D2.1 – Report on the review of imagined speech decoding approaches

BINGO

Brain Imaged-Speech Communication

H.F.R.I. Hellenic Foundation for Research & Innovation

Greece 2.0 NATIONAL RECOVERY AND RESILIENCE PLAN

Funded by the European Union NextGenerationEU

Funded by the European Union
NextGenerationEU

Greece 2.0
NATIONAL RECOVERY AND RESILIENCE PLAN

H.F.R.I.
Hellenic Foundation for Research & Innovation

The research project is implemented in the framework of H.F.R.I call “Basic research Financing (Horizontal support of all Sciences)” under the National Recovery and Resilience Plan “Greece 2.0” funded by the European Union – NextGenerationEU (H.F.R.I. Project Number: 15986).

Dissemination level:	Public (PU)
Contractual date of delivery:	Month 5, 27/04/2024
Actual date of delivery:	Month 5, 26/04/2024
Work Package:	WP2 Neuroengineering for EEG-based Imagined Speech Decoding
Task:	T2.1 - Investigating existing practices
Type:	Report
Approval Status:	Final
Version:	v1.0
Number of pages:	24
Filename:	D2.1_Report_Review_ImaginedSpeech_Decoding_Approaches_v1.docx

Executive Summary: Deliverable D2.1 'Report on the review of imagined speech decoding approaches' summarizes the existing practices with respect to EEG-based imaged speech decoding with the aim to identify the most suitable methods and publicly available EEG-based datasets for imagined speech decoding necessary for the development of informed decoding algorithms. The report outlines the challenges of Brain Computer Interface (BCI) systems, the advancements in EEG signal processing and the shift towards deep learning techniques in data processing; specifically, deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to assess temporal and spatial dependencies within EEG signals.

HISTORY

Version	Date	Reason	Revised by
v0.1	13/02/2024	Table of contents **(ToC)** to be checked and revised by the research team	Fotis P. Kalaganis
v0.2	11/03/2024	Populated all sections – **Alpha version**	Georgiadis Kostas
v0.3	27/03/2024	Refinements across all sections – **Beta Version**	Georgiadis Kostas
v0.4	12/04/2024	Address comments from internal review	Georgiadis Kostas
v0.5	26/04/2024	Ready for submission - **Final version**	Nikolopoulos Spiros

AUTHOR LIST

Organization	Name	Contact Information
CERTH	Fotis P. Kalaganis	fkalaganis@iti.gr
CERTH	Kostas Georgiadis	Kostas.georgiadis@iti.gr
CERTH	Vangelis Oikonomou	viknmu@iti.gr
CERTH	Vasilis Kitsios	vkitsios@iti.gr
CERTH	Nikos Laskaris	laskaris@csd.auth.gr
CERTH	Spiros Nikolopoulos	nikolopo@iti.gr
CERTH	Ioannis Kompatsiaris	ikom@iti.gr

ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
BCI	Brain Computer Interface
CNNs	Convolutional Neural Networks
CTN	Connectionist Temporal Classification
DQN	Deep Q-Network
EEG	ElectroEncephaloGraphy
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional Near-Infrared Spectroscopy
GRU	Gated Recurrent Unit
ICA	Independent Component Analysis
LLMs	Large Language Models
LSTM	Long Short-Term Memory
MLPs	Multi-Layer Perceptors
RNNs	Recurrent Neural Networks
UDA	Unsupervised Domain Adaptation

Contents

1 INTRODUCTION

Inner speech, also known as imagined speech, refers to the silent, mental monologue individuals engage in without physically articulating words. In essence, it is an internal monologue associated with various cognitive processes (e.g. language planning, analyzing, and problem solving) triggered in almost any everyday activity, whether that is preparing for a challenging conversation, strategizing how to address a potential customer, or practicing responses to interview questions. Unlike spoken communication, inner speech does not involve any movement of the articulators. Instead, it relies on the voluntary imagination of speech, where an individual can mentally simulate the act of speaking without vocalizing or engaging any part of their vocal tract. Therefore, lack of articulators' engagement does not hinder inner thinking, provided that the cognitive function of the speaker remains intact.

The ability to decode and externalize human thought, commonly referred to as "mind reading" (i.e. the interpretation of internally-generated speech) has been a long-held ambition of Brain-Computer Interfaces (BCIs). Imagined speech has recently been studied as an intuitive paradigm (Herff et *al*., 2017), where the neural responses that are generated by imagining pronunciation, without the involvement of the articulators, are registered and subsequently decoded. This paradigm is particularly suitable for building communication systems; its performance may be lacking compared to other BCI paradigms, yet it has multiclass scalability, which allows building extensible BCI systems. BCIs are systems designed to translate neural activity into commands that control external devices, initially developed to assist individuals who have lost voluntary muscle control in tasks such as operating wheelchairs or robotic arms. In recent years, BCIs have expanded into various fields, including interactive applications in gaming and neuroergonomics. BCIs utilize different neuroimaging modalities, with electroencephalography (EEG) emerging as the preferred choice, as EEG stands out as an non-invasive and low-cost method compared to alternatives like fMRI and fNIRS. While it may offer lower spatial resolution, EEG excels in capturing rapid brain activity changes, providing high temporal resolution crucial for real-time applications.

Over the years, the development of EEG-based BCIs for inner speech decoding has been based on the systematic exploration of diverse imagined speech elements, including phonemes, syllables, and lexical items. Consequently, the corresponding classification task requires the interpretation of the brain's neural responses during the processing of these elements. The majority of the first inner decoding studies were usually confined to binary classification tasks, with the scope of differentiating between different types of speech elements, such as consonants and vowels, nasals and bilabials, or phonemes and words. Recently, research progressed to focus on multiclass classification issues, with the objective to distinguish among a wide variety of words or to identify specific phonemes from a larger set.

The variety of speech components investigated reflects the complexity of inner speech, as it encompasses not only individual speech elements but also more structured linguistic phenomena. By examining these different speech elements, researchers aim to refine and enhance the accuracy and scalability of BCI systems designed to interpret the brain's neural signals, with the ultimate goal being the formulation of a universal lexicon. Formulating such a lexicon requires the development of decoding models that should be able to handle multiclass data and differentiate them efficiently in (near) real-time to enable seamless communication and control for users.

Similarly to other signal processing domains, the field of decoding inner speech from EEG signals has made significant advancements in the last years due to the shift towards deep learning techniques in data processing. In more detail, deep learning techniques such as convolutional neural networks (CNNs)

and recurrent neural networks (RNNs) have demonstrated significant advantages in capturing temporal and spatial dependencies within EEG signals, surpassing traditional machine learning approaches in both accuracy and performance. In comparison to traditional machine learning methods, which have often struggled with the complexity of inner speech signals, deep learning models are better equipped to handle the variability and subtlety inherent in this task. As the field progresses, deep learning is expected to play a central role in advancing the accuracy, efficiency, and practical application of inner speech-based BCIs.

