *2021*, 2226–2229. https://doi.org/10.1109/EMBC46164.2021.9630699

Panachakel, J. T., Ramakrishnan, A. G., & Ananthapadmanabha, T. V. (2020). A novel deep learning architecture for decoding imagined speech from EEG. *arXiv preprint arXiv:2003.09374*. https://doi.org/10.48550/arXiv.2003.09374

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Perrone-Bertolotti, M., Rapin, L., Lachaux, J. P., Baciu, M., & Lœvenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural brain research*, *261*, 220–239. https://doi.org/10.1016/j.bbr.2013.12.034

Pfurtscheller, G., & Neuper, C. (2001). Motor imagery and direct brain-computer communication. In: *Proceedings of the IEEE 89*, pp. 1123–113. http://dx.doi.org/10.1109/5.939829

Proix, T., Delgado Saa, J., Christen, A., Martin, S., Pasley, B. N., Knight, R. T., Tian, X., Poeppel, D., Doyle, W. K., Devinsky, O., Arnal, L. H., Mégevand, P., & Giraud, A. L. (2022). Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. *Nature communications,* *13*(1), 48. https://doi.org/10.1038/s41467-021-27725-3

Rekrut, M., Sharma, M., Schmitt, M., Alexandersson, J., & Krüger, A. (2021, February). Decoding semantic categories from eeg activity in silent speech imagination tasks. In *2021 9th International Winter Conference on Brain-Computer Interface (BCI)* (pp. 1-7). IEEE. https://doi.org/ 10.1109/BCI51272.2021.9385357

Roy, S., Kiral-Kornek, I., & Harrer, S. (2018). ChronoNet: A deep recurrent neural network for abnormal EEG identification. In *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17* (pp. 47-56). https://doi.org/10.48550/arXiv.1802.00308

Ruffini, G., Ibañez, D., Castellano, M., Dubreuil-Vall, L., Soria-Frisch, A., Postuma, R., Gagnon, J. F., & Montplaisir, J. (2019). Deep Learning With EEG Spectrograms in Rapid Eye Movement Behavior Disorder. *Frontiers in neurology*, *10*, 806. https://doi.org/10.3389/fneur.2019.00806

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533-536. https://doi.org/10.1038/323533a0

Saryazdi, R., Bannon, J., Rodrigues, A., Klammer, C., & Chambers, C. G. (2018). Picture perfect: A stimulus set of 225 pairs of matched clipart and photographic images normed by Mechanical Turk and laboratory participants. *Behavior research methods*, *50*(6), 2498–2510. https://doi.org/10.3758/s13428-018-1028-5

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, *38*(11), 5391–5420. https://doi.org/10.1002/hbm.23730

Sereshkeh, A. R., Trott, R., Bricout, A., & Chau, T. (2017). EEG classification of covert speech using regularized neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2292-2300. https://doi.org/10.1109/TASLP.2017.2758164

Shah, U., Alzubaidi, M., Mohsen, F., Abd-Alrazaq, A., Alam, T., & Househ, M. (2022). The role of artificial intelligence in decoding speech from EEG signals: a scoping review. *Sensors*, *22*(18), 6975. https://doi.org/10.3390/s22186975

Stone J. V. (2002). Independent component analysis: an introduction. *Trends in cognitive sciences*, *6*(2), 59–64. https://doi.org/10.1016/s1364-6613(00)01813-1

Sutter, E. E. (1992). The brain response interface: communication through visually-induced electrical brain responses. *Journal of Microcomputer Applications*, *15*(1), 31-45. https://doi.org/10.1016/0745-7138(92)90045-7

Suyuncheva, A., Saada, D., Gavrilenko, Y., Schevchenko, A., Vartanov, A., & Ilyushin, E. (2021). Reconstruction of words, syllables, and phonemes of internal speech by EEG activity. In *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics: Proceedings of the 9th International Conference on Cognitive Sciences, Intercognsci-2020, Moscow, Russia 9* (pp. 319-328). https://doi.org/10.1007/978-3-030-71637-0_37

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9). https://doi.org/10.1109/CVPR.2015.7298594

Tamm, M. O., Muhammad, Y., and Muhammad, N. (2020). Classification of vowels from imagined speech with convolutional neural networks. *Computers 9*(46). https://doi.org/10.3390/computers9020046

Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1422-1432). https://doi.org/10.18653/v1/D15-1167

