IIST BCI Dataset-4 for Selected 100 Telugu words

Chittaloori Likhitha[1], Shubham Tayade[2], Parvathy S S[3], Nancy Sunil[3], Shivani Sahoo[2], S Sumitra[2], and B S Manoj[2]

[1]Chhattisgarh Swami Vivekanand Technical University(CSVTU), Bhilai Chhattisgarh
[2]Indian Institute of Space Science and Technology(IIST), Trivandrum
[3]A.J College of Science and Technology, Thonnakkal, Trivandrum

April 23, 2024

IIST BCI Dataset-4 for Selected 100 Telugu words

*Chittaloori Likhitha, †Shubham Tayade, **Parvathy S S, **Nancy Sunil, †Shivani Sahoo, †S.Sumitra, †B.S.Manoj
* *Chhattisgarh Swami Vivekanand Technical University(CSVTU), Bhilai, Chhattisgarh, 491107, India.*
†*Indian Institute of Space Science and Technology (IIST), Trivandrum, 695547, India.*
**A J College of Science and Technology, Thonnakkal, 695317, Trivandrum.*
likhithachittaloori11@gmail.com, shubham.sc20b123@ug.iist.ac.in, parvathypanicker01@gmail.com,
nancysunil2000@gmail.com, shivani.sc23m077@pg.iist.ac.in, sumitra@iist.ac.in, and bsmanoj@iist.ac.in

***Abstract*—To overcome the challenges faced by people with neurodegenerative diseases, Brain-Computer Interface (BCI) systems must make use of datasets relevant to patient's spoken languages. However, BCI research frequently faces setbacks due to the absence of such datasets for the target language population. This paper deals with the BCI datasets for one of the major Indian languages, Telugu, used by more than 90 million people, yet to obtain essential BCI datasets capturing the linguistic characteristics. To solve the unavailability of the Telugu BCI datasets, we created a dataset featuring EEG signal samples corresponding to frequently used Telugu words, aiming to fill this void and facilitate advancements in BCI research for native Telugu speakers. Through the utilization of Machine Learning and other classifiers, BCI systems can potentially translate and classify EEG signals into a display of or pronouncing Telugu words. Our dataset consists of both vocal and sub-vocal datasets of Telugu words and an English dataset of the corresponding English equivalents of the Telugu words. This IIST-BCI Dataset for Telugu is the first of its kind. It is dedicated to improving the accessibility of BCI technologies for Telugu-speaking individuals and fostering further research progress in this particular area.**

I. Introduction

Effective communication is often a challenge for patients with dementia, neurocognitive disorders, and neurodegenerative diseases. Their degree of care is greatly diminished by the communication challenges. There exists a possibility of translation of the Electroencephalograph (EEG) signals from these patients' scalps into meaningful speech or text by using non-invasive Brain-Computer Interfaces (BCIs). Better communication between patients and healthcare professionals can, therefore, be made possible by the ability to transform the scalp EEG signals into speech or display them on a screen.

Large datasets are needed in BCI research to build classifiers that reliably categorize gathered EEG data into particular phrases. Only a limited number of BCI datasets in Indian languages are available [1] [2] [3] [4]. In many Indian languages, there exist no BCI datasets. For example, Telugu, one of the major Indian languages spoken by around 90 million people, has no BCI datasets according to our knowledge and belief. The absence of a comprehensive research dataset, particularly for patients whose cognitive functions are performed with strong training in Telugu is one of the biggest obstacles to BCI research and solutions for Telugu speakers. This dataset development, which is detailed in this study, was motivated by this shortage of BCI research resources in Indian languages.

This paper describes the Telugu language BCI dataset to help develop tools and solutions to assist native Telugu-speaking individuals in coping with neurodegenerative conditions. The dataset comprises EEG data associated with both vocalized and subvocalized words. Incorporating sub-vocal speech data holds significant importance as it allows for developing classifiers capable of distinguishing between vocalized and sub-vocalized words. The sub-vocal dataset is beneficial for developing tools that can support patients who are unable to speak aloud. Leveraging this dataset, researchers can pioneer advancements in Brain-Computer Interface (BCI) solutions, thereby, improving communication of Patients.