As research in this area continues to advance, deep learning is estimated to play a pivotal role (see Figure 1) in improving the practical application and deployment of inner speech-based BCIs, offering substantial benefits in the accuracy of inner speech interpretation and their incorporation to real-life applications. In this direction, the present report discusses in detail the deep learning inner speech decoding approaches available during the past decade.

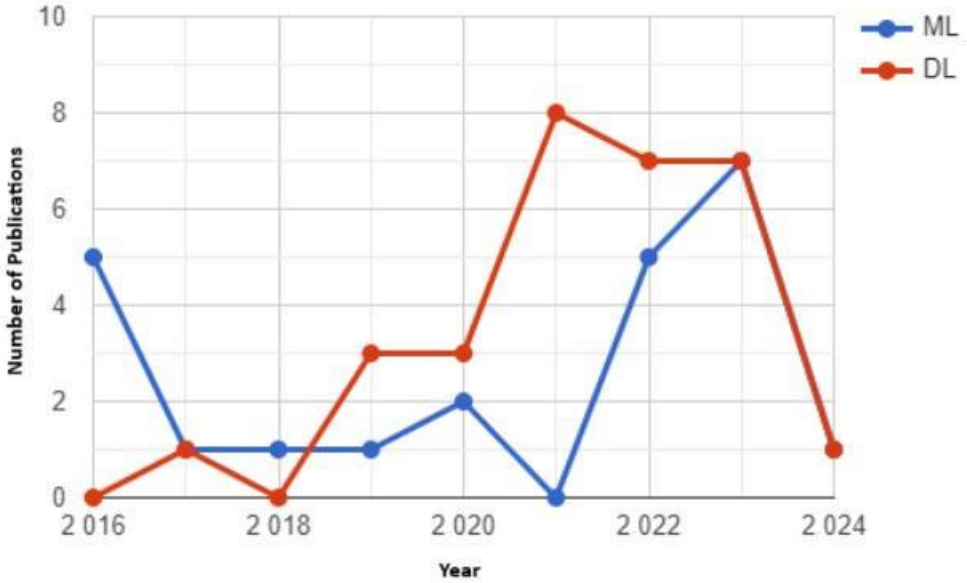

Figure 1. Number of publications per year based on Machine Learning (ML) and Deep Learning (DL) for EEG-based inner speech decoding.

The remainder of this document as follows: Section 2 presents the methodology employed; Section 3 presents the relevant datasets currently available; Section 4 describes the preprocessing steps for EEG data analysis; Section 5 details deep learning; and lastly, Section 6 concludes the report identifying the gaps that persist in inner speech research, which BINGO will address.

2 METHODOLOGY FOR LITERATURE SEARCH

We conducted the literature review for the present report with the aim to identify peer-reviewed research papers on the classification of inner speech using EEG with added attention to deep learning methodologies. The following steps outline the process followed to ensure a comprehensive and systematic search of relevant literature.

2.1 SEARCH DATE AND PLATFORMS

The search for relevant articles was conducted on February 5, 2024. The primary platform used for this search was Google Scholar, which aggregates content from a wide array of reputable academic sources. Specifically, the search included journals and conference papers available through IEEE Xplore, MDPI, ScienceDirect, IOPscience, Frontiers and SpringerLink. To capture relevant research, we used a combination of the following keywords: EEG, inner speech, covert speech, electroencephalography, classification and deep learning.

2.2 INCLUSION CRITERIA

The following inclusion criteria were considered for the present report:

- *Period*: The search confined to articles published between 2014 and 2024.
- *Language*: Only publications written in English considered.
- *Peer-Reviewed Articles*: Only peer-reviewed journal articles and conference papers considered to ensure the quality and reliability of the findings.
- *EEG-Related Studies*: Articles that specifically involved EEG data.
- *Deep Learning Focus*: Review studies employing deep learning techniques for the classification of inner speech.

The search process resulted in 30 papers focusing on deep learning and fitting all criteria which were considered for the literature review. We organize the review as follows: Section 3 discusses the publicly available datasets utilized in this field, detailing the essential characteristics and descriptions of each dataset. Section 4 outlines the preprocessing steps employed by researchers, including signal filtering, channel selection and feature extraction techniques. Section 5 provides an in-depth analysis of the deep learning architectures that have emerged in recent years, presenting findings on the trends and performance of the models used in inner speech classification. Lastly, Section 6 concludes the review underlying the main gaps in the literature.

3 DATASETS

In this section, we describe the open databases referenced in the literature. Each database consists of multiple tokens, each pronounced silently by the participants while their brain activity is monitored. The following inner speech tokens have been studied in the literature: phonemes (e.g. consonants and vowels), syllables, words and phrases. Note that some studies further examine features of lexical items such as word length or set of phrases for multiclass classification (Lee et *al*., 2020). Classification tasks are categorized in binary and multiclass classification. Binary classification includes, recognizing phonological categories, differentiating between short and longer words or classifying between token pairs. Meanwhile, multiclass is performed on a set of tokens and can include sets of words, syllables, or a combination of both. Bellow we describe each dataset:

KaraOne – Zhao & Rudzicz (2015)

The KaraOne dataset (Zhao and Rudzicz, 2015), developed at the Toronto Rehabilitation Institute, is designed for inner speech decoding using EEG signals. It includes data from 14 participants (4 females and 10 males), who were administered trials involving seven phonemic/syllabic tokens (/iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/) and four words (pat, pot, knew, gnaw). Each token was repeated 12 times per subject, resulting in a total of 132 trials per participant. EEG signals were recorded using a 64-channel cap, following the 10-20 electrode placement system, with a 1 kHz sampling rate. The acquisition protocol for each trial includes four states: a 5-second rest period, a 2-second stimulus presentation, a 5-second speech imagery phase, and a speaking state in which the participant spoke the token aloud, with both audio and facial features recorded. The dataset has been utilized in multiple binary classification tasks, including the detection of consonants, bilabials, nasals, single vowels, and the classification of words *vs*. phonemes (Macías-Macías et *al*., 2023; Panachakel & Ramakrishnan, 2021). Additionally, Datta & Boulgouris (2021) explored the classification of the grammatical class of tokens. Bakhshali et *al*. (2020) employed the dataset to classify all different pairs of the four spoken words. Research on multiclass classification has also been conducted utilizing all 11 classes (Hernandez-Galvan et al., 2022; Mini et al., 2021; Panachakel et al., 2019). Due to the limited number of trials per participant and class, the dataset may pose challenges for subject-dependent classification tasks using deep neural networks. This constraint suggests that approaches such as few-shot learning could be particularly well suited to address the issue, as they are designed to perform effectively with limited training data.

