IIST BCI Dataset-3 for 100 Malayalam Words

*Parvathy S S, *Nancy Sunil, †Shubham Tayade, **Chittaloori Likhitha, †S. Sumitra, and †B. S. Manoj
*A J College of Science and Technology, Thonnakkal, 695317, Trivandrum.
†Indian Institute of Space Science and Technology (IIST), Trivandrum, 695547, India.
**Chhattisgarh Swami Vivekanand Technical University(CSVTU), Bhilai, Chhattisgarh, 491107, India.
parvathypanicker01@gmail.com, nancysunil2000@gmail.com, shubham.sc20b123@ug.iist.ac.in,
likhithachittaloori11@gmail.com, sumitra@iist.ac.in and bsmanoj@iist.ac.in

***Abstract*—This paper introduces a dataset capturing brain signals generated by the recognition of 100 Malayalam words, accompanied by their English translations. The dataset encompasses recordings acquired from both vocal and sub-vocal modalities for the Malayalam vocabulary. For the English equivalents, solely vocal signals were collected. This dataset is created to help Malayalam speaking patients with neuro- degenerative diseases. This dataset not only contributes to the advancement of brain-computer interface technology but also holds promise in fostering effective communication solutions for individuals with restricted verbal abilities.**

I. Introduction

Communication is fundamental to human interaction, yet for individuals who are bedridden and unable to verbalize their needs or feelings, this basic human function becomes a significant challenge. In addressing this critical issue, non-invasive Brain-Computer Interfaces (BCIs) offer a promising avenue by leveraging EEG signals easily sensed from the scalp of patients. These signals, when processed and interpreted, can potentially provide a means for individuals to communicate without the need for traditional speech.

However, the development and refinement of BCI technology depend heavily on the availability of large, diverse datasets for training classifiers. Historically, datasets for BCI research have been predominantly available in widely spoken languages such as English [6] and Chinese [7], leaving a gap in accessibility for speakers of Indian languages.

Recognizing this gap and motivated by a desire to aid patients in Kerala, a state in southern India with 35 million population where Malayalam is predominantly spoken, we embarked on a pioneering initiative to create a comprehensive dataset in the Malayalam language. Authors of [3] and [4] introduced a dataset comprising 26 Malayalam words, demonstrating the feasibility and potential impact of their approach.

Building upon this foundation, we present an expanded dataset consisting of 100 Malayalam words, collected through both vocal and sub-vocal modalities. Additionally, English translations of these words were included to facilitate broader understanding and usage of the dataset within the research community. While only vocal signals were captured for the English words, this decision was made to ensure accessibility and ease of use for researchers unfamiliar with the Malayalam language.

Importantly, the 100 words selected for this dataset collection are the Malayalam equivalents of the same words used in [5] of a Marathi dataset. This ensures consistency and comparability across different languages, facilitating cross-linguistic research and benchmarking in BCI development.

By bridging the gap in BCI dataset availability for Indian languages, our research aims to not only advance BCI technology but also provide tangible benefits to individuals facing communication challenges, particularly within the Malayalam-speaking community of Kerala. This paper details our methodology, dataset collection process, and the potential implications of our work for both BCI research and practical applications in assistive communication.

II. Data Collection

We present the data collection methodology followed in the creation of this dataset. We discuss briefly the following: (a) Words Selection (b) Vocabulary Compilation (c) Electrode Arrangement and (d) Data Gathering Approach.

A. Words Selection

The selection of words for our dataset was aimed at ensuring relevance and accessibility for the target population. We identified commonly used Malayalam words to maximize utility for patients with communication challenges. By incorporating locally spoken words, we aimed to bridge the communication gap more effectively and cater to the specific linguistic needs of individuals in the region.

B. Vocabulary Compilation

We curated a list of 100 commonly used Malayalam words and their corresponding English translations to form the basis of our dataset. These words were carefully selected based on their simplicity, colloquial nature, and practical utility for addressing the daily needs of incapacitated patients. By prioritizing words that are familiar and relevant to the target population, we aimed to ensure the efficacy and applicability of our dataset in facilitating effective communication solutions for individuals with limited verbal abilities. Figures 1 and 2 present the comprehensive list of words utilized in the study, with Figure 1 depicting words numbered 1-50 and Figure 2 illustrating words numbered 51-100.