Tian, X., & Poeppel, D. (2010). Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in psychology*, *1*, 7029. https://doi.org/10.3389/fpsyg.2010.00166

Tian, X., & Poeppel, D. (2013). The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. *Journal of cognitive neuroscience*, *25*(7), 1020-1036. https://doi.org/10.1162/jocn_a_00381

Wang, L., Zhang, X., & Zhang, Y. (2013). Extending motor imagery by speech imagery for brain-computer interface. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and*

Biology Society. Annual International Conference, *2013*, 7056–7059. https://doi.org/10.1109/EMBC.2013.6611183

Willett, F. R., Kunz, E. M., Fan, C., Avansino, D. T., Wilson, G. H., Choi, E. Y., Kamdar, F., Glasser, M., Hochberg, L., Druckmann, S., Shenoy, K. & Henderson, J. M. (2023). A high-performance speech neuroprosthesis. *Nature*, *620*(7976), 1031-1036. https://doi.org/10.1038/s41586-023-06377-x

Wise, R., Chollet, F., Hadar, U. R. I., Friston, K., Hoffner, E., & Frackowiak, R. (1991). Distribution of cortical neural networks involved in word comprehension and word retrieval. *Brain*, *114*(4), 1803-1817. https://doi.org/10.1093/brain/114.4.1803

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical neurophysiology*, *113*(6), 767-791. https://doi.org/10.1016/s1388-2457(02)00057-3

Wu, S., Bhadra, K., Giraud, A. L., & Marchesotti, S. (2024). Adaptive LDA Classifier Enhances Real-Time Control of an EEG Brain-Computer Interface for Decoding Imagined Syllables. *Brain sciences*, *14*(3), 196. https://doi.org/10.3390/brainsci14030196

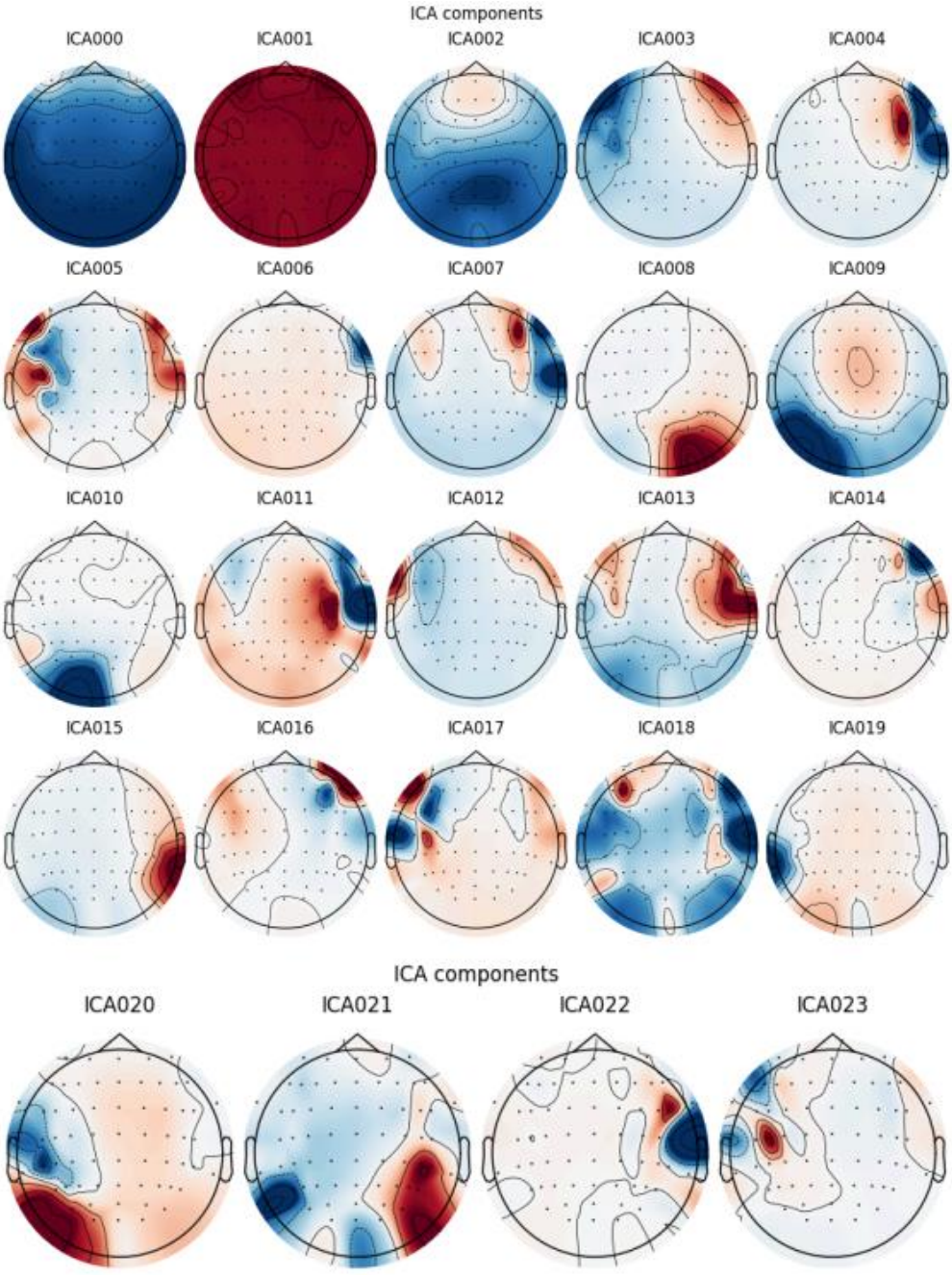
Xu, Y., Huang, X., & Lan, Q. (2021). Selective cross-subject transfer learning based on riemannian tangent space for motor imagery brain-computer interface. *Frontiers in Neuroscience*, *15*, 779231. https://doi.org/10.3389/fnins.2021.779231