Nguyen et *al*. (2017)

The study involved 15 healthy participants (4 females and 11 males) performing imagined speech tasks. The tasks included imagining short words (’in’, ’out’, ’up’), long words (’cooperate’, ’independent’), and vowels (/a/, /i/, /u/). Each subject participated in sessions where they silently pronounced these prompts following 4 visual cues on a computer monitor. Each session consisted of 100 trials per prompt, with a rhythmic beep to guide the timing of the imagined speech. EEG signals were recorded using a 64-electrode system, sampled at 1000 Hz and downsampled to 256 Hz. The data underwent preprocessing with a bandpass filter (8-70 Hz), a notch filter at 60 Hz, and artifact removal for eye movements. In recent work, the dataset was shown to be suitable for both binary and multiclass classification tasks such as

classification between two long words (Biswas & Sinha, 2022; Hernandez-Galvan et *al*., 2023; Jiménez-Guarneros & Gómez-Gil, 2021; Kamble et *al*., 2023) or of short *vs*. long words (Hernandez-Galvan et *al*., 2023; Kamble et *al*., 2023). Furthermore, multiclass classification has been performed on 8 classes (Hernandez-Galvan et *al*., 2023) as well as between three short words and three vowels (Hernandez-Galvan et *al*., 2023; Kamble et *al*., 2023).

Coretto et *al*. (2017)

The dataset includes EEG recordings from participants performing imagined and pronounced speech tasks. 15 participants (7 females and 8 males) were involved in the study, each undergoing a single recording session to reduce intra-subject variance. The dataset consists of two main classes of tokens: Spanish vowels (/a/, /e/, /i/, /o/, /u/) and command words ('arriba' "up", 'abajo "down", 'derecha' "right", 'izquierda' "left", 'adelante' "forward", and 'atras' "backward"). Each participant performed two types of tasks: imagining and pronouncing the tokens. For each item, 50 trials were conducted, with 40 trials dedicated to imagined speech and 10 to pronounced speech. EEG signals were recorded using an 18-channel system, with electrodes positioned according to the 10-20 international system. The data were collected at a sampling rate of 1024 Hz. During preprocessing, a band-pass filter between 2 Hz and 40 Hz was applied to the EEG signals to eliminate low frequency noise and high frequency artifacts. Additionally, a fourth-order Butterworth low-pass filter with a cutoff frequency of 10 kHz was used for voice signals. Cooney et *al*. (2019) further examined the classification of all word-pairs included in the set. Additionally, the multiclass problem has also been explored using separately the available words (Lee et *al*., 2020; Simistira Liwicki et *al*., 2022), the vowels (Mahapatra & Bhuyan, 2022a; Sarmiento et al., 2021; Simistira Liwicki et *al*., 2022) and all the 11 classes together (Mahapatra & Bhuyan, 2022b).

Torres-García et *al*. (2012)

The dataset is composed of EEG signals collected from 27 native Spanish-speaking participants, recorded using a 14 electrode headset with a sampling frequency of 128 Hz. The data are filtered using a finite impulse response (FIR) band-pass filter at the range 4 to 25 Hz. Each subject performed 33 inner speech trials of each one of the 5 Spanish words 'arriba', 'abajo', 'izquierda', 'derecha', 'seleccionar' - "up", "down", "left", "right", "select" respectively. The classification of these 5 words has been explored by Jiménez-Guarneros & Gómez-Gil (2021) and Torres-García et *al*. (2016). García-Salinas et *al*. (2019) initially used the data of 4 classes and with transfer learning experimented on increasing the vocabulary of the model by adding the remaining class to the model.

Thinking out loud – Nieto et *al*. (2022)

In this study, 10 healthy participants (4 females and 6 males) engaged in an experiment involving 4 Spanish words: 'arriba' "up", 'abajo' "down", 'derecha' "right", and 'izquierda' "left". Each participant performed 50-60 trials for each word. EEG data were collected using a 128-channel setup with a sampling rate of 1024 Hz. The EEG signals were subsequently preprocessed and filtered within a frequency range of 0.5 Hz to 100 Hz, with a Notch filter applied at 50 Hz to eliminate power line noise. The data were then downsampled to a final sampling rate of 254 Hz, and Independent Component Analysis (ICA) was employed for artifact removal. This dataset is particularly suitable for 4-way classification of the words; additionally, the extensive number of trials per participant facilitates the development of subject-dependent models using deep learning techniques (van den Berg et *al*., 2021).

Sarmiento et *al*. (2021)

The dataset comprises EEG recordings from 50 native speakers of Spanish (20 females and 30 males). The experiment involved imagining five vowels (/a/, /e/, /i/, /o/, /u/), with each vowel constituting a separate class. For each vowel, participants completed 25 trials, resulting in 125 trials per participant. EEG data was collected using a headset with 14 electrodes with a sampling frequency of 128 Hz. The researchers also explored the classification of the data using convolutional neural networks.

LaRocco et *al*. (2023)

The dataset includes EEG recordings from 16 participants (4 females and 12 males). While most participants were native speakers of English, 5 were non-native speakers but demonstrated functional proficiency in English. The study involved participants imagining the pronunciation of 44 different phonemes presented through a combination of visual and auditory stimuli to mitigate distractions. Each participant underwent three 40-minute sessions, each involving five trials per phoneme, resulting in 15 trials per phoneme across all sessions. Data acquisition was performed using a headset capturing signals from 16 EEG channels at a sampling rate of 250 Hz. The EEG channels followed the 10–20 International System. The dataset has been used for a one-versus-all phoneme-based classification and it is also suitable for research on multiclass based models due to its large number of classes.

Da Salla et *al*. (2009; 2012)

Three right-handed fluent speakers of English with no neurological disorders (1 female and 3 males) participated in the study. The participants performed three tasks, imagined mouth opening and vocalization for vowel /a/, imagined lip rounding and vocalization for vowel /u/, and a control task (rest state) with 50 trials per task (150 trials total per subject). Continuous EEG was recorded using a system with 64+8 electrodes at 2048 Hz, downsampled to 256 Hz, and electrodes were placed according to the 6 10-20 system. Data were visually inspected, and trials with artifacts were marked for rejection and repeated if necessary. Preprocessing involved zero-phase bandpass filtering (1–45 Hz) and extracting 3-second epochs (1 second pre-stimulus, 2 seconds stimulus), resulting in 150 epochs per subject.