C. Electrode Arrangement

The placement of electrodes was crucial for obtaining accurate EEG signals. We utilized an OpenBCI Cyton board

Sl. No.	MALAYALAM WORDS	CORRESPONDING ENGLISH TRANSLATIONS
1.	അച്ഛൻ	PAPA
2.	അമ്മ	MUMMY
3.	സഹോദരൻ	BROTHER
4.	ഡോക്ടർ	DOCTOR
5.	വെള്ളം	WATER
6.	മരുന്ന്	MEDICINE
7.	തണുപ്പ്	COLD
8.	ചൂട്	HOT
9.	തലവേദന	HEADACHE
10.	പനി	FEVER
11.	പേശി	MUSCLE
12.	കൈ	HAND
13.	കാല്	LEG
14.	ആഹാരം	FOOD
15.	ഉയർത്തുക	LIFT
16.	വെയ്ക്കുക	KEEP
17.	നിർത്തുക	STOP
18.	സമയം	TIME
19.	കേൾക്കുക	LISTEN
20.	പാട്ട്	SONG
21.	സഹായം	HELP
22.	രാവിലെ	MORNING
23.	കുളിക്കുക	BATH
24.	കക്കൂസ്	TOILET
25.	വാതിൽ	DOOR
26.	തുറക്കുക	OPEN
27.	ഓഫ് ആക്കുക	SHUTDOWN
28.	പ്രവർത്തിപ്പിക്കുക	TURN ON
29.	നിർത്തുക	TURN OFF
30.	കാപ്പി	BREAKFAST
31.	പാൽ	MILK
32.	ചായ	TEA
33.	ജനൽ	WINDOW
34.	സൂര്യപ്രകാശം	SUNLIGHT
35.	ധ്യാനം	MEDITATION
36.	പഴങ്ങൾ	FRUITS
37.	ആശുപത്രി	HOSPITAL
38.	വീട്	HOME
39.	വിശ്രമം	REST
40.	അതെ	YES
41.	വേണ്ട	NO
42.	അനുഭവം	FEEL
43.	കണ്ണ്	EYES
44.	കുടുംബം	FAMILY
45.	ഇന്ന്	TODAY
46.	നാളെ	TOMORROW
47.	ഇന്നലെ	YESTERDAY
48.	അഴുക്ക്	DIRTY
49.	വൃത്തി	CLEAN
50.	സ്വപ്നം	DREAM

Fig. 1. List of words from 1 to 50.

Sl. No.	MALAYALAM WORDS	CORRESPONDING ENGLISH TRANSLATIONS
51.	ഓർക്കുക	REMEMBER
52.	മറക്കുക	FORGET
53.	മുകളിൽ	UP
54.	താഴെ	DOWN
55.	ശുഭം	AUSPICIOUS
56.	അപ്പൂപ്പൻ	GRANDFATHER
57.	അമ്മൂമ്മ	GRANDMOTHER
58.	തുണി	CLOTHES
59.	പണം	MONEY
60.	ശരീരം	BODY
61.	വൈകുന്നേരം	EVENING
62.	രാത്രി	NIGHT
63.	ഉച്ചകഴിഞ്ഞ്	AFTERNOON
64.	കട്ടിൽ	BED
65.	വിനോദം	ENTERTAINMENT
66.	പ്രകൃതി	NATURE
67.	പാത്രങ്ങൾ	UTENSILS
68.	കസേര	CHAIR
69.	ബന്ധു	RELATIVE
70.	ക്ലോക്ക്	CLOCK
71.	സോപ്പ്	SOAP
72.	പൈപ്പ്	TAP
73.	വ്യായാമം	EXERCISE
74.	ക്ഷീണം	WEAK
75.	സങ്കൽപ്പം	IMAGINE
76.	ഉറക്കം	SLEEP
77.	മുറി	ROOM
78.	തറ	FLOOR
79.	പച്ചക്കറി	VEGETABLES
80.	യാത്ര	TRAVEL
81.	വണ്ടി	VEHICLE
82.	മാംസഭുക്ക്	NON-VEGETARIAN
83.	സസ്യഭുക്ക്	VEGETARIAN
84.	തോന്നൽ	ILLUSION
85.	സമൂഹം	SOCIETY
86.	പുസ്തകം	BOOK
87.	ചോദ്യം	QUESTION
88.	ശീലം	HOBBY
89.	പാട്	HARD
90.	മനസ്സിലാക്കുക	UNDERSTAND
91.	കളഞ്ഞു	LOST
92.	ആലോചന	THINKING
93.	ഇരിക്കുക	SIT DOWN
94.	നിക്കുക	STAND UP
95.	ചിരിക്കുക	LAUGH
96.	കരയുക	CRY
97.	വികാരം	EMOTION
98.	തീരുമാനം	DECISION
99.	പ്രവേശനം	ENTRY
100.	അവസാനം	END

Fig. 2. List of words from 51 to 100.