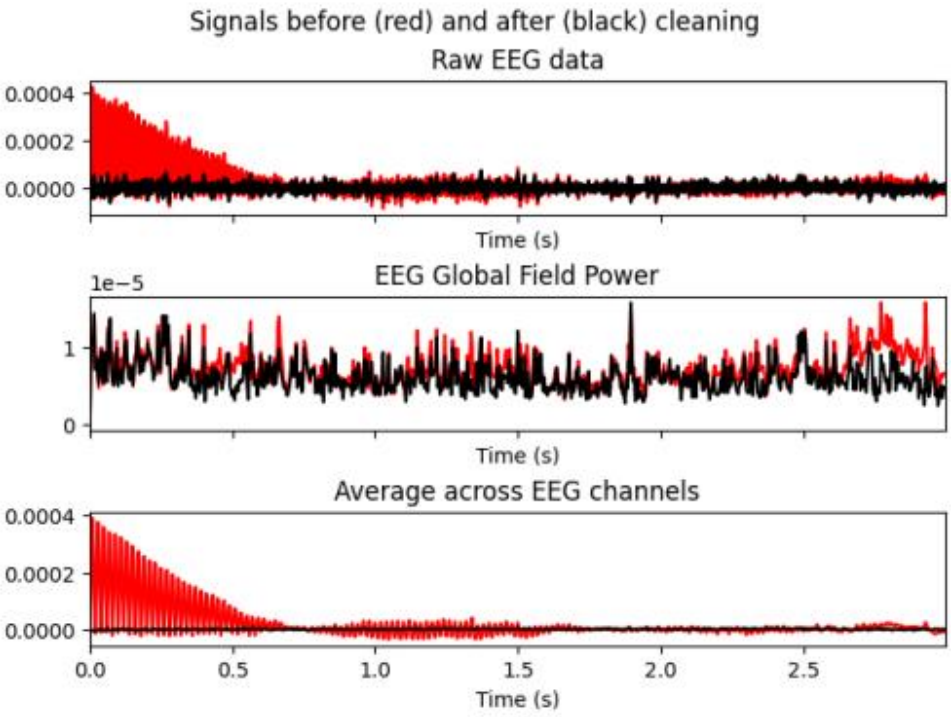
Zhao, S., & Rudzicz, F. (2015, April). Classifying phonological categories in imagined and articulated speech. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 992-996). IEEE. https://doi.org/10.1109/ICASSP.2015.7178118

Supplementary Material

Figure 1

Independent Component Analysis of the Pilot Participant's Data.




