Predicting numerical processing in naturalistic settings from controlled experimental conditions

J. Schrouff*†, C. Phillips‡, J. Parvizi*† and J. Mourão-Miranda§
*Laboratory of Behavioral and Cognitive Neurology, Stanford University, Stanford, California, USA
†Stanford Human Intracranial Cognitive Electrophysiology Program (SHICEP)
‡Cyclotron Research Centre, University of Liège, Belgium
§Department of Computer Science, University College London, United Kingdom

***Abstract*—Machine learning research is interested in building models based on a training set that can then be applied to new data, whether this unseen data comes from new examples (e.g. new subjects, other tasks) or new features (e.g. different modalities). In this work, we present a simple approach to transfer learning using intracranial EEG (also known as electrocorticographic, ECoG) data from three patients. More specifically, we aimed at detecting numerical processing during naturalistic settings based on a model trained with controlled experimental conditions. Our results showed significant prediction accuracy of numerical events in naturalistic settings when considering a priori knowledge of the target task.**

***Index Terms*—Electrocorticography; Transfer learning; Multiple Kernel Learning**

I. Introduction

Machine learning techniques have been applied to brain data in order to decode a variable of interest (e.g. the cognitive state of a subject [1] or a type of disease [2]), with a certain generalization performance often derived from cross-validation schemes. More recently, advanced techniques have been developed to improve generalization ability, especially in the context of multiple subject [3] or multiple modality learning [4].

Another interesting application is to generalize a model from one task to another. In this case, the training and test data might not have been drawn from the same feature space and the same distribution. This field of research, referred to as Transfer Learning [5], has received increased attention from the machine learning community, with applications in diverse fields including Web document classification and marketing. Transfer learning can be defined as the ability of a system to recognize and apply knowledge learned in previous domains/tasks to novel domains/tasks, which share some commonality [5]. Transfer learning approaches therefore assume that the train and target tasks are *related*. Using such a scheme on neuroimaging data implies that the brain activity generated in one or more tasks could help us identify/characterize brain activity generated during the task to predict [6]. This could yield important insights for cognitive neuroscience.

The present work provides a first step in the direction of transfer learning using ECoG recordings of numerical processing in experimental and in naturalistic settings. Previous work has revealed similarities in the electrical brain signal generated during numerical processing in both tasks [7]. In the current study, two main questions were investigated using an approach inspired from transfer learning: (1) *Can we detect numerical processing in naturalistic settings based on a model trained on experimental conditions?*, (2) *How does the detection performance depend on the selected features?* In this work, the transfer between the two tasks was direct: we assumed that the hypothesis space of the experimental settings could be used to model the naturalistic settings.

II. Material

A. Data

The material considered in this work is the same as in [7]. Therefore, only a brief description of the population and experimental design will be provided.

Three subjects were implanted with intracranial electrodes to localize the source of drug-resistant seizures[1]. Signal was continuously recorded for clinical purposes for 7-10 days, during which simultaneous video monitoring was performed (Nihon Kohden Technology, sampling rate: 1000Hz for P1 and P2, 500Hz for P3). Electrodes containing artifacts or pathological activity were discarded from further analyses.

Data was recorded when the patients performed simple true/false judgments of memory sentences or mathematical equations (Fig.1A). The memory sentences comprised self-episodic (e.g. 'I ate pizza this week'), self-semantic (e.g. 'I eat pizza often'), and self-judgment (e.g. 'I am a curious person') statements. Basic mathematical additions were presented along with a result (e.g. '4 + 49 = 53', further referred to as 'Math' condition). Interleaved across trials were 5s cued-rest periods, during which a centered cross sign was displayed on the screen and patients were instructed to fixate and rest. The experiment comprised 96 randomized trials of each condition (except for rest, ∼66 trials) and was divided in two sessions.

In addition to the data acquired during experimental conditions, we identified periods of naturalistic condition during which the patients interacted with their environment (e.g. talking with medical staff or on the phone with family members). Windows of 10 minutes (P1) or 6 minutes (P2, P3) were transcribed (i.e. each word was written down along with its timing, temporal resolution of the transcription: 1s) and selected for further analysis.