[1] with 8 electrodes for data collection. Each electrode was differentiated accordingly with a different colored wire for ease of identification. Conductive gel was applied to each electrode to enhance conductivity and ensure reliable signal acquisition. Additionally, White medical tape was utilized to securely fixate the electrodes in their designated positions. This approach was adopted due to the unavailability of specialized equipment for electrode placement on the head.

Figure 3 provides a detailed illustration of electrode placement, depicting views from the front, back, left side, and right side for optimal channel positioning.

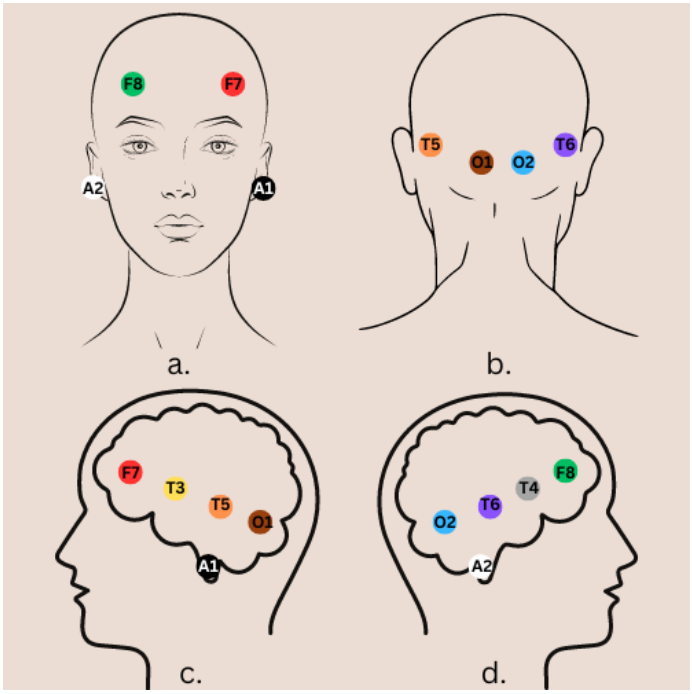

Fig. 3. Representation of electrode placement. The colored markers indicates the specific channels selected for EEG signal collection. a. frontal view, b. back view, c. left side view and d. right side view

The selected electrode positions corresponded to key cortical regions, serving specific functional roles in cognitive processes and neural activity. For instance, electrodes placed at the frontal sites (F7 and F8) capture activity associated with executive functions, decision-making, and emotional processing. Temporal electrodes (T3 and T4) record neural activity related to auditory processing and language comprehension. Occipital electrodes (O1 and O2) monitor visual processing and spatial awareness. Furthermore, the placement of electrodes on the earlobes (A1 and A2) served as reference points, ensuring accurate measurement of EEG signals relative to the surrounding cortical activity. Figure 4 provides detailed information on channel representation, delineating each channel's respective colors for ease of reference during data collection. By strategically positioning electrodes at these critical locations, we aimed to capture a comprehensive representation of neural activity, enabling detailed analysis and interpretation of EEG signals in relation to cognitive and perceptual processes. This electrode placement approach was essential for obtaining high-quality data for our study.

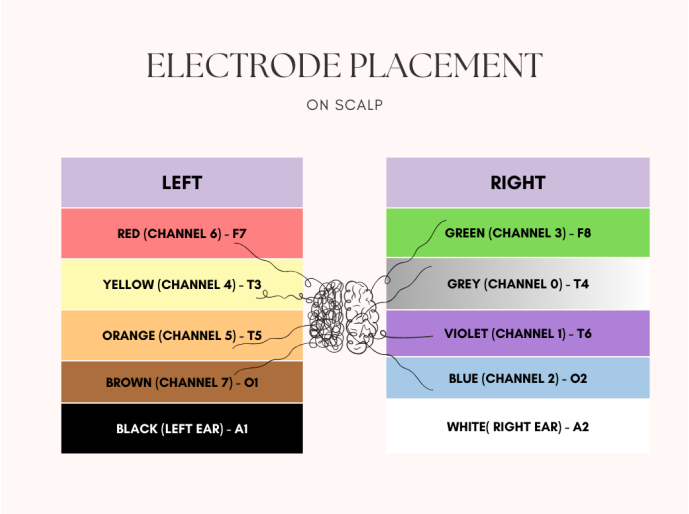

Fig. 4. Table representing electrode postioning.