Signals before (red) and after (black) cleaning
Raw EEG data
0.0004
0.0002
0.0000
Time (s)
EEG Global Field Power
1e−5
1
0
Time (s)
Average across EEG channels
0.0004
0.0002
0.0000
0.0
0.5
1.0
1.5
2.0
2.5
Time (s)

Out[ ]:

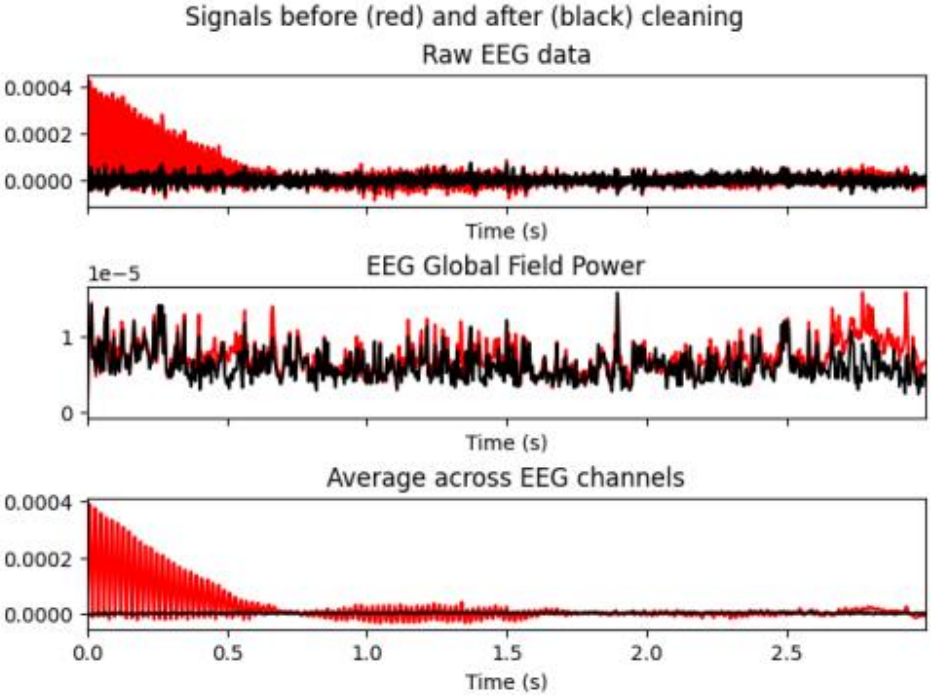

Signals before (red) and after (black) cleaning
Raw EEG data
0.0004
0.0002
0.0000
Time (s)
EEG Global Field Power
1e−5
1
0
Time (s)
Average across EEG channels
0.0004
0.0002
0.0000
0.0
0.5
1.0
1.5
2.0
2.5
Time (s)

Figure 2

Diagram flow of stimuli presentation, EEG preprocessing, and wavelet classification.