The structure of the rest of this paper is as follows: Section II describes the processes employed for gathering the data. Subsequently, Section III clarifies the structure of files within the dataset. Finally, Section IV concludes the paper.

II. Data Collection Methods

In this section, we present the methods used in data acquisition for creating the dataset. We offer an outline of the (a) Arrangement of electrodes, (b) Data acquisition methods, and (c) The chosen list of Telugu words.

A. Arrangement of Electrodes

We utilized the OpenBCI Cyton Biosensing board [6], which offers either eight channels by default or up to sixteen channels with the extended board option. In our setup, we used 8 channels, each equipped with gold-plated electrodes strategically placed across different regions of the volunteer's scalp. Adhering to the widely recognized 10-20 electrode placement system, we carefully positioned electrodes at specific regions on the volunteer's head. On the right side, we placed electrodes at O2, F8, T4, and T6, while on the left side, we positioned them at O1, F7, T3, and T5. This strategic arrangement ensured accurate data collection during our experiment. Furthermore, Black and White color-coded electrodes were incorporated into the setup as reference electrodes. The reference electrode-1 (Black) was affixed to the left earlobe, while the reference electrode-2 (White) was affixed to the right earlobe. The BCI Cyton board assigned distinctive colors to the electrodes, simplifying the specification process throughout the experiment.

Figure 1 shows a pie Chart illustrating the electrode positions and their corresponding wire colors and channel num-

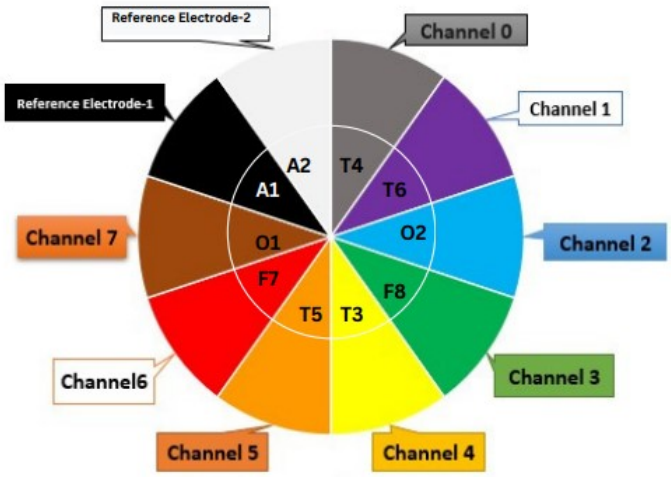

Fig. 1. Pie Chart showing the label of electrode positions and corresponding colored channels.

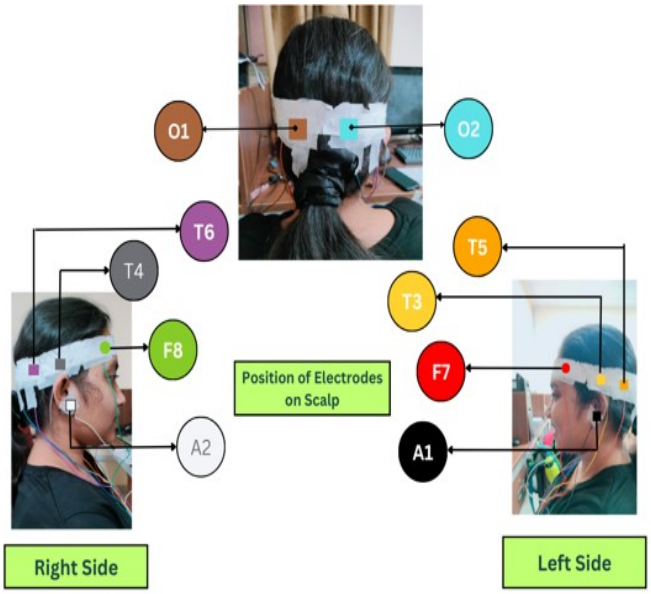

Fig. 2. Arrangement of electrode positions on the human scalp.

bers. The decision to use eight electrodes was driven by the need to develop affordable BCI solutions. The electrode locations were deliberately designed to replicate the areas of the scalp that an implantable Electro Encephalogram (EEG) device contacts when it is worn on the volunteer's head. This ensures optimal signal capture quality.

