Neurophysiology of perceived confidence

Martin Graziano, Lucas C. Parra (*IEEE Senior Member*), Mariano Sigman

Abstract— In a partial report paradigm, subjects observe during a brief presentation a cluttered field and after some time – typically ranging from 100 ms to a second – are asked to report a subset of the presented elements. A vast buffer of information is transiently available to be broadcasted which, if not retrieved in time, fades rapidly without reaching consciousness. An interesting feature of this experiment is that objective performance and subjective confidence is decoupled. This converts this paradigm in an ideal vehicle to understand the brain dynamics of the construction of confidence. Here we report a high-density EEG experiment in which we infer elements of the EEG response which are indicative of subjective confidence. We find that an early response during encoding partially correlates with perceived confidence. However, the bulk of the weight of subjective confidence is determined during a late, N400-like waveform, during the retrieval stage. This shows that we can find markers of access to internal, subjective states, that are uncoupled from objective response and stimulus properties of the task, and we propose that this can be used with decoding methods of EEG to infer subjective mental states.

I. Introduction

A vast ensemble of stimuli are continuously being processed in parallel by the sensory system, most of which elicit only a brief transient sensory response which fades after few hundred milliseconds without reaching working memory, executive control and consciousness [1], [2]. What determines the subset of this ensemble that reaches awareness has been a matter of intensive research.

An intriguing dissociation between conscious and unconscious information processing is that they seem to operate in different temporal scales. While conscious information can be sustained for seconds, the bulk of unconscious information processing decays very fast and lasts only for a few hundred milliseconds. For instance, using the partial report paradigm [3], Sperling showed that when observers saw briefly presented displays composed of several alphanumeric characters, only a few (3 to 5) elements reach consciousness and working memory. However, observers had a much better memory when asked to identify a specific subset of the characters at an interval (Inter Stimulus Interval, ISI) after the presentation of the visual display. This indicated the existence of a high capacity initial memory of the stimulus display which decayed a few hundred milliseconds after stimulus presentation, referred as Iconic Memory [4].

More recently, several studies have addressed empirically the contents of introspective and subjective estimates. Combining classic probes of metacognition with additive factor methodology, it was shown that introspective measures are highly reliable and thus that understanding which aspects of information processing are accessible to introspection and which are opaque can be determined with accurate quantitative precision, a methodology referred to as quantitative introspection [5]-[8].

In previous work we found that this methodology could detect strong dissociations between the objective response (the capacity to report the correct object) and the subjective response (the conscious perception of the subject about its response) [9]. Here we capitalize on this finding to understand the neurophysiological markers and neural dynamics of the construction of subjective confidence. We recorded high density EEG during an iconic memory experiment. The main focus of this experiment was to infer subjective confidence from EEG responses, i.e to identify neurophysiolgical responses which selectively distinguish trials in which subjects feel confident about their response (regardless of accuracy) from those in which they feel uncertain. We investigate these effects during the two critical stages of the experiment, the encoding of the cluttered scene and the retrieval after the presentation of the cue (Fig. 1).

Manuscript received April 30, 2010. This work was supported in part by the Human Frontiers Science Program and by the SECYT, PICT 38366. MG has a fellowship of the National Research Council of Argentina (CONICET).

M. G. is with the Laboratory of Integrative Neuroscience, School of Exact and Natural Science, University of Buenos Aires, Argentina. E-mail: marting@df.uba.ar.

L. C. P. is with the City College of the City University.of New York, 160 Convent Ave., New York, NY 10031. E-mail: parra@ccny.cuny.edu.

M. S. is with the Laboratory of Integrative Neuroscience, School of Exact and Natural Science, University of Buenos Aires, Argentina. Phone: 54-11-4576-3300 (282). Fax: 54-11-4576-3357. E-mail: sigman@df.uba.ar.