ZuCo (Hollenstein et al., 2018; Hollenstein et al., 2019)

The ZuCo dataset comprises EEG and eye-tracking data from 12 right-handed native speakers of English (5 females and 7 males). The dataset includes sentences from the Stanford Sentiment Treebank – 400 sentences: 123 neutral, 137 negative, 140 positive – and the Wikipedia relation extraction corpus – 650 sentences for normal reading and 407 for task-specific reading with partial overlap. Participants performed three reading tasks: (a) normal reading of sentiment-labeled sentences, with control questions about movie quality; (b) reading sentences with semantic relations, followed by multiple-choice questions; and (c) task-specific reading to identify specific relation types – e.g., award, education, employer). Sentences were presented on a computer screen in blocks, with text formatted for natural reading. Participants used a control pad to navigate the sentences at their own pace. EEG data were recorded using a 128-channel system at 500 Hz. Data preprocessing included artifact removal, band-pass filtering (0.1-100 Hz), and impedance checks every 30 minutes. The study provides insights into natural reading processes and the neural correlates of semantic and sentiment processing. The ZuCo dataset's extensive vocabulary and diverse corpus make it suitable for training models aimed at open EEG-to-text decoding, as demonstrated by Wang & Ji (2022) and Duan et *al*. (2023).

4 PREPROCESSING

In this section, we briefly go through the key preprocessing techniques are used in the literature.

4.1 FILTERING

Filtering of the signals is one of the most important steps of preprocessing. Band pass filters are used for noise reduction in different frequency bands. Most popular bands that are used for EEG signals are between 0.5-50Hz (Bisla & Anand, 2023; Macías-Macías et *al*., 2023;Tiwari et *al*., 2024; Zheng et *al*., 2023) while there are some previous works that expand their filtering and keep frequencies up to 128Hz (Abdulghani et *al*., 2023; LaRocco et *al*., 2023; Pawar & Dhage, 2023). Each of the main frequency bands that are relevant to EEG are associated with specific mental tasks. Hernandez-Galvan et *al*. (2023) performed connectivity analysis on EEG data and concluded that the beta band is the most efficient in classifying mental tasks. Finally, the 50 Hz frequency called powerline frequency is usually filtered out, as it contains noise that interferes with the data quality, using a notch filter at 50 Hz (Bakhshali et *al*., 2022; Tiwari et *al*., 2024).

4.2 ARTIFACT REMOVAL

The first step in signal preprocessing is removing noise and artifacts. Various methods are employed to achieve this, with blind source separation (Bisla & Anand, 2023) and Wavelet Enhanced Independent Component Analysis (WICA) (Mini et *al*., 2021) being among the most effective techniques. Blind source separation works by decomposing the EEG signals into statistically independent components, allowing for the identification and removal of artifacts such as eye blinks and muscle movements. WICA combines wavelet transformation with ICA, leveraging the strengths of both methods to enhance the separation of artifacts from neural signals. This dual approach helps in effectively isolating and eliminating non-neural signals, thereby significantly improving the clarity and quality of the EEG data. Other basic techniques such as normalization, average re-referencing of the data and ICA have also been used with the aim to increase the performance of the models (Abdulghani et *al*., 2023; Kamble et *al*., 2023; Retnapandian & Anandan, 2023).

4.3 CHANNEL SELECTION

Inner speech involves specific brain regions and neural mechanisms that are typically different from the ones involved in overt speech, as inner speech is associated primarily with brain areas involved in speech perception (Alderson-Day et *al*., 2015; Alderson-Day et *al*., 2016; McGuire et *al*., 1996 among others). The goal in channel selection is to identify and focus on the EEG channels that capture the most relevant neural activity associated with inner speech, thus enhancing the performance of the BCI. The following brain regions are consistently implicated in inner speech processing:

- *Broca's area* located in the left frontal lobe is associated with speech production and EEG channels over the left frontal regions are often selected to capture activity in the area.
- *Prefrontal cortex* is involved in executive functions and working memory and helps in formulating inner speech.
- *Wernicke's area* is located in the left superior temporal gyrus and is involved in language comprehension.
- *Parietal lobe* integrates sensory information and is involved in spatial processing and attention, which are relevant to inner speech.

4.4 FEATURE EXTRACTION

Feature extraction plays a pivotal role in the analysis of EEG signals for inner speech classification, as it aims to capture relevant patterns and characteristics from the data. Earlier research has explored various techniques to extract discriminative features from EEG signals, catering to different aspects of neural activity. Statistical features have been widely utilized (Macías-Macías et *al*., 2023; Pawar & Dhage, 2023; Zheng et *al*., 2023), including measures such as mean, variance, skewness, and kurtosis, which provide insights into the distribution and dynamics of the EEG signals. Riemannian features offer an alternative approach, leveraging the geometry of covariance matrices to represent the intrinsic structure of EEG data (Nguyen et *al*., 2017; Bakhshali et *al*., 2022). Wavelet-based methods (Abdulghani et *al*., 2023; Mahapatra & Bhuyan, 2023) exploit the multi-resolution nature of wavelet transforms to extract time-frequency features, capturing both temporal and spectral information from EEG signals. Additionally, spatiotemporal features (LaRocco et *al*., 2023) consider the spatial distribution and temporal dynamics of EEG signals, offering a comprehensive representation of neural activity across different brain regions.

5 DEEP LEARNING

Deep Learning, an advanced subdomain of machine learning inspired by the neural mechanisms of the human brain, has emerged as a pivotal framework for identifying complex structures within high-dimensional data. Through the utilization of multilayered neural architectures, it systematically abstracts hierarchical representations from unstructured inputs, including acoustic signals and textual data, thereby enabling substantial advancements in domains necessitating sophisticated analytical capabilities (Zhang et *al*., 2022). This paradigm exhibits a high degree of adaptability, leveraging large-scale datasets to capture effectively contextual subtleties (Li & Wang, 2023). Its ability to process intricate patterns has facilitated transformative developments across a broad spectrum of disciplines, reinforcing its position as a fundamental component of contemporary artificial intelligence research (Devlin et *al*., 2021).

This section presents deep learning architectures and relevant data based on 30 research articles discussing current trends in the field and models' performance. 16 studies employ CNNs to classify inner speech; 5 studies employ RNNs with 2 of them examining both CNNs and RNNs for classification; lastly, 11 studies incorporate various architectural types such as Multi-Layer Perceptrons (MLPs), siamese networks, capsule networks, shallow neural networks, and Large Language Models (LLMs). Note that other important characteristics other than the network type, are activation functions and optimization algorithm.

5.1 CNNS

CNNs are particularly favored for EEG analysis due to their efficacy in capturing spatial hierarchies in data, making them well suited for processing the complex patterns inherent in EEG signals. EEG signals are recorded from multiple electrodes placed across the scalp, creating a spatially structured dataset. CNNs use convolutional layers to focus on local spatial features, allowing the network to learn important patterns between neighboring electrodes. Multiple studies employ either standalone models or hybrid models paired with other methods, consistently achieving high accuracies in the field.

