EEG ERP classification and class balance [Apr 19]

This text continues the discussion of Apr 12th. The conclusion was: to present a classification model with accuracy 0.9.

The computational experiment and classification accuracy

We solve the **two-class** classification problem.

- The object of classification: an Event from the list of events. See the 1st and 2nd columns of Table 1
- The labels of classes: **Correct** or **Incorrect** response after the Event onset
- The epoch: 300 ms before onset and 1170 ms after onset, at 128 channels, 256 Hz sample rate, all electrodes included
- The quality criterion: average accuracy on the cross-validated data
- One data set contains one event for all users. See Table 1
- List of users: 1034, 1037, 1045, 1158, 1363, 1368, 1385, 2038, 6639, 7974, 7977, 7980,1327

Table1. The accuracy of event classification (these are non-verified preliminary results, and they could change drastically after verification and error analysis)

Task	Event	Class 1 label	Class 2 label	Class 1 size	Class 2 size	Accuracy
Encoding	Larger-than word	Incorrect response	Correct response	37	487	0.923
Encoding	Smaller-than word	Incorrect response	Correct response	29	555	0.950
Recognition	Old word	Incorrect response	Correct response	59	481	0.885
Recognition	New word	Incorrect response	Correct response	99	426	0.786

Questions about the dataset

1. **Do the users allow to change their responses during the experiment?** Was it planned in the experiment intentionally? What is more important for modeling: the first response, the second of the fact of change? See Table 2 for details.

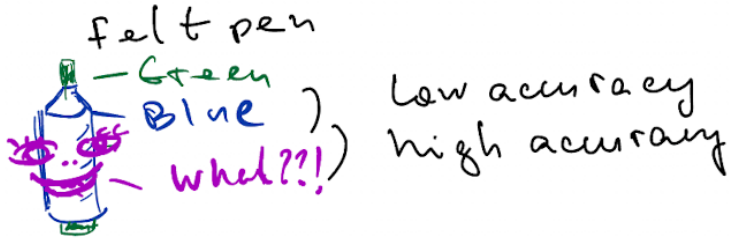
2. **How do we deal with significant class imbalance?** See Section Class imbalance illustration. Will it be the same after the final data collection? Or will the user be less sure about the answers during the final experiment? See Section User embarrassment hypothesis.

Questions about the experiment (our considerations painted blue)

1. **Does the accuracy apply to the event or to the user?** Default answer: we use the Bayesian rule and apply it to the user. Agree with your point; we need to check with psychiatrists and other people. I am guessing we can start with Stimulus detection!
2. **Is this data final?** Default answer: the data format will be the same, and the classes will be simplified. Not necessarily; we start with this dataset
3. **Is the user's emotional state change?** Default answer: the user feels the same way so that it will give results of the same accuracy. Yes; this is the idea behind of DGD, because we have a latent process which capture emotional changes in the data. For now, we assume stationarity of the emotion.
4. **Is user embarrassment welcome in the experiments?** See Section User embarrassment hypothesis. Your observation is true that there might be trial or response dependence across trials! But this should not be the concern as the first round of the analysis; but, if our DGD is the correct model, then it might bring “embarrassment” through the latent process.
5. **What kind of proof do you expect to show the model works?** Default answer: we propose a list of tests for error analysis. What sort of visualization do you expect?
   At the current phase, Performance! Eventually, 1. Physiology, consistency of the leading features across patient, 2. Performance again (cross-validation), 3. Scalability if a system can work over different times for one patient or across patients.

User embarrassment hypothesis

To make a high-quality dataset is important to put a user in a surprised or embarrassed state. The hypothesis is that a user's feelings bring more variety in signals than the perceptions of different familiar stimuli. This hypothesis is now preliminary and supported by the data. Under this condition, almost any machine learning model gives decent stable accuracy.



Examples:
Task 1 encoding Large Correct vs Large Incorrect accuracy: 0.923
Task 1 encoding Large Correct vs Small Correct accuracy: 0.533
Task 3 recognition Old Correct vs Old Incorrect accuracy: 0.885
Task 3 recognition Old Correct vs New Correct accuracy: 0.581
Actually, they are designing user specific experiment to evoke the signal of interest.

The user’s double response

In the experiment collected from all users, the total number of events discovered: 4372. Events lost: 362, usually due to no response from the user. Events collected in the dataset: 4010. Users change their decision after the 1st response 872 times (each fifth event). The average delay is 0.160 ms.