D. Data Gathering Approach

The dataset collection process involved dividing the 100 words into four separate lists, each containing 25 words. Figure 5 represents how the data collection trials are conducted. This method facilitated systematic data collection and organization, ensuring efficiency and coherence throughout the process. The collection process lasted for two days, providing enough time for careful recording and verification of the EEG signals corresponding to each word.

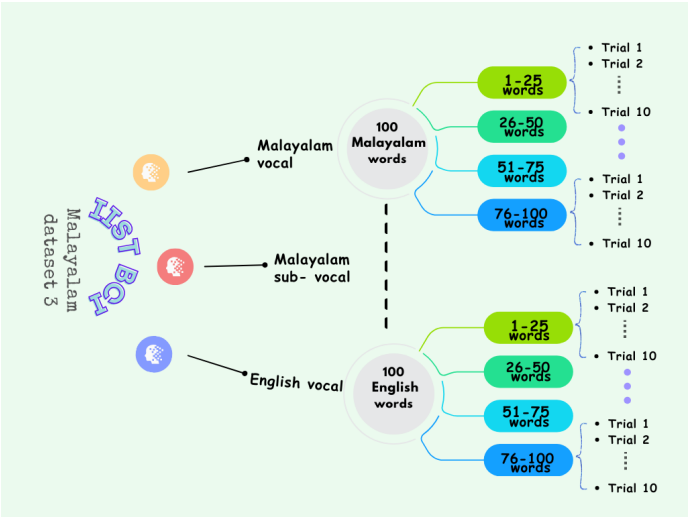

Fig. 5. Representation of how the data collection trials are conducted

The deliberate choice of a 22-year-old female volunteer aimed to understand potential variations between female and

male participants in neural responses during the articulation of Malayalam words. This decision allowed for comparative analysis across genders, as a male volunteer was selected for the dataset collections in (IIST BCI Malayalam Dataset-1 for Selected Common Malayalam Words and IIST BCI Dataset-2 for Selected Common Marathi Words) [3] and [5].

A Microsoft PowerPoint presentation containing Malayalam words with its English translation, provided a structured framework for the volunteer's engagement and pronunciation accuracy. The automatic slideshow, synchronized with predefined timings, facilitated the seamless articulation of each word. Figure 6 gives a detailed view of how the volunteers collected the dataset.

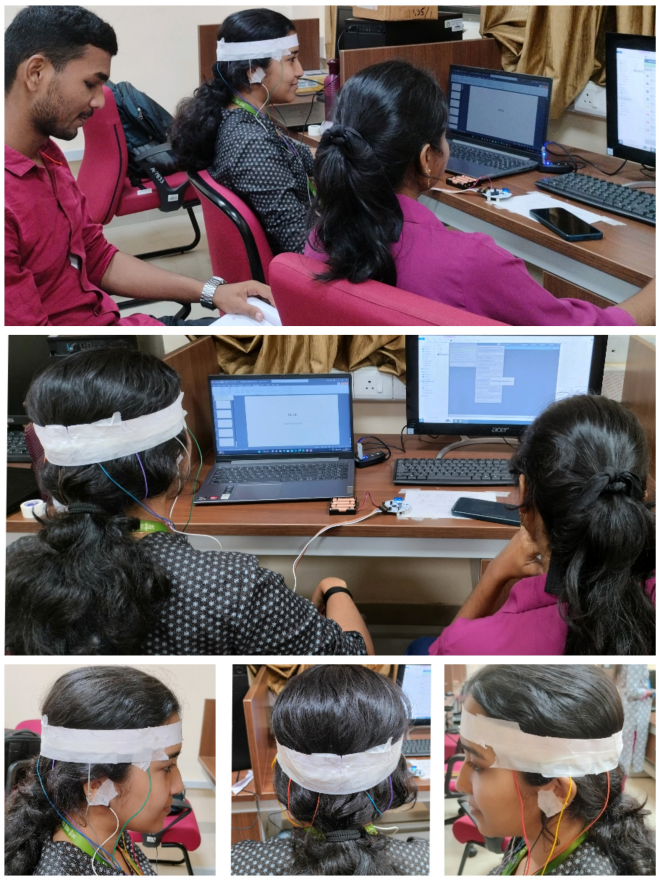

Fig. 6. The views of the volunteer with a white band on her head can be seen. The electrodes filled with gel are placed inside the band.