Basic CNN open models

CNN architectures typically consist of a series of convolutional and pooling layers followed by fully connected layers, effectively learning spatial features from the EEG data. These typical CNN representations have proven to be extremely effective for tackling complex problems. For EEG classification, a popular and effective CNN architecture is the EEGNet (Lawhern et al., 2021). EEGNet is particularly well-known for its lightweight and efficient architecture, specifically tailored for EEG-based applications. EEGNet employs depth-wise and separable convolutions, which significantly reduce the number of parameters while maintaining robust feature extraction capabilities. Various studies have used EEGNet based classification for the task of inner speech decoding. Within this context, EEGNet was used by van den Berg et *al*. (2021) to discriminate among 4 words resulting in an average accuracy of 29.67% while the best subject reached as high as 34.50%, indicating that performance is highly affected by individual differences and signal noise.

DeepConvNet authors (Schirrmeister et al., 2017) aimed to develop CNN-based models specifically tailored to classify between EEG signals. The output of this study was two widely used distinct architectures named Deep ConvNet and Shallow ConvNet. The Deep ConvNet comprises of 4 convolutional blocks aiming to extract the relevant features from the signals followed by a softmax

classification unit whereas Shallow ConvNet features a single convolutional layer followed by a temporal filter layer and is specifically tailored for decoding band power features. Cooney et *al*. (2020) have researched the effect of hyperparameters on such networks for imagined speech decoding. The EEGNet, Deep ConvNet and Shallow ConvNet models were tested with various hyperparameter configurations using nested Cross-Validation. The tests found a statistically significant the effect of various HP for DL models. These models are often referred to as benchmarks models in many studies.

For inner speech recognition specifically, the iSpeech-CNN model introduced by Simistira Liwicki et *al*. (2022) employs consecutive 2D convolutional blocks with varying filter numbers, followed by a softmax layer for multiclass classification. This model achieved accuracies of 35.20% for five vowels and 29.21% for six words, both significantly above chance. Furthermore, Nitta et *al*. (2023) tested five vowels using a network that consisted of serialized 2D convolutional layers followed by activation functions, dropout or fully connected layers. The model achieved on average 72.6% accuracy. Based on simple CNNs García-Salinas et *al*. (2023) introduced the ability for a pretrained CNN network to learn new previously unseen classes without it resulting in catastrophic forgetting. The newly introduced algorithm managed to maintain the total accuracy of the network even after introducing the new class, resulting in a stable model.

Advanced works

More advanced CNN architectures have been explored utilizing CNNs with more complex structures and/or aiming for a better understanding of brain signal patterns. Li et *al*. (2021) explored the hybrid-scale spatial-temporal dilated convolution network (HS-STDCN), which similarly to the EEGNet and DeepConvNet networks uses temporal and spatial convolution layers for extracting characteristics from the EEG signals. HS-STDCN's difference is that the temporal convolution is done by a hybrid convolutional layer instead of a standard one, utilizing multiple kernel sizes to capture features on different scales. The main addition of this work is the use of a dilated convolution block post spatiotemporal features extraction. The dilated block allows the network to capture multi-scale context thus improving contextual understanding over larger scales. Li et *al*. (2021) trained and tested their method, on a 8-word vocabulary dataset, and compared the findings with common benchmark algorithms, including various machine learning classifiers as well as the EEGNet, with the proposed algorithm achieving an average accuracy of 54.31%, significantly outperforming other methods.

Staying on the unique architecture area of DL studies, Tiwari et *al*. (2024) employed multi-headed CNNs for solving the classification of five vowels with each head trained on different extracted features from the signals and each output passed into a last fully connected later for the final classification. A 97.67% accuracy was reported, a measurable improvement over the 92.8% accuracy achieved by a 1D-CNN model. A similar approach, but with a multi-channel CNN model, was employed by Datta & Boulgouris (2021) who classified the grammatical class (verb/noun) of eight silently spoken words. Each of the three CNN channels was trained with data from different electrodes resulting in an accuracy of 85.7% even when the classified word was never used during training (Leave One Out validation). Similarly, Sarmiento et *al*. (2021), deployed an ensemble model to classify between 5 vowels. For each EEG input they trained 10 distinct CNNs in parallel, each corresponding to one of the possible vowel-pairs; the 10 outputs were subsequently passed to a voting scheme function, which outputs the final prediction. The model predicted correct output at 85.66% when tested on their publicly available database (see Section 3).

The studies reported so far used solely EEG data to train a classifier and decode the participant's inner speech. Meanwhile, Cooney et *al*. (2021) used EEG data along with fNIRS signals with two CNNs trained on each data type for feature extraction, and then both outputs were concatenated and used for classification. Classification tasks included (a) 4 different words and (b) 4 word-pairs – two-word phrases. The average accuracy of the bimodal model was 34.29% suggesting that such methods could help improve

the open problem of inner speech detection. Hybrid models have also achieved promising results; the plan is to use CNNs along with other DL models, to better capture the information hidden in the complex signals. In this research setting, four works appear to use such hybrid architectures in the field of inner speech classification and are presented bellow. Saha et *al*. (2019) used both the CNN architecture as well as LSTM, a popular type of RNN, in parallel in order to extract relevant features from the signals. Both of the outputs were concatenated and passed as a single input to a traditional machine learning classifier for classification. This hybrid model was tested in multiple binary classification problems using the KaraOne dataset, noting significant improvements in classification over previous SoTA approaches.

So far, the studies mentioned have tested their performance either on binary classification problems or on a vocabulary of 4-6 tokens (words or vowels). A larger vocabulary size was taken into account by Vorontsova et *al*. (2021) with participants instructed to pronounce 8 different words along with 1 pseudoword. The proposed network consisted of a CNN-based architecture named ResNet, that is commonly used for image classification, followed by an RNN consisting of multiple GRUs. This new approach successfully differentiated between the 9 tokens with 85% accuracy showcasing strong potentials. Mahapatra & Bhuyan (2022b) attempted an even larger vocabulary size with 11 different classes including both words and vowels. The researchers included the use of temporal convolutional networks (TCN) in the model's pipeline, responsible for learning temporal features. The TCN blocks are trained in parallel with CNN blocks and their outputs are taken into account for the final classification. An impressive accuracy of 96.49% was reported which is considerably higher compared to other models. In the interim, technology of transformers and self-attention become central in the field. Lee & Lee (2022) paired the EEGNet with a self-attention module for inner speech classification. A 13-class dataset (12 words/phrases + rest) was used for testing the proposed classifier and the research resulted in a 35.07% accuracy. Concerning this review, this is the largest close-word vocabulary used for inner speech classification.

5.2 RNNS

The nature of EEG data being sequential with respect to time makes the use of RNNs appealing in their processing. RNNs are particularly effective for tasks where the temporal dynamics and context of data are important, such as time series prediction, language modeling, and sequence classification. They are characterized by their ability to maintain a hidden state that captures information from previous time steps, which is then used to influence the processing of subsequent time steps. This makes them uniquely suited to handle data where the order of inputs is crucial.