[1] The procedure was approved by the Stanford Institutional Review Board and the subjects provided written informed consent to participate in the study.

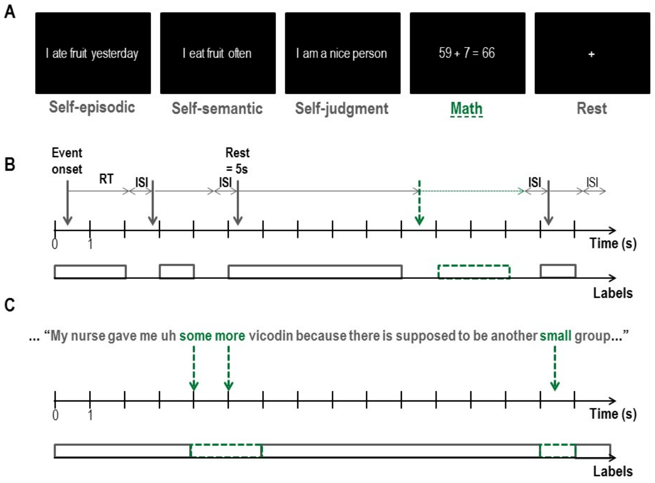

Fig. 1. **Experimental design and event extraction.** **A**. Examples of events presented during experimental condition. 48 events of each category (except rest = 33) are presented during each of 2 distinct sessions. **B**. Illustration of event presentation (arrows) and extraction (axes). RT stands for response time, ISI for inter-stimulus interval (200ms). For each event, its onset is rounded to the closest second and its duration floored to the previous second, giving categorical labels (math, in green, or others, in grey) to each 1s time window. **C**. According to the transcripts of the videos, words assessed as ‘quantitative’ were marked as ‘Math’ events while everything else (including silences) was marked as ‘Other’.

B. *Definition of events*

1) Experimental settings: For the experimental settings, a precise onset can be derived for each event, based on the visual presentation of the stimulus. However, due to the temporal resolution of the transcription (1s), events in the naturalistic cannot be extracted with such timing accuracy. Two feature sets were hence defined: (1) considering knowledge of the training task only, and (2) including knowledge about the target task.

1) Events were extracted based on the precise onset of visual stimulus presentation (in ms).
2) Each precise onset was rounded to the closest second, and the duration of the visual presentation was rounded towards negative infinity to the closest second (i.e. Matlab ‘floor’). Each 1s time window defined by this strategy was labeled as ‘Math’ if an equation was presented, or ‘Others’ in the case of self-episodic, self-semantic, self-judgment or rest trials. This is illustrated in Fig.1.B.

Feature sets #1 and #2 hence corresponded to the same trials, with the only difference being in the temporal resolution of their extraction.

2) Naturalistic settings: As the temporal resolution of the transcription is low (1s), events in the naturalistic settings were extracted at each second of the recording. Each event was labeled as ‘Math’ if quantities or numbers were evoked (e.g. ‘some’, ‘a bottle of, ‘small’, ‘all the nurses’, ‘once every hour’, ‘10 days ago’), and ‘Other’ otherwise (including silences). This is illustrated in Fig.1.C. It is important to note that the labeling of naturalistic events is prone to imprecision, due to the low temporal resolution of the transcription compared to the flow of words, as well as very different sampling rates between the video and the ECoG signal. However, since only the video can bring information about the behavior of the patient, we assumed that the derived labels represented the ‘ground truth’.

C. *Pre-processing*

Signal pre-processing was performed using Matlab[2] and SPM8[3]. The continuous signal was first filtered for line noise and harmonics, re-referenced to the average of all channels and downsampled to 436Hz.