During the data collection sessions, EEG signals corresponding to spoken Malayalam words were acquired using the OpenBCI Cyton board, with eight channels carefully positioned on the volunteer's scalp. EEG recording commenced as each word appeared on the screen, capturing neural responses evoked by both the vocal and sub-vocal articulations. Visualization of EEG data acquisition, along with some features such as Time Series, Accelerometer, FFT Plot, and Head Map from Open BCI is shown in figure 7.

Before each recording session, a new session was started using the OpenBCI Graphical User Interface (GUI) to ensure the best possible signal acquisition conditions. The EEG recording stopped smoothly after the pronunciation of each word, resulting in ten distinct folders containing EEG data corresponding to the 100 Malayalam words.

A concurrent approach was employed for collecting vocal English words, ensuring a comprehensive and balanced dataset acquisition strategy across both languages.

III. Strategies to mitigate AC noise in OpenBCI Recordings

To minimize AC noise in OpenBCI recordings, several steps can be taken. Adjusting the notch filter to target the 50 Hz frequency can effectively filter out AC interference. Additionally, it is essential to keep other electronic devices away from the OpenBCI Cyton board to reduce electromagnetic interference. Ensuring that electrode cables are kept still during recording sessions can prevent mechanical noise artifacts. Finally, securely connecting electrodes to the scalp can improve signal quality by reducing impedance and preventing electrode movement. By implementing these measures, researchers can enhance the quality of EEG recordings and minimize the impact of AC noise on data analysis.

IV. Organization of Files of the Dataset

The dataset files contain raw data without preprocessing, with each file comprising several columns. The first column contains sample data, while columns 2 through 9 correspond to EEG recordings from 8 channels, capturing neural activity associated with articulating Malayalam words by the volunteer.

Columns 10 through 22 and 24 are considered irrelevant as they do not contain pertinent data for analysis. However, columns 23 and 25 are significant, containing data in Unix timestamp and formatted timestamp formats, respectively. These timestamp columns provide valuable information about the recording time of each sample data point in both machine-readable and human-readable formats.

Accurate organization and clear documentation of the dataset help researchers studying brain- computer interfaces and communication solutions for people who struggle with verbal communication.

Each GUI session has its directory, with every recording stored in separate files within that directory. Our dataset is organized into three main folders: Malayalam-Vocal, English-Vocal, and Malayalam Subvocal. Each folder contains four subfolders representing different sets of words: Set-1 (1-25 words), Set-2 (26-50 words), Set-3 (51-75 words), and Set-4 (76-100 words). Within each set, 10 trials were conducted for data collection.

V. Conclusion

This report describes a dataset containing brain signals generated by speaking 100 Malayalam words, alongside English translations. This dataset is helping bedridden individuals in communication. The methodology involved word selection,