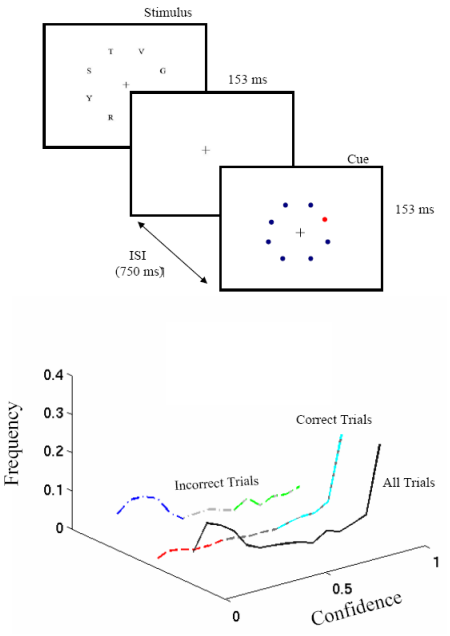

Fig. 1. Experimental Design. A circular array of eight letters was presented during 153 ms (participants fixated in a cross at the center of the array before the stimulus presentation for 1000-1500 ms, not shown). After a fixed delay of 750 ms (ISI), a small red circle (the cue) was presented in one of the locations of the array indicating the letter that had to be reported, after and additional 1000 ms waiting period following the cue. Participants had to also report with a mouse click the confidence level of their response on a visual analog scale (a horizontal bar placed at the center of the screen). The response ranged between 0% confidence (random guessing) and 100% (completely certain). The distribution of confidence reports for one representative subject is shown at the bottom of the panel, and it is also shown the same distribution restricted to correct or incorrect trials (in blue and green it is marked the low and high confidence range taken for this subject, and in red and cyan the same for correct trials).

II. METHODS

A. Data Collection - EEG

EEG was recorded at 512 Hz with a 128-electrode ActiveII Biosemi device. Fifteen adult subjects participated in the experiment. Each subject performed a session of 480 trials, divided in 6 blocks of 80 trials each. All signals were band-pass filtered (1-30Hz), re-referenced to a common mean, and the mean baseline activity during 300 ms before stimulus presentation was subtracted from each trial and each electrode. We rejected trials with voltage exceeding ±150μV and electrooculogram activity exceeding ±70μV. Three subjects were discarded because a significant fraction of trials contained noise from a bad reference electrode (all sessions conducted in the same day). In all other experiments the number of rejected trials was < 15% for all participants (average 6%).

B. Data Analysis

Data analysis was performed using Matlab and the EEGLAB toolbox.

Multiple projections: To capture the main components of the signal as they evolve in time several first principal components were computed collapsing across subjects with data from multiple times points in the stimulus-locked EEG. We divided these principal components (see Fig. 2) according to their degree of correlation and we selected two components that were least correlated (almost orthogonal), the P1 component (about 156 ms) for the stimulus, and the N400 component that appeared 342ms after the cue (1242ms after the stimulus). We regressed the trial averaged EEG – the evoked response potentials (ERP) – to these two components for each subject and computed mean values depending on the subject's behavioral response dividing correct from incorrect trials and confident from uncertain reports focusing on the N400 period during recall (Fig. 3B).

We use two different techniques to find single-trial correlations between subjective confidence and EEG activity: Linear Regression in space and time of raw EEG activity, and wavelet denoising of single trial projections.

Linear Regression: The activity of the 128 electrodes was correlated at the single trial level with the level of confidence at each trial, to seek for regions of significant correlation between EEG activity and confidence level of the subject Averaged activity in a 50ms window for each trial was regressed with reported confidence (Fig. 4A). Significance of the R^2 value against the null hypothesis was estimated for each time window by randomizing confidence values and repeating the regression. In this manner 500 random R^2 values are computed for each window and p-value is determined as the fraction of these random bootstrap trials above the correct R^2. P-values <0.01 were taken as significant (asterisks in Fig 2A). The same procedure was used for individual subjects but adding a repeated sample methodology to improve the estimation of the linear regression., namely, estimating regression with all samples within a time windows (instead of the mean) and combining all trials. This increases the number of samples available to compute regressors which is required here as the number of parameters (128) is large relative to the number of trials (480).

Wavelet denoising: We use the EP_den v2 program [10], [11], to de-noise the single trial ERP activity projected to the N400 component. We use Biorthogonal B-spline as the basic wavelet, and decompose the signal at 8 scales. The coefficients to de-noise the signal were selected by inspection. After de-noising, mean signal at a given time window was calculated to correlate with the subjective response at each trial (Fig 4B).

III. RESULTS

Participants were asked to indicate the identity of the letter which had appeared in the cued location. Subsequently, they were asked to rate their confidence in their response using a visual analog scale which ranged from: 100 (absolutely certain) to 0 (guessing). The distribution of confidence was bimodal for the majority of individual subjects and hence it could be easily parsed in high-confidence and low-confidence categories. While high-confidence errors and low-confidence correct responses had on average less counts than the compatible conditions (high-confidence correct, low-confidence errors), and because our point was not to characterize the distributions of subjective confidence over a population but to measure neurophysiological covariates of these distributions, we selected all the subjects that had sufficient trials in each category to assure an unbiased analysis (10/12).