Figure 2 shows the positioning of electrodes on the human scalp. The signals collected from the occipital lobe are crucial for visual sensation, where electrodes O1 and O2 are positioned. Meanwhile, Signals from the frontal lobe, critical for speech function and proximate to the Broca area, are captured by electrodes situated at F7 and F8. Electrodes T6, T5, T4, and T3 capture signals originating from the temporal lobe. Particularly, electrodes T5 and T6 are strategically arranged near Wernicke's area, crucial for linguistic growth, within the rear view of the temporal lobe.

Conductive gel was applied to fill the gold cups of the electrodes, which were then secured to the scalp using tape at their designated positions. This application of conductive gel enhances measurement accuracy by minimizing the signal loss caused by resistance or improper contact with the gold cup.

B. Data Acquisition Methods

A 21-year-old Female, whose mother tongue is Telugu and proficient in English, volunteered for the data collection experiment. Electrodes were positioned on the volunteer's scalp according to Figure 2, following an 8-channel EEG configuration aligned with the standard 10-20 electrode placement system as shown in Figure 3. Additionally, both reference electrodes were placed on each side of the head, specifically on the earlobes. The female volunteer avoided using oil on the hair because it can disrupt EEG recordings by acting as an insulator. This precaution maintained recording accuracy, emphasizing the need for carefulness in EEG research to ensure reliable data.

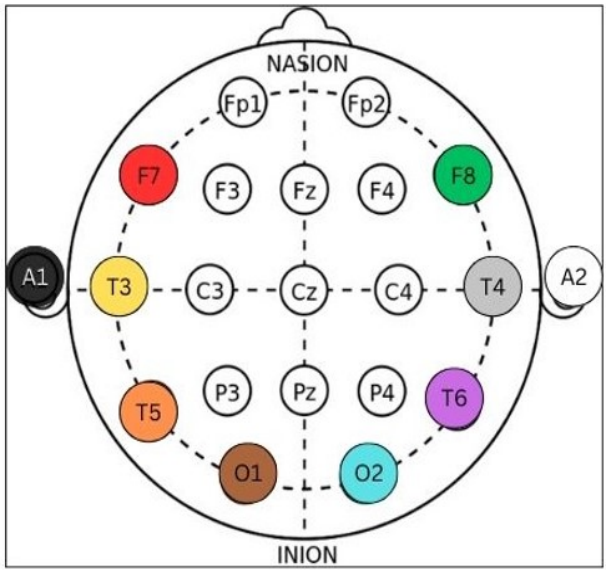

Fig. 3. The colored positions signify the channels chosen for gathering Electroencephalogram (EEG) signals through the international 10-20 system.

The dataset, comprising spoken and sub-vocalized Telugu and English words, was collected over several days from the volunteer. We separated the list of words into four sets of 25 words each. One day was used to record Set-1, which consists of words 1 to 25, and another day was used to record Set-2, which consists of words 26 to 50. Similarly, Sets 3 (51 to 75 words) and 4 (76 to 100 words) were recorded on different days.