For both feature sets, the defined epochs were extracted using a $[-200ms, 1200ms]$ time window around ‘onset’ and a time-frequency decomposition was performed using Morlet wavelets (7 wavelets), with frequencies of interest log-spaced between 1 and 110Hz. The resulting decomposition was scaled (point-wise) by the logarithm of its value. The instantaneous power of the signal in the 0 to 1000ms time-window was then averaged in each of the following frequency bands: δ (1-4Hz), θ (4-8Hz), α (8-12Hz), β (15-25Hz), low-γ (30-55Hz) and a narrow band of High Frequency Broadband (HFB), high-γ (70-110Hz), by averaging the frequency bins within those bands. For each channel, 436 features were hence considered, representing the average power in a chosen frequency band in the selected 1s time window.

III. Methods

A. *Model*

Linear kernels were built for each channel and each frequency band. Modeling was performed using the simpleMKL algorithm [8], a multiple kernel learning (MKL) approach based on support vector machines (L1 regularization on kernels). A binary MKL classifier was trained to discriminate between ‘Math’ and ‘Other’ trials, based on the two sessions of experimental settings. The estimated model was then applied to each 1s epoch of the naturalistic settings to detect ‘Math’ events.

B. *Performance assessment*

1) Experimental settings: To apply the model estimated using data from experimental settings to naturalistic settings, we first need to show that such model was able to significantly discriminate between ‘Math’ and ‘Other’ trials. This was achieved using a 10-folds cross-validation on the events (by keeping events extracted from a single visual presentation in blocks). An inner cross-validation was performed to optimize the soft-margin hyper-parameter ($C = 10^{-2:1:3}$). Model performance was estimated via the computation of balanced accuracy (i.e. the average of class accuracies). Significance was assessed using 1000 permutations, with results associated to a p-value smaller than 0.05 reported as significant.

[2] www.mathworks.com

[3] www.fil.ion.ucl.ac.uk/spm

2) Naturalistic settings: In the present case, the model was trained on the two sessions of experimental settings and tested on each epoch of the naturalistic settings (both on 'Math' and 'Other' trials). The hyper-parameter value was selected through an inner cross-validation based on the sessions (i.e. train on one experimental session and test on the other for each value of C). The value of C leading to the highest balanced accuracy was selected for further modeling. Since we are interested in the detection of numerical processing, we focused the results on the 'Math' condition and computed the class accuracy for 'Math', as well as its positive predictive value (PPV). In view of the imprecision of the labeling of the naturalistic condition, we computed those values based on a direct correspondence between the predictions and the labels, as well as based on a three seconds correspondence, i.e. considering a correspondence whenever a labeled 'Math' behavioral event was preceded $(-1s)$, corresponding (direct $= 0s$) or followed $(+1s)$ by a 'Math' prediction. As for experimental settings, the significance of the obtained results was assessed using 1000 permutations.

IV. Results

A. Events and features

After discarding electrodes containing noisy or pathological signal, 40 channels were selected for P1, 102 for P2 and 97 for P3. This led to the computation of 40*6 = 240 kernels for P1 (612 for P2, 582 for P3), used in the simpleMKL model.

1) Feature set #1 - Experimental settings: 96 'Math' events were extracted for each subject, based on the two experimental settings. They were balanced with 96 events randomly selected in the four other conditions, i.e. 24 events labeled as self-episodic, 24 labeled as self-judgment, 24 labeled as self-semantic and 24 labeled as rest.

2) Feature set #2 - Experimental settings: Regarding the two sessions of experimental condition, 184 'Math' events were extracted for P1, 255 for P2, and 185 for P3. They were balanced with an equal number of epochs from each of the four other conditions, i.e. 46 epochs labeled as self-episodic, 46 labeled as self-judgment, 46 labeled as self-semantic and 46 labeled as rest (total 184) were randomly selected for P1 (63 per category for P2, 46 per category for P3).

3) Naturalistic settings: Regarding the naturalistic conditions, 33 math epochs were defined for P1 (total: 611 epochs), 96 for P2 (total: 358) and 73 for P3 (total: 358) based on the transcripts. Among the total number of epochs, 230 (resp. 47, 103) epochs corresponded to silences (i.e. patient not speaking) for patient P1 (resp. P2, P3).

B. Model performance

1) Feature set #1 - Experimental settings: This model was trained and tested on experimental settings data using feature set #1. Model performance is displayed in Table I, in terms of balanced and class accuracy for each subject.