Those characteristics drove researchers to include RNNs in the task of inner speech classification from EEGs. Besides the aforementioned works where RNNs were used along with CNNs for decoding, there is research work utilizing RNN-only architectures. Abdulghani et *al*. (2023) used a popular type of RNN, the LSTM, for the classification of 4 words. The architecture consisted of 3 LSTM layers with dropout layers in between that were responsible for capturing the relevant information. Training the LSTM models, resulted in a 92.5% accuracy for the 4 words. In 5 vowels classification, Retnapandian & Anandan (2023) employed CNN training with EEG signals and tested the performance in all five popular frequency bands. The study reported a 88.9% accuracy for the classification of the 5 vowels with theta and beta bands achieving the highest accuracies. Addressing the challenge of limited availability of EEG signal data, Hernandez-Galvan et *al*. (2023) introduced a novel approach using few-shot learning with a prototypical network combined with RNNs for classification. The prototypical network was chosen for its effectiveness in learning from a small number of samples. RNNs are employed to extract features from the EEG signals, which are subsequently embedded and classified using the prototypical network algorithm. The study

evaluates the performance using both LSTM and GRU as RNN architectures, achieving accuracies of 92.04% and 96% respectively.

5.3 OTHER ARCHITECTURES

In the literature, we also identify studies employing other architectures such as training an MLP for the purpose of inner speech classification. Sereshkeh et *al*. (2017) conducted research on 'yes' *vs*. 'no' *vs*. rest classification examining pair classification and 3-way classification. A multi-layer perceptron with a single hidden layer was developed and trained with the EEG data yielding a binary classification accuracy of yes' *vs*. 'no' at 63.2%. The task of rest *vs*. covert speech trial had a higher accuracy that reached 75.7% while the 3-way classification surpassed the chance level with an accuracy of 54.1%. Additionally, Kiroy et *al*. (2022) aimed to solve a more complex problem where the proposed MLP network aimed to classify between 7 classes consisting of 6 mentally pronounced words plus the rest state of the participants; the trained ANN was evaluated with 33–40% accuracy on classifying all seven classes.

Promising are also the results by Mini et *al*. (2021) who explored multiple feature extraction methods from the EEG signals and then used those features for training a MLP. Their 11-way classification task achieved 77.37% accuracy noting significant improvements over earlier studies and with a larger corpus. The same problem was challenged four years earlier by Panachakel et *al*. (2019); a NN was trained on wavelet features and managed to drastically outperform previous state-of-the-art results with 57.15% for all 11 classes. In a similar fashion, a shallow NN was tested for the classification between nasal and bilabial phonemes from the K1 dataset (Panachakel et *al*., 2021). The study surpassed the previous benchmark methods, with accuracies reaching up to 95% on some participants.

Recent work on EEG decoding of inner speech has explored further unique methods. Macías-Macías et *al*. (2023) using the K1 dataset examined the binary tasks of bilabial, nasal, consonant, and vowel detection. Using capsule neural networks, they attempted to classify the signals based on their statistical attributes. A Capsule Network is a type of neural network that addresses some limitations of traditional CNNs, mainly handling spatial information and relations among various parts of the signal or image. Their experimental results suggest successful classification in all binary paradigms with accuracy over 90%, reaching up to 94% in certain tasks. In addition, Mahapatra & Bhuyan (2023) deployed another distinct NN with a Siamese Neural Network utilized for the classification of 6 words. These type of networks perform well for a small number of classes due to the pairwise comparison. The idea is that two identical sub-networks are trained for pairs of input, extracting characteristics and measuring the similarity (distance) between them; the weights are updated so that same class inputs have a small distance, and different classes fall far from each other. Even with a small dataset, the Siamese Network managed to outperform previous methods by a margin of 10% in accuracy, achieving an accuracy of 31.4% for 6-way classification.

The above articles were based on supervised learning methods with the networks updating their weights based on the true class of the input with the aim to match inputs to the correct outputs. Mahapatra & Bhuyan (2022a) test a different type of learning, namely reinforcement learning, along with DNNs; a deep Q-network (DQN) was developed to classify signal inputs into 5 different vowels. This reinforcement learning method proved to have promising results as it achieved an accuracy of 81.69% when classifying all of the 5 vowels. Jiménez-Guarneros & Gómez-Gil (2021) addressed one of the most important issues in EEG classification, that of Inter-Subject Variability which results in EEG signals varying significantly between different individuals. They proposed a transfer learning technique named Unsupervised Domain Adaptation (UDA); a machine learning technique that adapts a model trained on a source domain (e.g., one subject's EEG data) to a target domain (e.g., another subject's EEG data) without requiring labeled data in the target domain. The Deep Learning-Based UDA utilizes deep learning models

for domain adaptation improving the ability to learn complex features. The proposed approach of Jiménez-Guarneros & Gómez-Gil (2021) was to utilize adaptive batch normalization on the target domain data and to introduce a novel loss function designed to match the distributions on the source and target domains. The novel method was tested using a bidirectional RNN (bi-RNN) consisting of two GRU layers. Two different datasets where used to test the performance of the proposed approach, with 5 words on the first and 2 long words on the second. To test the transfer learning ability of the model, the training data consisted of all but one subject, and the models performance was evaluated with the data of the remaining subject (leave-one-out cross validation). The binary classification yielded slightly above chance level results, but the 5-word classification achieved an average accuracy of 61.02% showcasing the potential of such methods.

5.4 LARGE/OPEN CORPUS WORKS

The vast majority of inner speech decoding from EEG literature consists of multi class classification of tokens, which was the primary focus of 28/30 articles included in the present review. Next, we briefly present different paradigms of inner speech classification based on the two remaining studies.

Krishna et *al.* (2020) performed continuous inner speech recognition employing EEG signals from a corpus consisting of 30 unique English sentences. Four participants read silently 30 unique sentences from the USC-TIMIT database (Narayanan et al., 2013). For classification, a Connectionist Temporal Classification (CTC) model was employed; a type of network used for sequence prediction problems where the alignment between the input and output sequences is unknown. Two GRUs were used for encoding the EEG signals and the encoded information is passed to a dense layer for decoding. The model is trained to minimize the CTC loss function link. The proposed model was evaluated using Word Error Rate (WER) as the suitable metric, noting a WER of 83.34 when using the 30 unique sentences. This result means that more than 4/5 predicted words are wrong considering the ground truth sequence. Even with such high error rates, this study is the first where continuous inner speech decoding is attempted and may pave the way for future research.