Figure 4 illustrates the experimental setup involving the

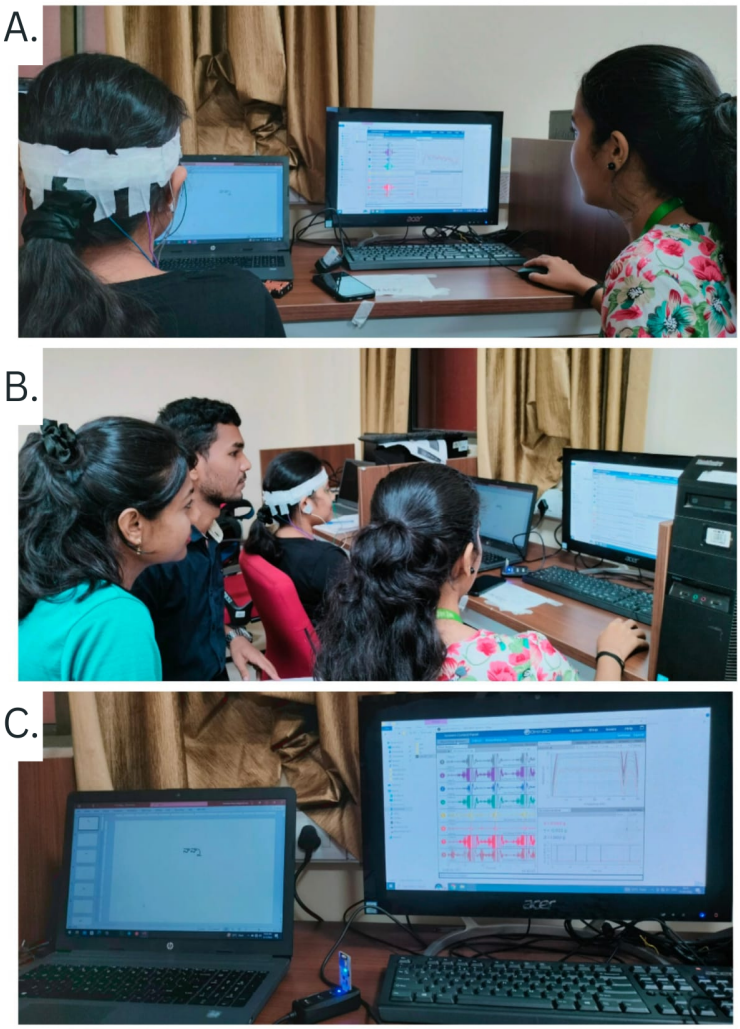

Fig. 4. This representation clarifies the setup used in an ongoing EEG data collection experiment. (A) A depiction of the initialization setup employed for EEG recording collection. (B) The collaborative team engaged in EEG recording collection for the dataset. (C) An image presenting Telugu word visuals on the left screen contrasted with corresponding EEG recordings displayed on the right screen.

participant and the collaborated team. A PPT file containing 100 Telugu words paired with their English translations was displayed on a laptop during the experiment. Each slide presented a single Telugu word followed by an empty slide. The female volunteer, positioned in front of the laptop, progressed to the next word by PowerPoint slideshow. Before the teammate commenced and upon completion of each recording for every word, this procedure was executed. A team of four collaborated to help the volunteer capture EEG signals associated with spoken Telugu words, their English translations, and sub-vocal (silent pronunciation). Subsequently, Data from each of the 8 channels was represented in separate columns within the text file.

Before commencing the slideshow, a fresh session was established using the Open BCI GUI. The EEG recording commenced in synchronization with the volunteer's spoken word at a consistent volume as each word was displayed on the screen. After fully articulating each word, the EEG recording concluded. Typically, each EEG recording lasted only a brief duration. The process was iterated until we got 10 trials in total. Figure 5 shows the interface of open BCI having different slots such as Time Series, Band Power, Headplot, and textfile. The use of collecting time series data in OpenBCI is to analyze and interpret the temporal dynamics of brain activity. Also, time series analysis plays a crucial role in extracting meaningful insights from EEG data acquired through the OpenBCI interface.

A head plot, provided by OpenBCI, allows users to visualize the placement of EEG electrodes on the scalp. It can display statistical measures such as t-values, p-values, or effect sizes on the scalp surface, indicating regions of significant differences or effects. This helps to identify brain regions that are involved in specific cognitive processes or are affected by experimental manipulations.

Band Power refers to the power of the EEG (electroencephalography) signal within specific frequency bands. EEG signals are typically decomposed into different frequency components having their ranges each associated with different brain states and cognitive processes. Bandpower analysis in OpenBCI involves quantifying the power of specific frequency bands within EEG signals recorded by the device.

C. The chosen List of Telugu words

The datasets were generated by incorporating 100 commonly used Telugu words paired with their English translations. These words were specifically selected for their simplicity, widespread usage, and practical relevance in addressing everyday needs. This meticulous curation ensures that the datasets cater to a diverse audience and are applicable in a variety of contexts, thereby, enhancing their utility and accessibility for different users.We carefully chose the words for our dataset to make sure they were both relevant and easy to understand for our target audience. Our goal was to select words that would be most helpful for individuals facing communication difficulties, ensuring they could benefit the most from our dataset.For the sub-vocal data collection, Both Telugu words and English words were utilized. We depended on the timings indicated by the PowerPoint slideshow presentation to start and terminate the recordings, as these words were sub-vocally articulated rather than audibly pronounced. Figures 7, 8, 9, and 10 present a detailed list of these words.