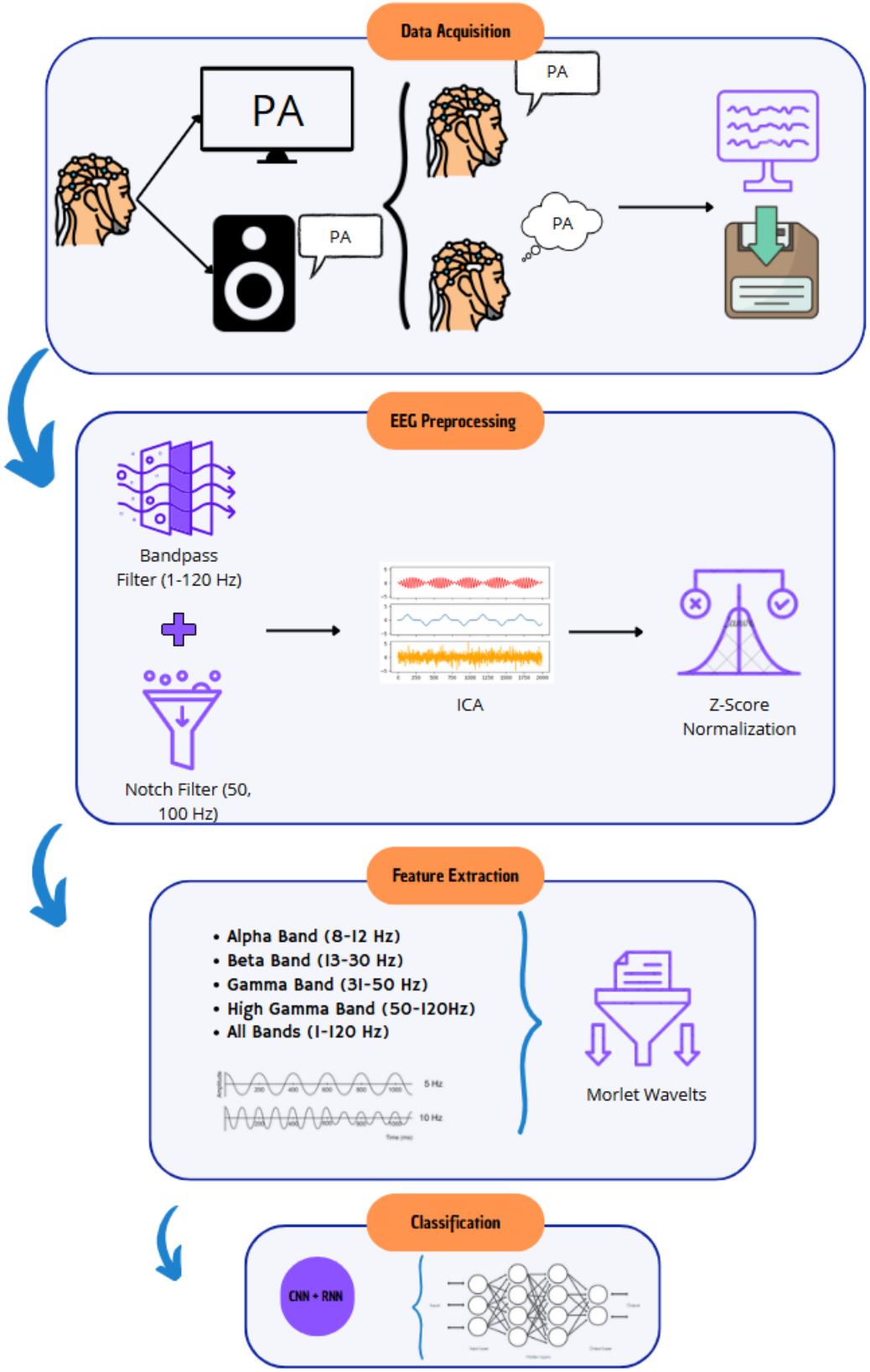


Figure 3

Average of signals from all channels for the three conditions of the pilot participant's data. First image corresponding to real speech, second image corresponding to imagined speech and third image corresponding to silence condition.