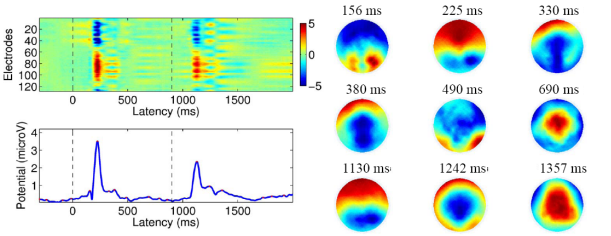

Fig. 2. Decomposing ERPs in a sequence of response components. Left: Raster plot of electrode activity averaged across subjects. Dashed lines indicates the time of stimulus (0ms) and cue onset (900 ms after stimulus onset). Right: Absolute mean activity averaged across electrodes. Red points indicate local peaks of this average amplitude signal. Right: Topography of components obtained analyzing the peaks in the previous plot.

To identify the main response components, we first examined the summed absolute value of voltages recorded over all electrodes from the grand average ERP (Evoked Response potentials averaged over all trials and all subjects, Figure 2). We identified a series of typical ERP response components aligned to the presentation of the stimulus and to the cue (900 ms after stimulus onset): we observed an early P1 and N1 waveforms sequence, and a P2 at 490ms for stimulus presentation. We also observed an N400-like waveform (central negativity) at 340ms and a relatively late P3-like (central positivity) at 600ms of stimulus onset and 450ms of cue onset, respectively. The sequence of these waveforms was reliable across individual subjects.

In order to investigate the neuro-physiologic correlates of the subjective report we first categorized the data in two binary factors: high and low confidence, and correct and incorrect responses. Analysis of absolute mean activity grouped by condition shows a clear distinction at 1180-1280ms between high and low confidence conditions irrespective of the objective response (Fig. 3A). This effect clearly distinguished confident from uncertain response, regardless of their correctness.

This result suggested that the topography of the N400 component may be a relevant direction in the 128-dimensional space of all channels encoding information about subjective confidence. To eliminate background noise, we performed a linear regression of the data to two nearly orthogonal components: P1 and N400 (most of the other components could be explained reliably by a linear combination of this basis functions). We then measured, for each individual subject and condition, the projection of the data to the N400 component and averaged these projections across subjects. An ANOVA analysis revealed a highly significant effect of confidence at (1180-1280ms after stimulus onset, $F = 7.98$, $df = 1$, $p < 0.01$) and no effect of the correct/incorrect factor ($F = 0.64$, $df = 1$, $p = 0.43$) (Fig. 3B). The analysis also revealed a less pronounced effect during the encoding stage.

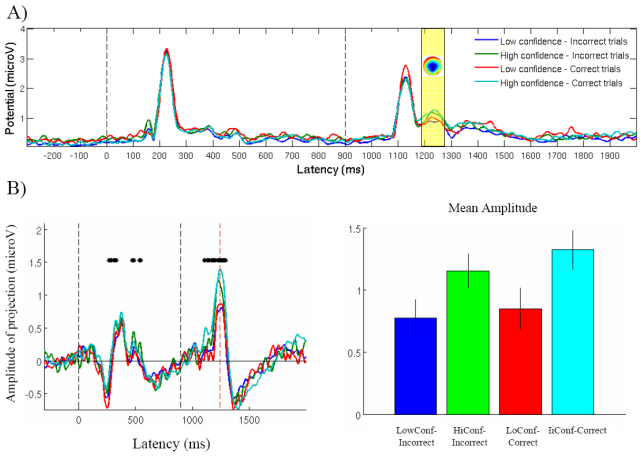

Fig. 3. N400 is modulated by subjective confidence. A) Absolute mean activity grouped by condition (blue: Low confidence-Incorrect trials, green: High confidence-Incorrect trials, red: Low confidence-Correct trials, cyan: High confidence-Correct trials). There is a clear difference at 1180-1280ms between high and low confidence conditions. B) Projection of the N400 component to the averaged subject activity, for each condition. Asterisks indicate significant ($p<0.01$) difference in a two-way ANOVA of confidence as main factor. Right panel: Mean amplitude of the projection for each condition in the 1180-1280 time window.

Our main motivation was to find, at the single trial level, a neural marker of the subjective value of confidence. We conducted this analytic effort in two complementary ways. First, we implemented a linear regression analysis of the raw data (Figure 4A) to find spatio-temporal regions of the signal that correlate linearly with confidence level in the trial (see methods and caption for a more detailed explanation of the technique and the procedure). In accordance with the factor analysis, we observed two main regions of interest: a moderate but significant effect during the initial sensory processing of the stimulus (overlapping with the P1 and P2 components), and a significant and large effect during the retrieval stage after the presentation of the cue (peaking at the N400 component). In a second stage, for each individual subject we performed a linear correlation analysis after

empiric based denoising of the data [11]. After restricting the single trial data in terms of spatial (linear projection to N400 topography), frequency and latencies as established by the average responses, we observed a very tight correlation between EEG responses and subjective confidence (mean 0.14 ± 0.03 - Pearson correlation ± err. standard; p-value<0.05 for 6 of 10 subjects).