Recently, in a more challenging endeavor, Wang & Ji (2022) tackled the goal of enhancing current brain-to-text systems beyond limited closed vocabularies. Their study introduces an open vocabulary EEG-to-Text Sequence-to-Sequence decoding framework leveraging established language models. Utilizing the publicly available ZuCo EEG dataset (for details on dataset see Hollenstein et *al.*, 2018; Hollenstein et *al.*, 2019), which includes participants' brain activity while reading movie reviews; the study also captured eye movements using an eye tracker. The proposed architecture employs transformers, a powerful neural network architecture, to map EEG data into embeddings. These embeddings are subsequently inputted into a pre-trained Large Language Model (LLM) to generate tokens corresponding to decoded text. This novel idea by Wang & Ji (2022) managed to achieve best accuracy on the field of open-vocabulary EEG-to-text decoding, an area of BCI that is still very much in its infancy.

6 CONCLUSION

The application of artificial intelligence, particularly Deep Learning, to inner speech holds considerable promise as a transformative approach, offering novel pathways to decode and harness internal linguistic and cognitive processes. Its capacity to model complex patterns positions it as a powerful tool for the BCI field; yet significant gaps persist in inner speech research, which require further investigation:

i. *Challenges in Data Collection for Model Training*
   Studying inner speech presents difficulties in gathering the large datasets needed for Deep Learning. Unlike external speech, which benefits from vast audio databases, inner speech relies on indirect signals gathered via speakers' recruitment. Consequently, generating large-scale data remains challenging.
ii. *Obstacles to Real-Time Use*
   Using inner speech for instant applications, such as silent communication, requires fast processing. Considering that Deep Learning models are computationally demanding, and non-invasive neural signals are often noisy make the implementation difficult, highlighting the need for more efficient systems with higher accuracy.
iii. *Issues with Generalization Across Speakers*
   Inner speech varies widely due to differences in languages tested and classes selected in experimental design. Models trained on a small number of participants often struggle are limited when applied to broader populations of multiple language backgrounds. Developing adaptable systems that can handle this variation without constant retraining remains a key challenge.
iv. *Limited Use of Multimodal Data*
   Most research focuses on brain signals, but inner speech may also involve subtle physical cues, like eye movement activity. Combining these signals in Deep Learning models could improve accuracy, yet this area remains largely unexplored.
v. *Ethical and Transparency Concerns*
   Applying AI to internal thought processes raises serious ethical questions, especially regarding privacy and consent. Additionally, Deep Learning models often function as "black boxes," making it unclear how decisions are made. Ensuring both ethical standards and model transparency requires further development.

These gaps highlight directions for future research, which could include applying methods from spoken language analysis, creating artificial inner speech datasets, or working more closely with neuroscience. Solving these challenges could lead to major advances, such as silent communication systems or tools.

REFERENCES

Abdulghani, M. M., Walters, W. L., & Abed, K. H. (2023). Imagined speech classification using EEG and deep learning. *Bioengineering*, *10*(6), 649.

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, *141*(5), 931.

Alderson-Day, B., Weis, S., McCarthy-Jones, S., Moseley, P., Smailes, D., & Fernyhough, C. (2016). The brain's conversation with itself: neural substrates of dialogic inner speech. *Social cognitive and affective neuroscience*, *11*(1), 110-120.

Bakhshali, M. A., Khademi, M., & Ebrahimi-Moghadam, A. (2022). Investigating the neural correlates of imagined speech: An EEG-based connectivity analysis. *Digital Signal Processing*, *123*, 103435.

Bakhshali, M. A., Khademi, M., Ebrahimi-Moghadam, A., & Moghimi, S. (2020). EEG signal classification of imagined speech based on Riemannian distance of correntropy spectral density. *Biomedical Signal Processing and Control*, *59*, 101899.

Bisla, M., & Anand, R. S. (2023, August). Analysis of imagined speech characteristics using phase-based connectivity measures. In 2023 IEEE 13th International Conference on Control System, Computing and Engineering (ICCSCE) (pp. 68-73). IEEE.

Biswas, S., & Sinha, R. (2022). Wavelet filterbank-based EEG rhythm-specific spatial features for covert speech classification. *IET Signal Processing*, *16*(1), 92-105.

Cooney, C., Folli, R., & Coyle, D. (2021). A bimodal deep learning architecture for EEG-fNIRS decoding of overt and imagined speech. *IEEE Transactions on Biomedical Engineering*, *69*(6), 1983-1994.

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, 20(16), 4629.

Cooney, C., Korik, A., Raffaella, F., & Coyle, D. (2019, September). Classification of imagined spoken word-pairs using convolutional neural networks. In *The 8th Graz BCI Conference, 2019* (pp. 338-343). Verlag der Technischen Universitat Graz.

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.

DaSalla, C. S., Kambara, H., Koike, Y., & Sato, M. (2009, April). Spatial filtering and single-trial classification of EEG during vowel speech imagery. In *Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology* (pp. 1-4).

DaSalla, C. S., Kasahara, K., Honda, M., & Hanakawa, T. (2012). MRI correlates of mu rhythm activity during EEG-based brain–computer interface control. In *42nd Annual Meeting of the Society for Neuroscience, New Orleans, LA*.

Datta, S., & Boulgouris, N. V. (2021). Recognition of grammatical class of imagined words from EEG signals using convolutional neural network. *Neurocomputing*, *465*, 301-309.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2021). BERT revisited: Pre-training of deep bidirectional transformers for language understanding. *Journal of Artificial Intelligence Research*, 71, 1023-1056.

Duan, Y., Zhou, J., Wang, Z., Wang, Y. K., & Lin, C. T. (2023). Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*.

García-Salinas, J. S., Torres-García, A. A., Reyes-Garćia, C. A., & Villaseñor-Pineda, L. (2023). Intra-subject class-incremental deep learning approach for EEG-based imagined speech recognition. *Biomedical Signal Processing and Control*, *81*, 104433.

García-Salinas, J. S., Villaseñor-Pineda, L., Reyes-García, C. A., & Torres-García, A. A. (2019). Transfer learning in imagined speech EEG-based BCIs. *Biomedical Signal Processing and Control*, *50*, 151-157.

Herff, C., de Pesters, A., Heger, D., Brunner, P., Schalk, G., & Schultz, T. (2017). Towards continuous speech recognition for BCI. *Brain-Computer Interface Research: A State-of-the-Art Summary 5*, 21-29.

Hernandez-Galvan, A., Ramirez-Alonso, G., & Ramirez-Quintana, J. (2023). A prototypical network for few-shot recognition of speech imagery data. *Biomedical Signal Processing and Control*, *86*, 105154.

Hernandez-Galvan, A., Ramirez-Alonso, G., Camarillo-Cisneros, J., Samano-Lira, G., & Ramirez-Quintana, J. (2022, October). Imagined speech recognition in a subject independent approach using a prototypical network. In *Congreso Nacional de Ingeniería Biomédica* (pp. 37-45). Cham: Springer International Publishing.