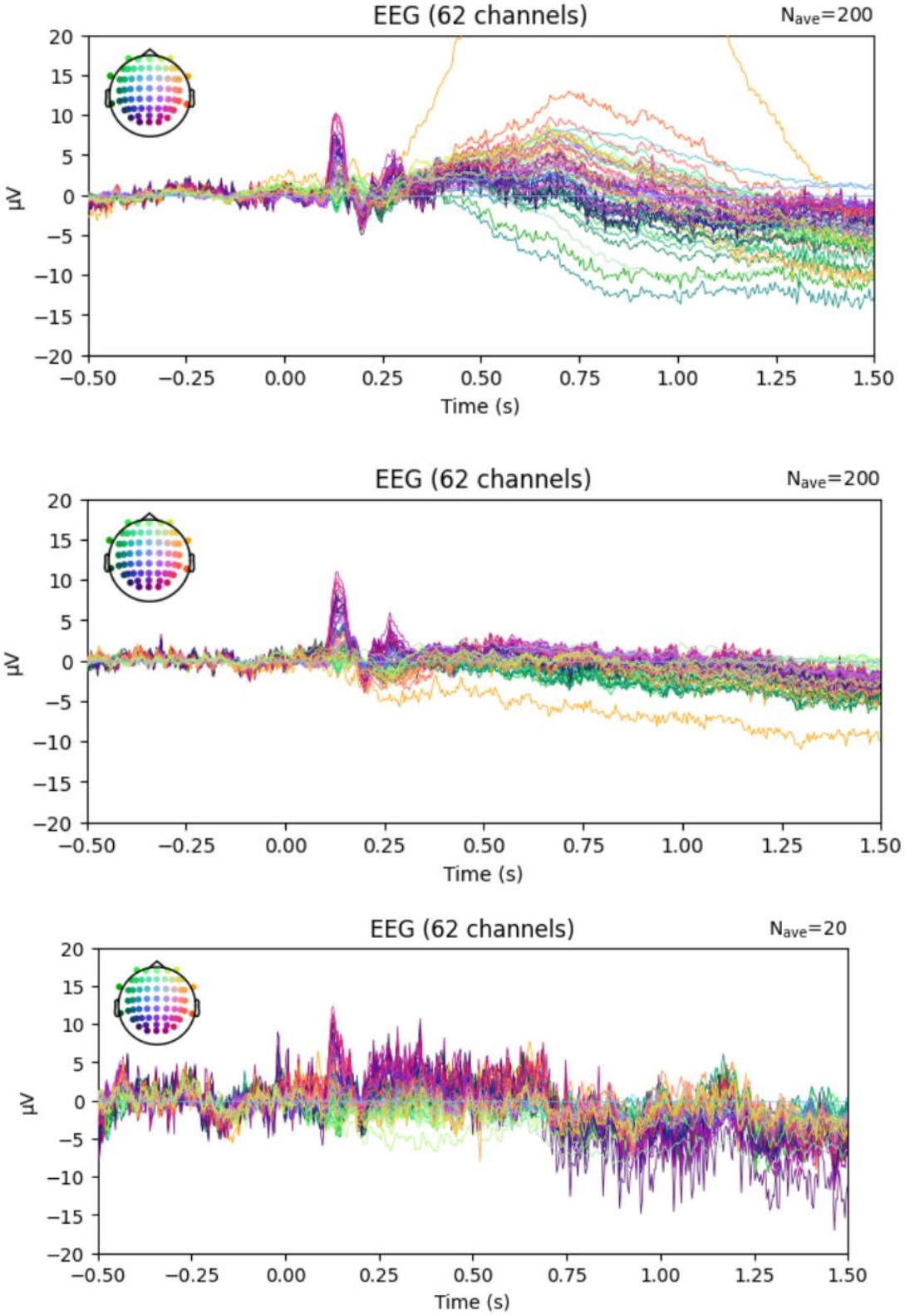


Figure 4

Violin plots representing the distribution of the accuracy data in classifying Rest vs. Imagined speech vs. Overt speech across across the Alpha, Beta, Gamma and High gamma frequency bands compared to a random classification on each band.