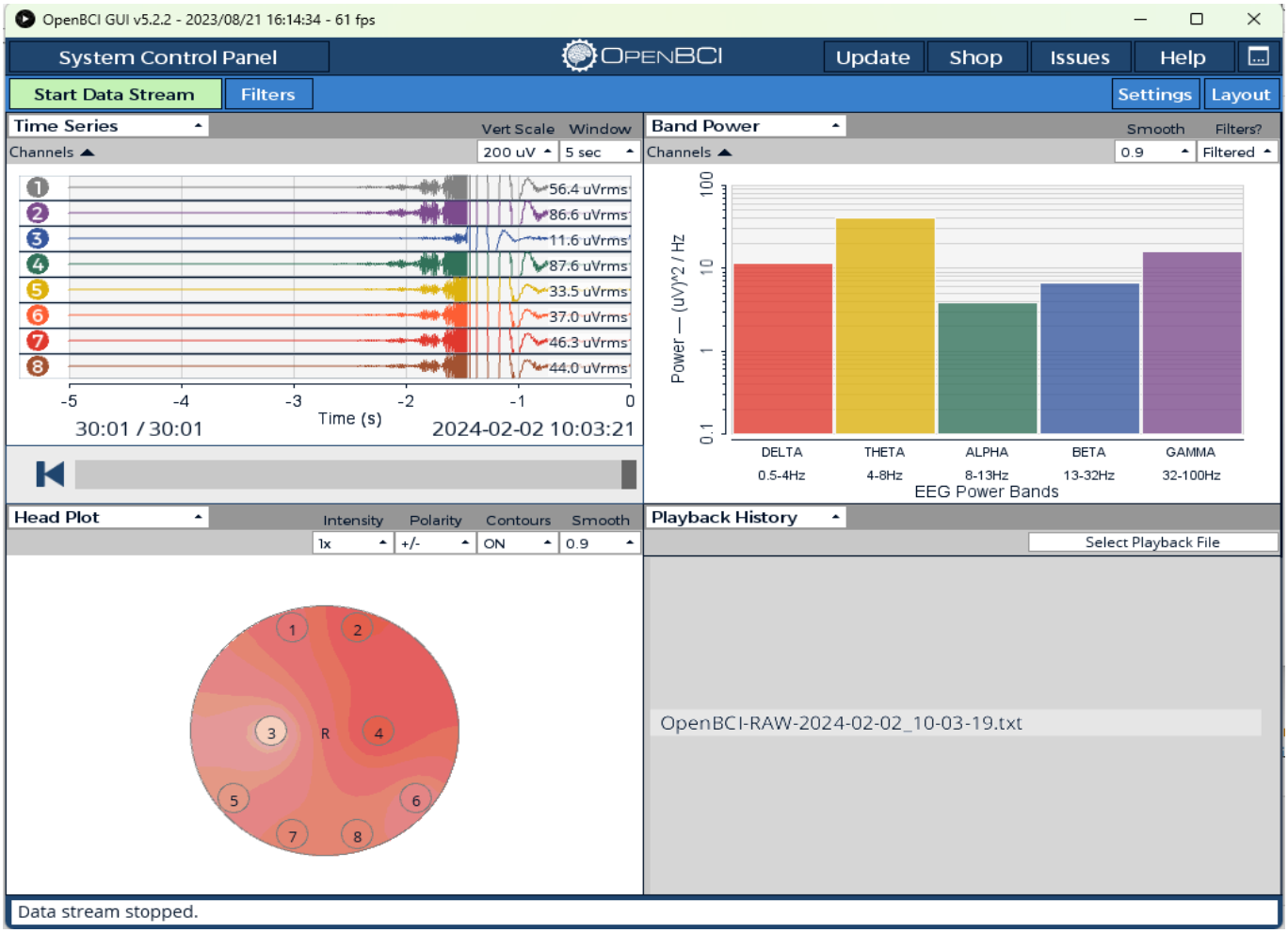

Fig. 5. The image illustrating the Interface of open BCI having time series, Band Power, Head Plot, and Text file.