Table 2. Delays of the users’ second responses

User	Number of events	Count changes **1->11**	Average delay, ms	Count changes **11->1**	Average samples	Count changes **2->12**	Average samples	Count changes **12->2**	Average samples
1034	**285**	0		3	488	1	23	38	133
1037	**275**	0		4	168	0		49	188
1045	**349**	1	242	9	258	0		34	242
1158	**355**	0		7	219	1	47	36	117
1363	**324**	1	211	3	156	1	23	69	125
1368	**296**	0		3	145	0		91	113
1385	**311**	0		0		0		72	160
2038	**278**	0		0		4	137	75	199
6639	**280**	1	102	16	172	2	148	69	164
7974	**289**	0		12	168	0		76	172
7977	**376**	4	203	39	211	2	172	53	188
7980	**330**	1	152	12	160	1	184	12	148
1327	**262**	0		0		0		70	137

Class imbalance illustration

The class imbalance significantly increases sufficient sample size. Each figure corresponds to one data set from Table 2. The x-axis shows the class labels: Incorrect response (1, 2) and Correct response. The y-axis shows the sample size of each class.

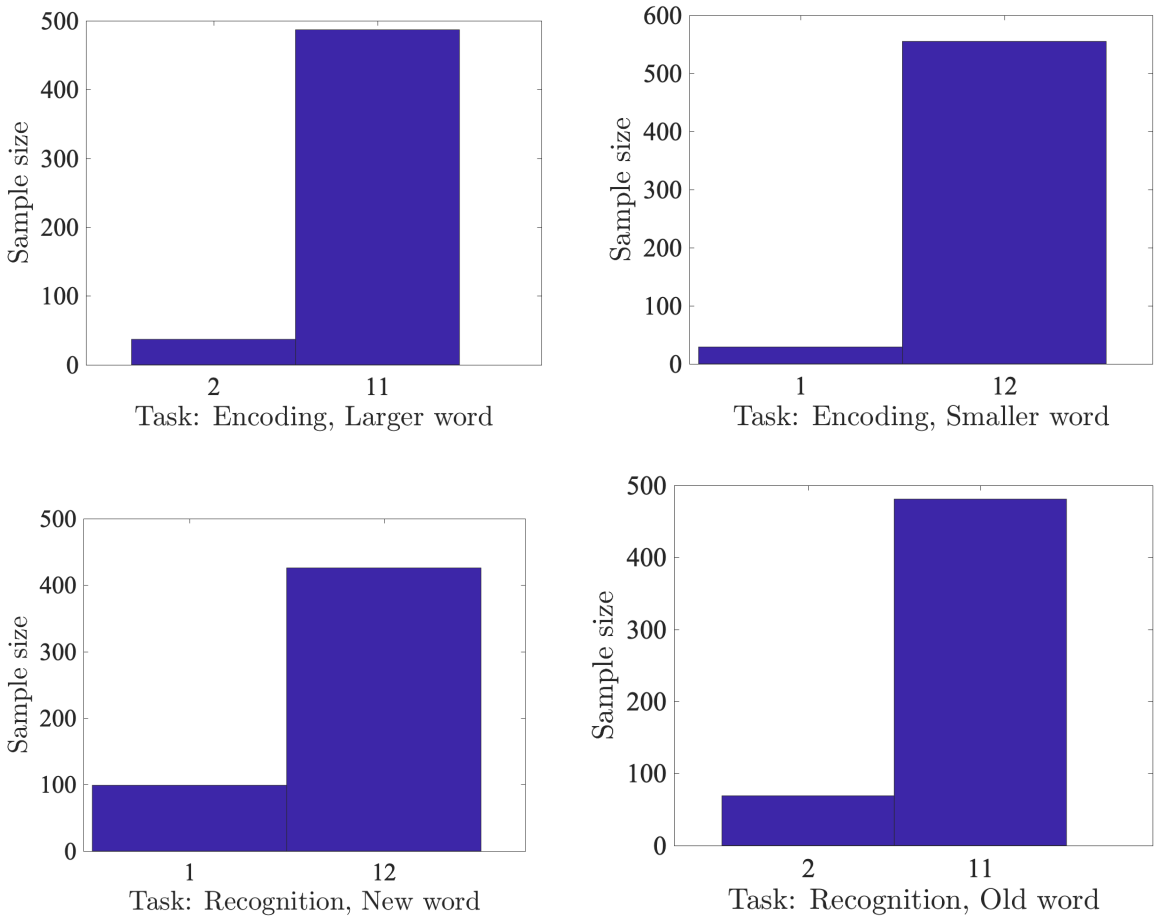


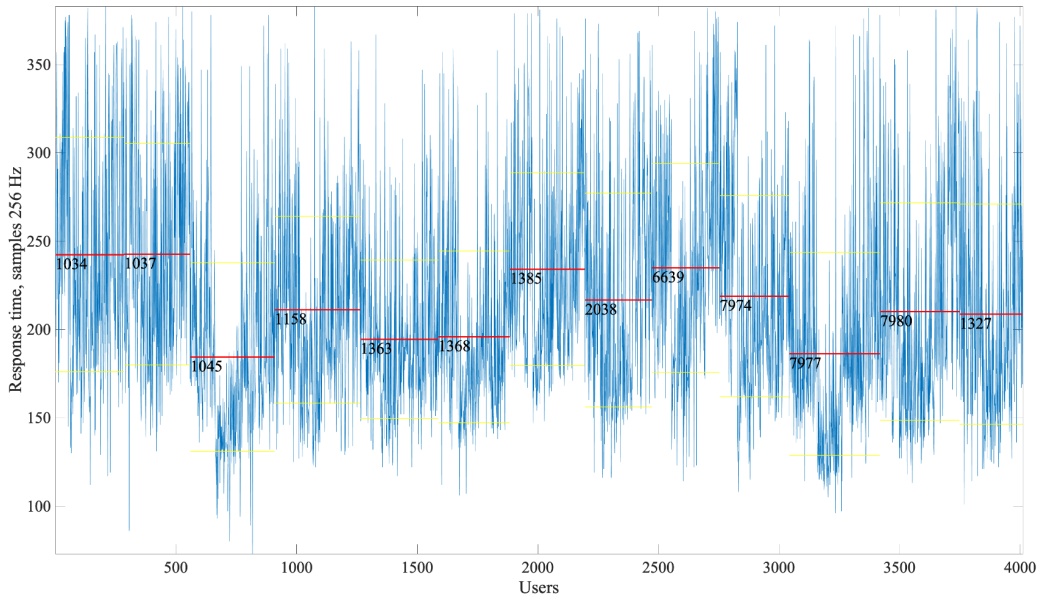
Table 3. The average Class balance is **1:15 Incorrect to Correct responses**

Task	**Event**	**Sample size of Incorrect class**	**Sample size of Correct class**
Encoding	Small	**29**	**555**
Encoding	Large	**37**	**487**
