Data Description – Track#3 Multi-class imagined speech classification

Every dataset is publicly available via the Open Science Framework (https://osf.io/pq7vb/). This manual describes data usage. The details of the dataset description and converted data are described below. You can check the contents by referring to the list. The subjects (S1-S15; aged 20-30 years) were asked to perform imagined speech of five different words/phrases that can be utilized for basic conversation: 'hello', 'help me', 'stop', 'thank you', and 'yes'.

Contents

1. Dataset description
2. Converted data

1. Dataset description

- EEG of five-class imagined speech words/phrases were recorded. 70 trials per class (70×5 = 350 trials) are released for training (60 trials per class) and validation (10 trials per class) purpose. Using the given validation set is not obligated. Validation for the training data can be performed not only by the given validation set but also with the competitors' choice (example: N-fold cross-validation over the whole data). The test data (10 trials per class) will be released later.
- The dataset was divided into epochs based on cue information (event codes).

1. **Background and purpose**

   1-1. Experimental setup

   The purpose of this experiment is to classify multi-class imagined speech with a robust performance. Imagined speech of five words/phrases for basic communication ('hello', 'help me', 'stop', 'thank you', and 'yes') was recorded.

   During the experiment, the subjects were seated in a comfortable chair in front of a 24-inch LCD monitor screen. The subjects were instructed to imagine the silent pronunciation of the given word as if they were performing real speech, without moving any articulators nor making the sound. Subjects were strictly directed not to perform any other brain activity but the given task. They were
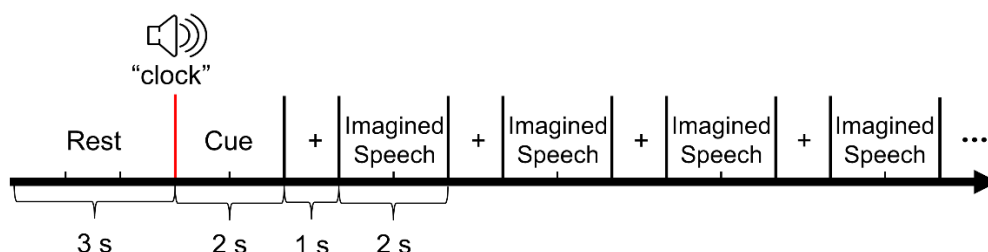

Fig. 1: Illustration of the experimental protocol

noted not to move and avoid dye blinks while imagining or receiving the cue. All imagination trials were performed while a black screen was provided, without receiving any stimulus in order to avoid any other factors affecting the brain activity.

An auditory cue of the five words/phrases was randomly presented for 2 s, followed by 0.8-1.2 s of a cross mark. The subjects were instructed to perform imagined speech of the given cue as soon as the cross mark disappears on the screen. Four times of cross mark (0.8-1.2 s) and imagined speech phase (2 s) were followed in a row per random cue. After performing four times of imagined speech, 3 s of the relaxation phase was given to clear up the mind for the next word/phrase.

1-2. Data acquisition

The EEG data were recorded with an EEG signal amplifier (BrainAmp, BrainProduct GmbH, Germany). The raw data were recorded using BrainVision (BrainProduct GmbH, Germany) with MATLAB 2019a (The MathWorks Inc., USA). 64 EEG electrodes following a 10-20 international configuration were used for the recording. Ground and reference channels were placed on the Fpz and FCz. The impedance of all electrodes between the sensors and the skin of the scalp were maintained below 15 kΩ.

2. Event codes

2-1. Each class

class	Hello	Help me	Stop	Thank you	Yes
event code	1	2	3	4	5

2-2. Start and end of the experiment (a whole session)

class	Start	End
event code	13	14

3. Channel labels

channel	label	channel	label	channel	label	channel	label	channel	label
1	Fp1	16	T8	31	O2	46	FT10	61	PO3
2	Fp2	17	TP9	32	PO10	47	C5	62	POz
3	F7	18	CP5	33	AF7	48	C1	63	PO4
4	F3	19	CP1	34	AF3	49	C2	64	PO8
5	Fz	20	CP2	35	AF4	50	C6		
6	F4	21	CP6	36	AF8	51	TP7		

7	F8	22	TP10	37	F5	52	CP3		
8	FC5	23	P7	38	F1	53	CPz		
9	FC1	24	P3	39	F2	54	CP4		
10	FC2	25	Pz	40	F6	55	TP8		
11	FC6	26	P4	41	FT9	56	P5		
12	T7	27	P8	42	FT7	57	P1		
13	C3	28	PO9	43	FC3	58	P2		
14	Cz	29	O1	44	FC4	59	P6		
15	C4	30	Oz	45	FT8	60	PO7		

2. Converted data

- If you want to download data of .mat file, you can find it from `Track3/Training Set,' and 'Track3/Validation Set'.

1. **Description of each file**

   1-1. File

   Uploaded data files as subject number
   ex) Data_sample01, Data_sample02, ..., Data_sample15.mat

   Each Data file contains two structs

   - epo: data file containing the data and the lables (epo_train or epo_validation or epo_test)
   - mnt: montage file of channels

   1-2. Folder

   The main folder 'Track3/' contains two sub folders named 'Training Set' and 'Validation Set'

   - training: Dataset of 300 trials (5 class × 60 trials) for subject 1–15. Please use it for training.
   - validation: Dataset of 50 trials (5 class × 10 trials) for subject 1–15. Please use it for validation, but you may use it as an additional training data.
   - test (release date; TBA): Dataset of 50 trials (5 class × 10 trials) for subject 1–15.

     Subject dependent analysis is required in the competition. Therefore, the model learned by Data_Sample01 in the Training Set is evaluated by Data_Sample01 in the Test Set.

1-3. Recording

The files contain EEG data of 15 subjects performing imagined speech of five words/phrases.

- o Class name

  Hello, Help me, Stop, Thank you, Yes

- o Paradigm

  Imagined speech

Name	Description
epo.x	Raw data (time × channels × trials)
epo.y	Class labels corresponding to the epo.x
epo.fs	Sampling frequency
epo.t	Time points (-500 ms ~ 2600 ms)
epo.className	Class name
epo.file	File name
epo.clab	Channel information
mnt.x	X coordinates for channel position
mnt.y	Y coordinates for channel position
mnt.pos_3d	3D coordinates for channel position
mnt.clab	Channel information

Thank you for reading. Good luck!