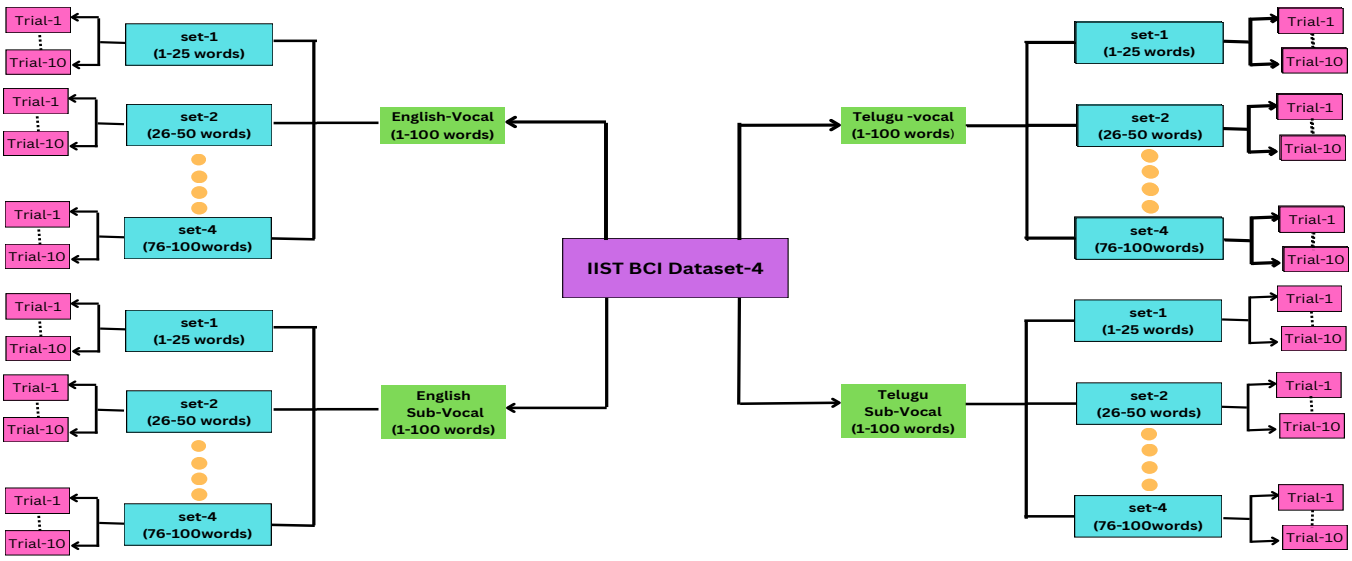

Fig. 6. Organization of files in the dataset folders.

S.no	Telugu Words	English Words corresponding to Telugu words
1.	నాన్న	PAPA
2.	అమ్మ	MUMMY
3.	అన్న	BROTHER
4.	వైద్యుడు	DOCTOR
5.	నీరు	WATER
6.	మందులు	MEDICINE
7.	చలి	COLD
8.	వేడి	HOT
9.	తలనొప్పి	HEADACHE
10.	జ్వరం	FEVER
11.	కండరం	MUSCLE
12.	చెయ్యి	HAND
13.	కాలు	LEG
14.	ఆహారం	FOOD
15.	ఎత్తడం	LIFT
16.	ఉంచు	KEEP
17.	ఆపు	STOP
18.	సమయం	TIME
19.	వినుు	LISTEN
20.	పాట	SONG
21.	సహాయం	HELP
22.	ఉదయం	MORNING
23.	స్నానము	BATH
24.	మరుగుదొడ్లు	TOILET
25.	తలుపు	DOOR