Hollenstein, N., Rotsztejn, J., Troendle, M., Pedroni, A., Zhang, C., & Langer, N. (2018). ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific data*, *5*(1), 1-13.

Hollenstein, N., Troendle, M., Zhang, C., & Langer, N. (2019). ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*.

Jiménez-Guarneros, M., & Gómez-Gil, P. (2021). Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition. *Pattern Recognition Letters*, *141*, 54-60.

Kamble, A., Ghare, P. H., & Kumar, V. (2023). Optimized rational dilation wavelet transform for automatic imagined speech recognition. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1-10.

Kiroy, V. N., Bakhtin, O. M., Krivko, E. M., Lazurenko, D. M., Aslanyan, E. V., Shaposhnikov, D. G., & Shcherban, I. V. (2022). Spoken and inner speech-related EEG connectivity in different spatial direction. *Biomedical Signal Processing and Control*, *71*, 103224.

Krishna, G., Tran, C., Carnahan, M., & Tewfik, A. (2020). Continuous silent speech recognition using eeg. *arXiv preprint arXiv:2002.03851*.

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.

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.

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.

Lee, Y. E., & Lee, S. H. (2022, February). EEG-transformer: Self-attention from transformer architecture for decoding EEG of imagined speech. In *2022 10th International winter conference on brain-computer interface (BCI)* (pp. 1-4). IEEE.

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.

Li, X., & Wang, H. (2023). Adaptive deep learning frameworks: A survey of recent developments. *Artificial Intelligence Review*, 56(3), 1895-1924.

Macías-Macías, J. M., Ramírez-Quintana, J. A., Chacón-Murguía, M. I., Torres-García, A. A., & Corral-Martínez, L. F. (2023). Interpretation of a deep analysis of speech imagery features extracted by a capsule neural network. *Computers in Biology and Medicine*, *159*, 106909.

Mahapatra, N. C., & Bhuyan, P. (2022a, November). Decoding of Imagined Speech Neural EEG Signals Using Deep Reinforcement Learning Technique. In *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)* (pp. 1-6). IEEE.

Mahapatra, N. C., & Bhuyan, P. (2022b). Multiclass classification of imagined speech vowels and words of electroencephalography signals using deep learning. *Advances in Human-Computer Interaction*, *2022*(1), 1374880.

Mahapatra, N. C., & Bhuyan, P. (2023). EEG-based classification of imagined digits using a recurrent neural network. *Journal of neural engineering*, *20*(2), 026040.

McGuire, P. K., Silbersweig, D. A., Murray, R. M., David, A. S., Frackowiak, R. S. J., & Frith, C. D. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychological medicine*, *26*(1), 29-38.

Mini, P. P., Thomas, T., & Gopikakumari, R. (2021). EEG based direct speech BCI system using a fusion of SMRT and MFCC/LPCC features with ANN classifier. *Biomedical Signal Processing and Control*, *68*, 102625.

Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., ... & Proctor, M. (2014). *USC-TIMIT: A database of multimodal speech production data*. USC, Tech. Rep., 2013.[Online] http://sail. usc. edu/span/usc-timit/usctimit_report. pdf.

Nguyen, C. H., Karavas, G. K., & Artemiadis, P. (2017). Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of neural engineering*, *15*(1), 016002.

Nieto, N., Peterson, V., Rufiner, H. L., Kamienkowski, J. E., & Spies, R. (2022). Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition. *Scientific Data*, *9*(1), 52.

Nitta, T., Horikawa, J., Iribe, Y., Taguchi, R., Katsurada, K., Shinohara, S., & Kawai, G. (2023). Linguistic representation of vowels in speech imagery EEG. Frontiers in Human Neuroscience, 17, 1163578.

Panachakel, J. T., & Ramakrishnan, A. G. (2021, November). Classification of phonological categories in imagined speech using phase synchronization measure. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 2226-2229). IEEE.

Panachakel, J. T., Ramakrishnan, A. G., & Ananthapadmanabha, T. V. (2019, December). Decoding imagined speech using wavelet features and deep neural networks. In *2019 IEEE 16th India Council International Conference (INDICON)* (pp. 1-4). IEEE.

Pawar, D., & Dhage, S. (2023). EEG-based covert speech decoding using random rotation extreme learning machine ensemble for intuitive BCI communication. *Biomedical Signal Processing and Control*, *80*, 104379.

Retnapandian, A. S., & Anandan, K. (2023). Phoneme-based Imagined Vowel Identification from Electroencephalographic Sub-Band Oscillations during Speech Imagery Procedures.

Saha, P., Fels, S., & Abdul-Mageed, M. (2019, May). Deep learning the EEG manifold for phonological categorization from active thoughts. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2762-2766). IEEE.

Sarmiento, L. C., Villamizar, S., López, O., Collazos, A. C., Sarmiento, J., & Rodríguez, J. B. (2021). Recognition of EEG signals from imagined vowels using deep learning methods. *Sensors*, *21*(19), 6503.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., ... & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, *38*(11), 5391-5420.

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.

Simistira Liwicki, F., Gupta, V., Saini, R., De, K., & Liwicki, M. (2022). Rethinking the methods and algorithms for inner speech decoding and making them reproducible. *NeuroSci*, *3*(2), 226-244.

Tiwari, S., Goel, S., & Bhardwaj, A. (2024). Classification of imagined speech of vowels from EEG signals using multi-headed CNNs feature fusion network. *Digital Signal Processing*, *148*, 104447.

Torres-García, A. A. , 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.

Torres-García, A. A., Reyes-García, C. A., Villaseñor-Pineda, L., & García-Aguilar, G. (2016). Implementing a fuzzy inference system in a multi-objective EEG channel selection model for imagined speech classification. *Expert Systems with Applications*, *59*, 1-12.

van den Berg, B., van Donkelaar, S., & Alimardani, M. (2021, September). Inner speech classification using eeg signals: A deep learning approach. In *2021 IEEE 2nd International Conference on Human-Machine Systems (ICHMS)* (pp. 1-4). IEEE.

Vorontsova, D., Menshikov, I., Zubov, A., Orlov, K., Rikunov, P., Zvereva, E., ... & Bernadotte, A. (2021). Silent EEG-speech recognition using convolutional and recurrent neural network with 85% accuracy of 9 words classification. *Sensors*, *21*(20), 6744.

Wang, Z., & Ji, H. (2022, June). Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 5, pp. 5350-5358).

Zhang, Q., Liu, Y., & Han, J. (2022). Deep learning for multimodal signal processing: Emerging trends and applications. *IEEE Signal Processing Magazine*, 39(4), 25-37.

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.

Zheng, X. B., Ling, B. W. K., Zheng, S. Y., & Li, C. J. (2023). Supervised categorized principal component analysis for imagined speech classification via applying singular value decomposition on a symmetry matrix. *Biomedical Signal Processing and Control*, *86*, 105324.