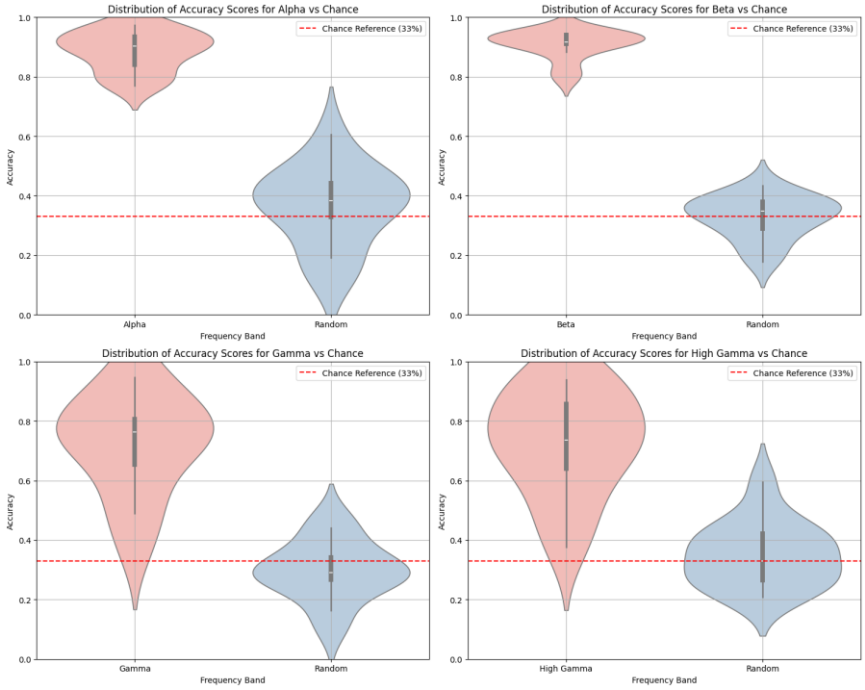


Figure 5

Mean classification accuracy for each participant across vowel classification

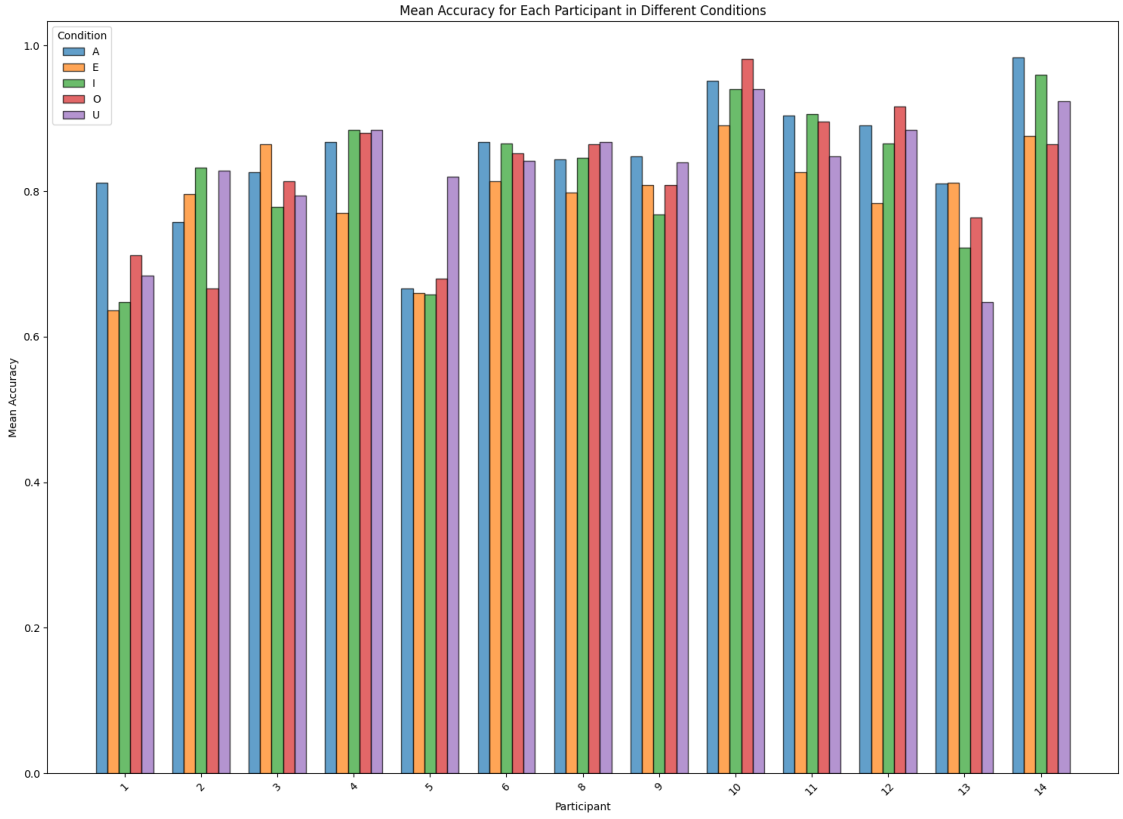


Figure 6

Violin plot representing the distribution of the accuracy data in classifying the vowels across the Alpha, Beta, Gamma and High gamma frequency bands compared to a random classification.