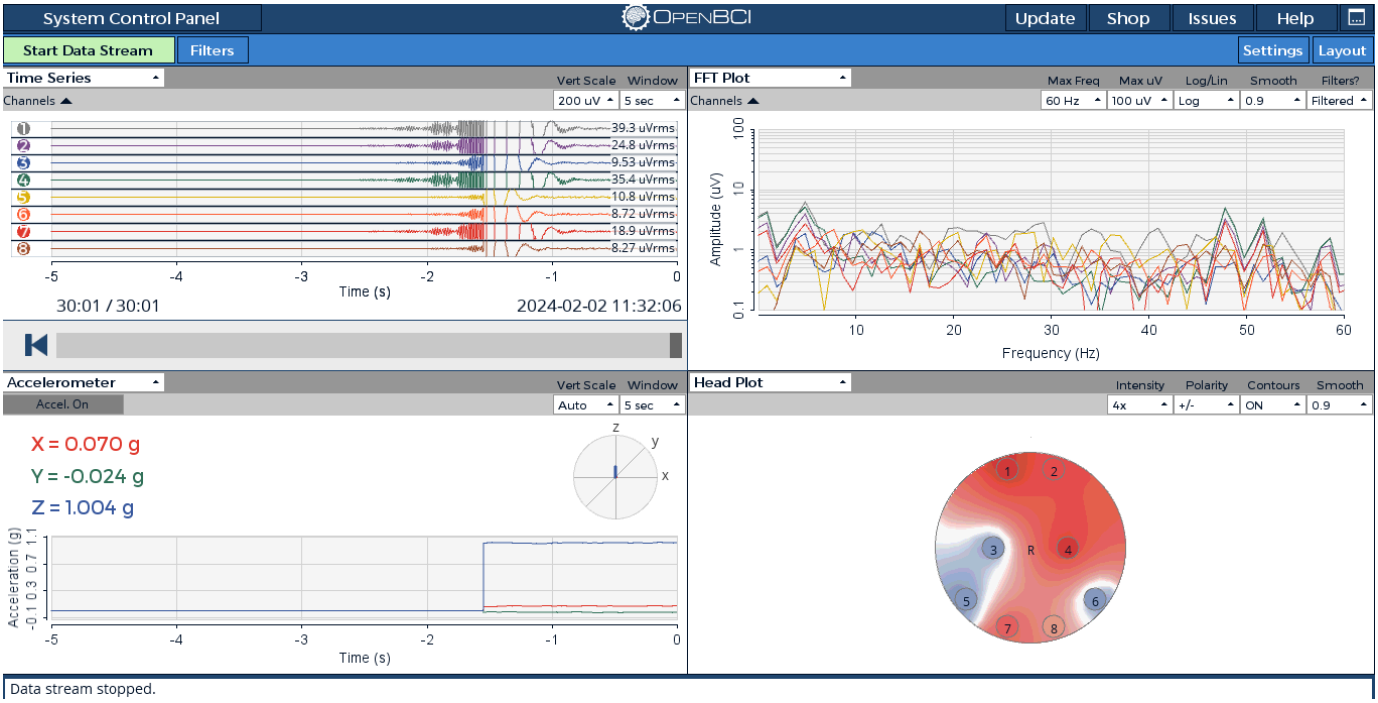

Fig. 7. This is a visualization of EEG data acquisition and analysis with Multi-Modal Insights such as Time Series, Accelerometer, FFT Plot, and Head Map Visualizations from OpenBCI 8-Channel EEG Recording

electrode placement, and data collection using an OpenBCI Cyton board with 8 electrodes. EEG recordings from 8 channels were collected for each word, with additional columns for timestamps. Organized into directories, each GUI session contains separate files for recordings, facilitating research on language processing dynamics. This dataset paves the way for emerging research in brain-computer communication.

Acknowledgements

We express our gratitude to the Director of the Indian Institute of Space Science and Technology (IIST), Trivandrum, and the entire administrative team for their steadfast support throughout this endeavor. This project was conducted during the internship program of the first four authors at the Brain-Computer Interface (BCI) Research group at IIST Trivandrum.

The dataset provided here is intended solely for academic and research purposes, without any implicit or explicit warranties or guarantees. Our motivation in sharing this dataset is driven by a commitment to advancing patient care. While we have taken steps to minimize errors, noise, and other potential issues, we acknowledge the possibility of unintended noises and ambient interference inadvertently affecting the dataset. It is important to recognize that the authors cannot be held responsible for any unintentional discrepancies found in the dataset. Therefore, users are advised to exercise caution and discretion when utilizing it.

For any further clarification or additional information not covered in this document, interested parties are encouraged to contact the authors. The corresponding author, B. S. Manoj, can be reached via email at bsmanoj@iist.ac.in.

References

[1] https://docs.openbci.com/Cyton/CytonLanding/
[2] https://en.wikipedia.org/wiki/10%E2%80%9320_system_(EEG)
[3] https://ieee-dataport.org/documents/iist-bci-dataset-1-selected-common-malayalam-words
[4] https://www.techrxiv.org/users/712048/articles/696682-iist-bci-dataset-1-for-selected-common-malayalam-words
[5] https://ieee-dataport.org/documents/iist-bci-dataset-2-selected-common-marathi-words
[6] DaSalla CS, Kambara H, Sato M, Koike Y. Single-trial classification of vowel speech imagery using common spatial patterns. Neural Netw. 2009 Nov;22(9):1334-9. doi: 10.1016/j.neunet.2009.05.008. Epub 2009 May 22. PMID: 19497710.
[7] Li Wang, Xiong Zhang, Xuefei Zhong, Yu Zhang, Analysis and classification of speech imagery EEG for BCI, Biomedical Signal Processing and Control, Volume 8, Issue 6, 2013, Pages 901-908, ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2013.07.011.
[8] Mariko Matsumoto, Junichi Hori, Classification of silent speech using support vector machine and relevance vector machine, Applied Soft Computing, Volume 20, 2014, Pages 95-102, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2013.10.023.
[9] https://en.wikipedia.org/wiki/Brain%E2%80%93computer_interface
[10] https://www.frontiersin.org/articles/10.3389/fnsys.2021.578875/full