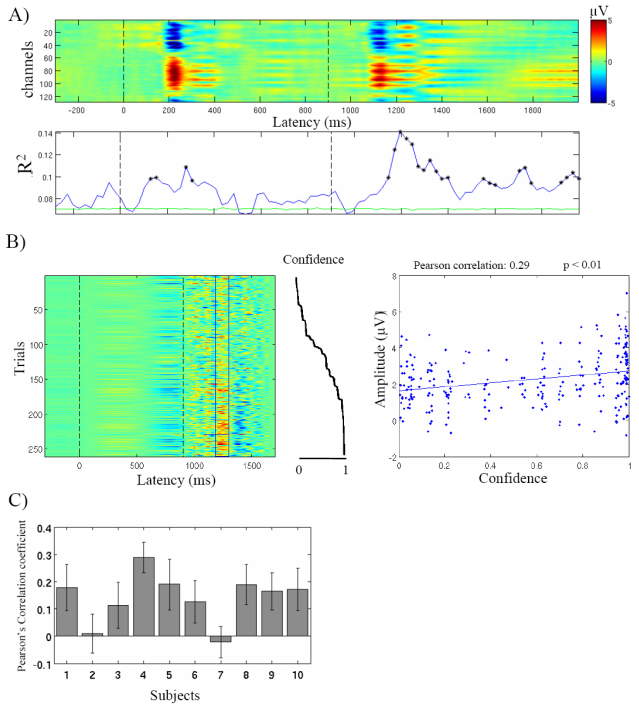

Fig. 4. Single Trial regression of subjective experience. Only correct trials were used for this analysis. A) Linear regression of confidence. Top panel: Electrode activity averaged across subjects. Bottom panel: R^2 values for the multiple linear regression using the 128 electrodes as independent variable and confidence as the dependent variable. Asterisks marks a significant correlation ($p<0.01$) between EEG and analog value of confidence. B) Example of projection using N400 component to single trials of a individual subject, sorted by the confidence level and filtered with appropriate wavelet coefficients (latency and frequency of N400). C) Pearson correlation coefficient for all subjects. We obtained a significant correlation ($p<0.05$) between confidence and the amplitude of the component in 6 of 10 participants.

IV. CONCLUSION

We investigated neurophysiologic markers of subjective report in humans, using a partial report paradigm in a cluttered visual field. Subjective confidence was partially indexed by a very early phase during the encoding process. However, the bulk of the determinant of subjective confidence was in a late N400-like wave during the retrieval process. A linear correlation analysis based on spatial and spectral filters was effective in indicating on a single trial basis, a subjective measure of confidence that was uncoupled from the explicit objective behavioral performance. This could be useful to build a linear decoder of EEG data to access internal, subjective states, which are uncoupled from objective response and stimulus properties of the task.

REFERENCES

[1] T. Shallice, *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press, 1988.

[2] M.I. Posner, Attention: the mechanisms of consciousness. *PNAS*, vol 91, 1994, pp. 15–64.

[3] G. Sperling, The information available in brief visual presentations. *Psychological Monographs: General and Applied*, vol 74, 1960, pp 1-29.

[4] U. Neisser, *Cognitive Psychology*. Englewood Cliffs, NJ: Prentice Hall, 1967.

[5] C. Kunimoto, J. Miller, H. Pashler, Confidence and accuracy of near threshold discrimination responses. *Conscious Cogn*, vol 10, 2001, pp. 294-340.

[6] N. Persaud, P. McLeod, A. Cowey, Post-decision wagering objectively measures awareness. *Nat Neurosci*, vol 10, 2007, pp 257-261.

[7] C. Sergent, S. Dehaene, Is consciousness a gradual phenomenon? Evidence for an all-or.none bifurcation during the attentional blink. *Psychol Sc*, vol 15, 2004, pp 720-728.

[8] G. Corallo, J, Sackur, S. Dehaene, M. Sigman, Limits on introspection: Distorted subjective time during the dual-task bottleneck. *Psychological Science*, vol 19(11), 2008, pp 1110-1117.

[9] M. Graziano, M. Sigman, The Spatial and Temporal Construction of Confidence in the Visual Scene. *PLoS One*, vol 4(3), 2009, pp. e4909.

[10] R. Quian Quiroga, Obtaining single stimulus evoked potentials with Wavelet Denoising. *Physica D*, vol 145, 2000, pp 278-292.

[11] R. Quian Quiroga, H. Garcia, Single-trial evoked potentials with wavelet denoising. *Clin Neurophysiol*, vol 114, 2003, pp 376-390.