Fig. 7. Word Set-1

S.no	Telugu Words	English Words corresponding to Telugu words
26.	తెరవడం	OPEN
27.	మూసివేయి	SHUTDOWN
28.	ఆరంభించండి	TURN ON
29.	ఆపివేయండి	TURN OFF
30.	అల్పాహారం	BREAKFAST
31.	పాలు	MILK
32.	టీ	TEA
33.	కిటికి	WINDOW
34.	ఎండ	SUNLIGHT
35.	ధ్యానం	MEDITATION
36.	పండ్లు	FRUITS
37.	ఆసుపత్రి	HOSPITAL
38.	ఇల్లు	HOME
39.	విశ్రాంతి	REST
40.	అవును	YES
41.	కాదు	NO
42.	అనుభూతి	FEEL
43.	కళ్లు	EYES
44.	కుటుంబం	FAMILY
45.	ఈరోజు	TODAY
46.	రేపు	TOMORROW
47.	నిన్న	YESTERDAY
48.	మురికైన	DIRTY
49.	శుభ్రంగా	CLEAN
50.	కల	DREAM

Fig. 8. Word Set-2

S.no	Telugu Words	English Words corresponding to Telugu words
51.	గుర్తుంచుకోవటం	REMEMBER
52.	మరచిపోవడం	FORGET
53.	పైకి	UP
54.	కింద	DOWN
55.	శుభప్రదమైనది	AUSPICIOUS
56.	తాతయ్య	GRANDFATHER
57.	అమ్మమ్మ	GRANDMOTHER
58.	బట్టలు	CLOTHES
59.	డబ్బు	MONEY
60.	శరీరం	BODY
61.	సాయంత్రం	EVENING
62.	రాత్రి	NIGHT
63.	మధ్యాహ్నం	AFTERNOON
64.	మంచం	BED
65.	వినోదం	ENTERTAINMENT
66.	ప్రకృతి	NATURE
67.	పాత్రలు	UTENSILS
68.	కుర్చీ	CHAIR
69.	బంధువు	RELATIVE
70.	గడియారం	CLOCK
71.	సబ్బు	SOAP
72.	కొళాయి	TAP
73.	వ్యాయామం	EXERCISE
74.	బలహీనమైన	WEAK
75.	ఊహించు	IMAGINE

Fig. 9. Word Set-3

S.no	Telugu Words	English Words corresponding to Telugu words
76.	నిద్ర	SLEEP
77.	గది	ROOM
78.	నేల	FLOOR
79.	కూరగాయలు	VEGETABLES
80.	ప్రయాణం	TRAVEL
81.	వాహనం	VEHICLE
82.	మాంసాహారి	NON-VEGETARIAN
83.	శాకాహారి	VEGETARIAN
84.	భ్రమ	ILLUSION
85.	సమాజం	SOCIETY
86.	పుస్తకం	BOOK
87.	ప్రశ్న	QUESTION
88.	అలవాటు	HOBBY
89.	కష్టం	HARD
90.	అర్థంచేసుకోవడం	UNDERSTAND
91.	కోల్పోవడం	LOST
92.	ఆలోచించుట	THINKING
93.	కూర్చో	SITDOWN
94.	నిలబడు	STANDUP
95.	నవ్వు	LAUGH
96.	ఏడుపు	CRY
97.	భావోద్వేగం	EMOTION
98.	నిర్ణయం	DECISION
99.	ప్రవేశం	ENTRY
100.	ముగింపు	END

Fig. 10. Word Set-4

III. ORGANIZATION OF FILES

The dataset files comprise unprocessed raw data, each containing multiple columns. Generated by the OpenBCI board, the dataset emphasizes the sample index, EEG data from eight channels (Channel 0 to Channel 7), and the corresponding timestamp.

As shown in Figure 11 dataset is structured as a text file containing values organized in rows and columns which are isolated by commas. The sample index occupies the first column, while columns 2 to 9 contain recordings from all the channels. Channels 10-22 and 24 contain extraneous data. Column 23 contains the timestamp in an unprocessed format, while column 25 displays the timestamp in a human-readable format.

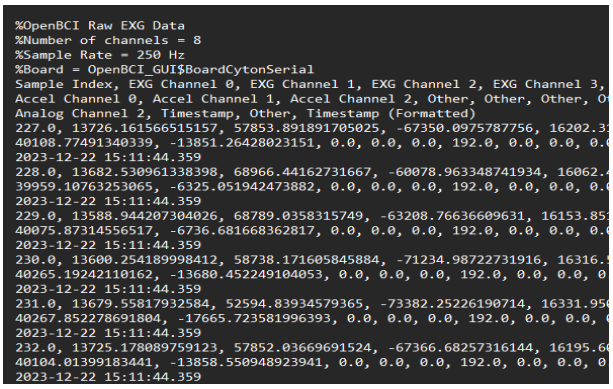

Fig. 11. A preview of the raw data in text file format.