TABLE I
Balanced (BA) and class accuracy (in %) for the 'Math' versus 'Other' classification for each subject, as well as selected value for the hyper-parameter C. Significant classification results are displayed in bold.

Subject	**BA**	**'Math' accuracy**	**'Other' accuracy**	**C**
P1	**97.92**	**97.92**	**97.92**	1
P2	**73.75**	**79.79**	**67.71**	1
P3	**81.68**	**85.71**	**77.65**	1

2) Feature set #1 - Naturalistic settings: When considering features based on the exact stimulus presentation in experimental condition, the predictions in naturalistic settings were poor, with almost everything classified as 'Math' for the three subjects in direct accuracy (see Table II) and 100% 'Math' accuracy for 3s-range accuracy.

TABLE II
Balanced (BA) and 'Math' accuracy (in %) for the detection of 'Math' events in naturalistic settings, for each subject, with positive predictive value (PPV). Significant classification results are displayed in bold.

Subject	**BA**	**'Math' accuracy**	**'Math' PPV**
P1	49.92	**99.51**	49.92
P2	48.18	**91.90**	48.96
P3	50.42	**81.68**	50.14

3) Feature set #2 - Experimental settings: This model was trained and tested on experimental settings data using feature set #2. Model performance is displayed in Table III, in terms of balanced and class accuracy for each subject.

TABLE III
Balanced (BA) and class accuracy (in %) for the 'Math' versus 'Other' classification for each subject, as well as selected value for the hyper-parameter C. Significant classification results are displayed in bold.

Subject	**BA**	**'Math' accuracy**	**'Other' accuracy**	**C**
P1	**89.51**	**88.06**	**90.96**	1
P2	**87.19**	**85.49**	**88.89**	1
P3	**76.80**	**72.89**	**80.71**	1

4) Feature set #2 - Naturalistic settings: Considering similar event extraction for both the training and target tasks led to improved results, with significant balanced and 'Math' direct accuracy for patients P1 and P2 (see Table IV). The three subjects displayed significant 3s-range 'Math' accuracy and PPV for this feature set (see Table V).

TABLE IV
Balanced (BA) and 'Math' accuracy (in %) for the detection of 'Math' events in naturalistic settings, for each subject, with positive predictive value (PPV). Significant classification results are displayed in bold.

Subject	**BA**	**'Math' accuracy**	**'Math' PPV**
P1	**68.36**	**42.42**	**29.79**
P2	**52.64**	**8.33**	**50.00**
P3	49.46	1.37	**12.5**

TABLE V

3S-RANGE 'MATH' ACCURACY (IN %) FOR THE DETECTION OF 'MATH' EVENTS IN NATURALISTIC SETTINGS, FOR EACH SUBJECT, WITH POSITIVE PREDICTIVE VALUE (PPV). SIGNIFICANT CLASSIFICATION RESULTS ARE DISPLAYED IN BOLD.

Subject	**'Math' accuracy**	**'Math' PPV**
P1	**72.73**	**44.68**
P2	**18.75**	**81.25**
P3	**4.11**	**12.5**

V. Discussion

In this work, we present a simple approach to transfer learning based on ECoG recordings of numerical processing in two settings (experimental and naturalistic).

The performance obtained on naturalistic settings can seem low. However, the results were significant for two patients in direct accuracy and for the three patients in 3s-range accuracy. This is surprising due to the imprecise labeling of math events in the naturalistic settings: the transcription was performed on 1s temporal resolution videos, with the evocation of quantities or numbers happening at any time during the 1s time windows. Furthermore, the labeling was performed manually, which led to further approximation in the timing of the events. Hence obtaining significant math accuracies, both for direct and 3-s range correspondence can be considered as a good result. We also computed the number of false positives and did not discard any period of silences (i.e. when the patient does not speak). The resulting positive predictive values were significant, showing that the model did not predict everything as 'Math' (which would also lead to high accuracy but poor PPV) or at random (chance accuracy and PPV). This might suggest a common space for numerical processing in the two tasks, hence revealing a relationship between the activity of neuronal populations in the human brain during controlled experimental conditions and during naturalistic settings.