Lexical	Old	7	627
Lexical	New	29	539
Lexical	Non	37	588
Recognition	Old	59 (69 three labels) 1 - 10; 2 - 59; 12 - 481;	481
Recognition	New	99	426

Appendix 1

Average response time per user

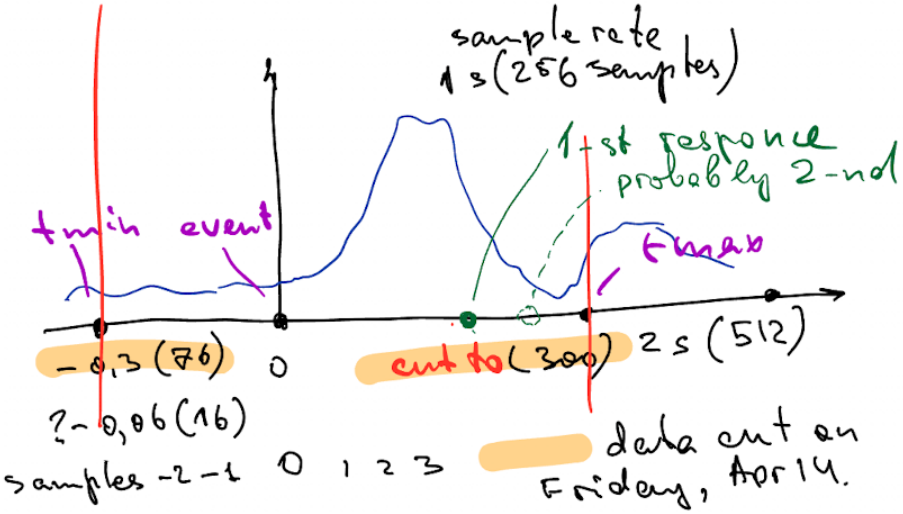
The x-axis represents the unsorted users' response events. The y-axis shows the response time in samples at 256 Hz. In conclusion, the response standard deviation is 300 samples.



The current data cut

The three-way matrix dimensionality is 4010 x 128 x 375 (events x electrodes x samples).

Number of samples is 375 = 512-212 + 76 - 1 at 256 Hz.



Model complexity, classification of Incorrect vs Correct responses

Feature extraction model: tangent hyperplanes in the Riemannian space of positive curvature.
Classification model: logistic regression.

Task 1 (encoding) Larger than
components: 5, accuracy: **0.925573**
components: 10, accuracy: 0.927481

Taks 2 Old word
components: 5 accuracy: 0.988959

Taks 2 New word
components: 5 accuracy: 0.947183

Taks 3 Non-word
components: 5 accuracy: 0.937600

Taks 3 Old word
components: 3 accuracy: 0.887037
components: 5 accuracy: **0.890741**
components: 10 accuracy: 0.885185

Task 3 New word
components: 10 accuracy: **0.813333**
components: 5 accuracy: 0.786667