Each data-gathering session carried out using the OpenBCI GUI is segregated into its directory. Within each session directory, every recording is saved as an individual file. Our dataset consists of four main directories: Telugu-Vocal, English-Vocal, Telugu-Subvocal, and English-Subvocal. Each of these directories includes four subdirectories named Set-1, Set-2, Set-3, and Set-4. In each set, there are ten instances for data collection. Figure 6 visually depicts the hierarchical grouping of files within the dataset.

IV. CONCLUSIONS

This paper introduced a new dataset consisting of a carefully curated Electro Eencephaloggraph (EEG) dataset generated when speaking 100 commonly used Telugu words and their English word equivalents. The dataset was created using the OpenBCI Cyton Biosensing board, using 8 channels following the international 10-20 system. The created dataset is invaluable for researchers and developers in the field of Brain-Computer Interface (BCI) applications, particularly for Telugu-speaking patients suffering from Neurodegenerative diseases. The dataset is available on IEEE Dataport.

ACKNOWLEDGMENT

We convey our thankfulness to the entire administrative team of the Indian Institute of Space Science and Technology (IIST), Trivandrum, including the Director, of IIST, for their invaluable assistance with this project. It is important to emphasize that the dataset is only meant to be used for academic and research purposes and that it is provided in its current state. The dataset is not subject to any explicit or implicit assurances or warranties. It is important to note that any inadvertent errors discovered in the dataset cannot be attributed to the authors. As a result, users are encouraged to utilize it with caution and wisdom. We invite interested parties to contact the authors if they require any further information or explanation beyond what is offered in this text. For communication with the corresponding author, B. S. Manoj, kindly use the email address bsmanoj@iist.ac.in.

REFERENCES

[1] Parvathi Nair, Parvathy S S, Nancy Sunil, Anurag Mukati, Charu Chauhan, S Sumitra, and B S Manoj, “IIST BCI Dataset-1 for Selected Common Malayalam Words,” IEEE Dataport, January 2024. doi: https://dx.doi.org/10.21227/83t5-tw46

[2] Parvathi Nair, Parvathy S S, Nancy Sunil, Anurag Mukati, Charu Chouhan, S Sumitra, and B. S. Manoj, “IIST BCI Dataset-1 for Selected Common Malayalam Words,” TechRxiv. January 2024. DOI: 10.36227/techrxiv.170473894.42689961/v1

[3] Shubham Tayade, Parvathy S S, Nancy Sunil,Charu Chouhan, S Sumitra, and B. S. Manoj. IIST BCI Dataset-2 for Selected Common Marathi Words. TechRxiv. March 14, 2024. DOI:https://doi.org/10.36227/techrxiv.171043118.80448751/v1

[4] Parvathy S S, Nancy Sunil, Shubham Tayade, Chittaloori Likhitha, S. Sumitra, B. S. Manoj. IIST BCI Dataset-3 for 100 Malayalam Words. TechRxiv. April 08, 2024. DOI: 10.36227/techrxiv.171259992.21826294/v1

[5] Parvathy S S, Nancy Sunil, Shubham Tayade, Chittaloori Likhitha, S. Sumitra, B. S. Manoj. IIST BCI Dataset-3 for 100 Malayalam Words.DOI: https://dx.doi.org/10.21227/wa4h-cr10

[6] https://docs.openbci.com/Cyton/CytonLanding/

[7] https://en.wikipedia.org/wiki/10%E2%80%9320_system_(EEG)

[8] 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.

[9] 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.

[10] 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.

[11] MIKE X COHEN - Analyzing Neural Time Series Data (Theory and Practice)

[12] https://www.scienceabc.com/innovation/what-is-a-brain-computer-interface.html

[13] https://www.frontiersin.org/research-topics/57633/brain-computer-interfaces-in-neurological-disorders-expanding-horizons-for-diagnosis-treatment-and-rehabilitation

[14] https://ieee-dataport.org/documents/iist-bci-dataset-1-selected-common-malayalam-words