This result was previously suggested by [7], in which the authors used thresholds on the high-γ power to identify 'Math' peaks on specific electrodes. Unfortunately, our results can hardly be compared to the sensitivity and specificity obtained using univariate techniques [7] since we obtained one value per subject and not one value per electrode. In addition, our results display direct correspondence between 1-second epochs of 'Math' trials and behavioral events while the authors of [7] have computed the co-occurrence of a high-γ peak within 5 seconds of a behavioral 'Math' event, discarding periods of silence.

Our results showed that the considered features had a large effect on the results. It seems that including prior knowledge on the features (implemented here through the selection of 1s successive epochs for the experimental settings) is a more sensible choice than selecting features based only on the training set (i.e. extracting epochs based on the stimulus presentation timing). This result suggests that although we added variance (i.e. noise) within the training set, using commonly defined features brought the two hypothesis spaces closer, which led to significant detection of numerical processing within the 3s-range for the three patients.

In this work, the transfer between the two tasks was direct: we assumed that the hypothesis space of the experimental settings could be used to model the naturalistic settings. In addition, the feature construction represents a manual attempt at uncovering a common feature space between the two tasks. Together, these two hypotheses aim at fulfilling the major assumption of traditional machine learning techniques: the train and target data are drawn from the same feature space and the same distribution. Although this simple approach led to significant results, similar transfer problems might benefit from more advanced algorithms, such as Multi-Task Learning which aims at uncovering the common (latent) features between multiple tasks to learn them simultaneously although they are different [9]. In the present case, only one task was available for training. In future experiments, multiple aspects of a classification problem could be divided in different tasks to increase the generalization ability to new tasks (e.g. reading numbers, listening to numbers, mental representation of numbers, . . .). Moreover, investigating what information is shared between tasks to form a common hypothesis space might provide valuable insights on the cognitive question of interest.

Acknowledgment

This work was supported by research grants from the Stanford NeuroVentures Program, US National Institute of Neurological Disorders and Stroke (R01NS078396), US National Science Foundation (BCS1358907) to J.P., a BAEF fellowship and Medical Foundation of the Liege Rotary award to J.S. (Belgium), and a F.R.S-F.N.R.S grant to C.P. (Belgium). J.M.M. was supported by the Wellcome Trust (U.K.).

References

[1] J.-D. Haynes and G. Rees, "Predicting the orientation of invisible stimuli from activity in human primary visual cortex," *Nature Neuroscience*, vol. 8, pp. 686–691, 2005.

[2] G. Garraux, C. Phillips, J. Schrouff, and E. Salmon, "Multiclass classification of FDG PET scans for the distinction between Parkinson's disease and atypical Parkinsonian syndromes," *NeuroImage Clinical*, vol. 2, pp. 883–893, 2013.

[3] A. Marquand, M. Brammer, S. Williams, and O. Doyle, "Bayesian multi-task learning for decoding multi-subject neuroimaging data," *NeuroImage*, vol. 92, pp. 298–311, 2014.

[4] M. Filippone, A. Marquand, C. Blain, S. Williams, J. Mourao-Miranda, and M. Girolami, "Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities," *The Annals of Applied Statistics*, vol. 6, no. 4, pp. 1883–1905, 2012.

[5] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, 2009.

[6] S. Majerus, N. Cowan, F. Peters, L. V. Calster, C. Phillips, and J. Schrouff, "Cross-modal decoding of neural patterns associated with working memory : Evidence for attention-based accounts of working memory," *Cerebral Cortex*, 2014.

[7] M. Dastjerdi, M. Ozker, B. L. Foster, V. Rangarajan, and J. Parvizi, "Numerical processing in the human parietal cortex during experimental and natural conditions," *Nature Communications*, vol. 4, 2013.

[8] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[9] T. Evgeniou, C. Michell, and M. Pontil, "Learning multiple tasks with kernel methods," *Journal of Machine Learning Research*, vol. 6, pp. 615–637, 2005.