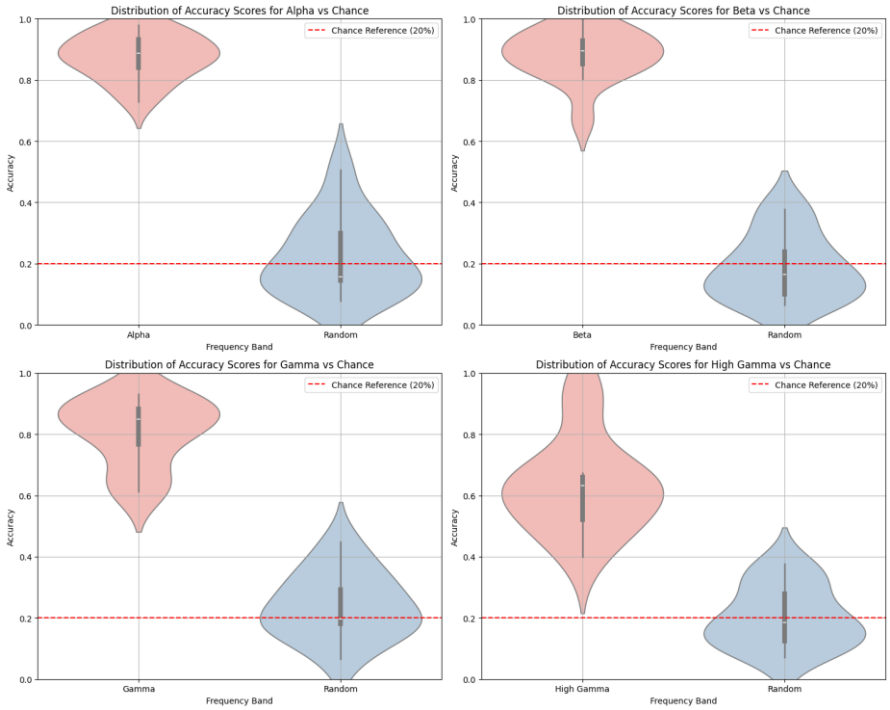


Figure 7

Mean classification accuracy for each participant across semanticl classification

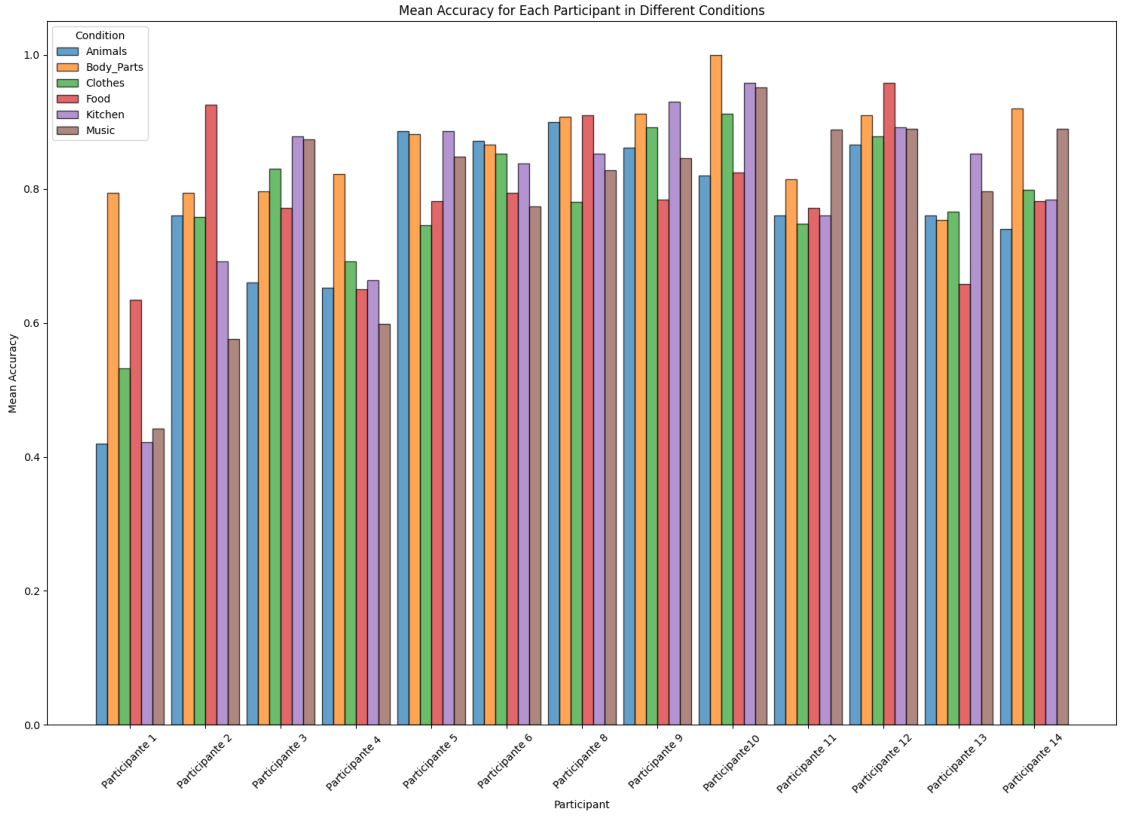


Figure 8

Violin plot representing the distribution of the accuracy data in classifying the semantic categories across the Alpha, Beta, Gamma and High gamma frequency bands compared to a random classification.