A template-projection approach to decode higher-order vision in realtime and at the perceptual threshold

Kai J Miller
Department of Neurosurgery
Stanford University, Stanford, CA,
kai.miller@stanford.edu

Dora Hermes
Department of Psychology
Stanford University, Stanford, CA, USA

Abstract— **The link between object perception and neural activity in visual cortical areas is a problem of fundamental importance in neuroscience. We measured brain surface physiology with implanted electrocorticography (ECoG) electrodes in humans. Physiological responses to visual stimuli in object-specific ventral temporal loci are highly polymorphic in different cortical loci, for both broadband and raw potential trace changes. To address this, we developed a template-projection method, where averaged responses from a localizer task are projected into the continuous datastream recorded from the brain. These projections are used to build a feature space. A classifier for decoding visual perception is applied to this feature space during training periods, and is applied to plain images, as well as noise masked images. This enables robust classification of visual perceptual state.**

Keywords-electrocorticography; perception; classification;

I. Introduction

We describe a new technique for decoding perception from electrical potentials measured from the human brain surface. All previous attempts have focused on the identification of classes of stimuli or behavior where the timing of experimental parameters is known or pre- designated. However, real world experience is spontaneous, and to this end we describe an experiment predicting the occurrence, timing, and types of visual stimuli perceived by human subjects from the continuous brain signal. In this experiment, human patients with electrodes implanted on the underside of the temporal lobe were shown pictures of faces and houses in rapid sequence. We developed a novel template-projection method for analyzing the electrical potentials, where, for the first time, broadband spectral changes and raw potential changes could be contrasted as well as combined.

II. METHODOLOGY

A. Subjects:

All patients in the study were epileptic patients at Harborview Hospital in Seattle, WA. Sub-dural grids and strips of platinum electrodes were clinically placed over frontal, parietal, temporal, and occipital cortex for extended clinical monitoring and localization of seizure foci. They performed the tasks at the hospital bedside, with 10cm-wide pictures were displayed on a bedside monitor at ~1m from the patients, indicating task choice using a separate keyboard. All patients participated in a purely voluntary manner, after providing informed written consent, under experimental protocols approved by the Institutional Review Board of the University of Washington (#12193). All patient data was anonymized according to IRB protocol, in accordance with HIPAA mandate.

B. "Localizer task"

Subjects performed a basic face and house observation task. Subjects were presented with simple, grayscale, pictures of whole-field faces and houses that were displayed in random order for 400ms each, with 400ms inter-stimulus interval (blank screen) between. There were 3 experimental runs with each patient, with 50 house pictures and 50 face pictures in each run. Subjects were instructed to report an upside-down house.

C. "Noisy task"

Subjects also performed a face-detection task using phase-scrambled close-up pictures of faces and houses. Using a phase-scrambling technique, images of faces and houses were degraded in increments from 0 to 100% in 5%. A total of 630 stimuli were shown to each patient, 15 of each type at each 5% noise increment. Patients were instructed to report with a keypress when they saw a face. These stimuli were generously shared by the Ungerleider laboratory at the NIH, from a previous study of theirs [3].

D. Recordings:

Experiments were performed at the bedside, using Synamps2 amplifiers (Neuroscan, El Paso, TX) in parallel with clinical recording. Stimuli were presented with a monitor at the bedside using the general purpose BCI2000 stimulus and acquisition program. The potentials were sampled at 1000 Hz, with respect to a scalp reference and ground, and had an instrument-imposed bandpass filter from 0.15 to 200 Hz.

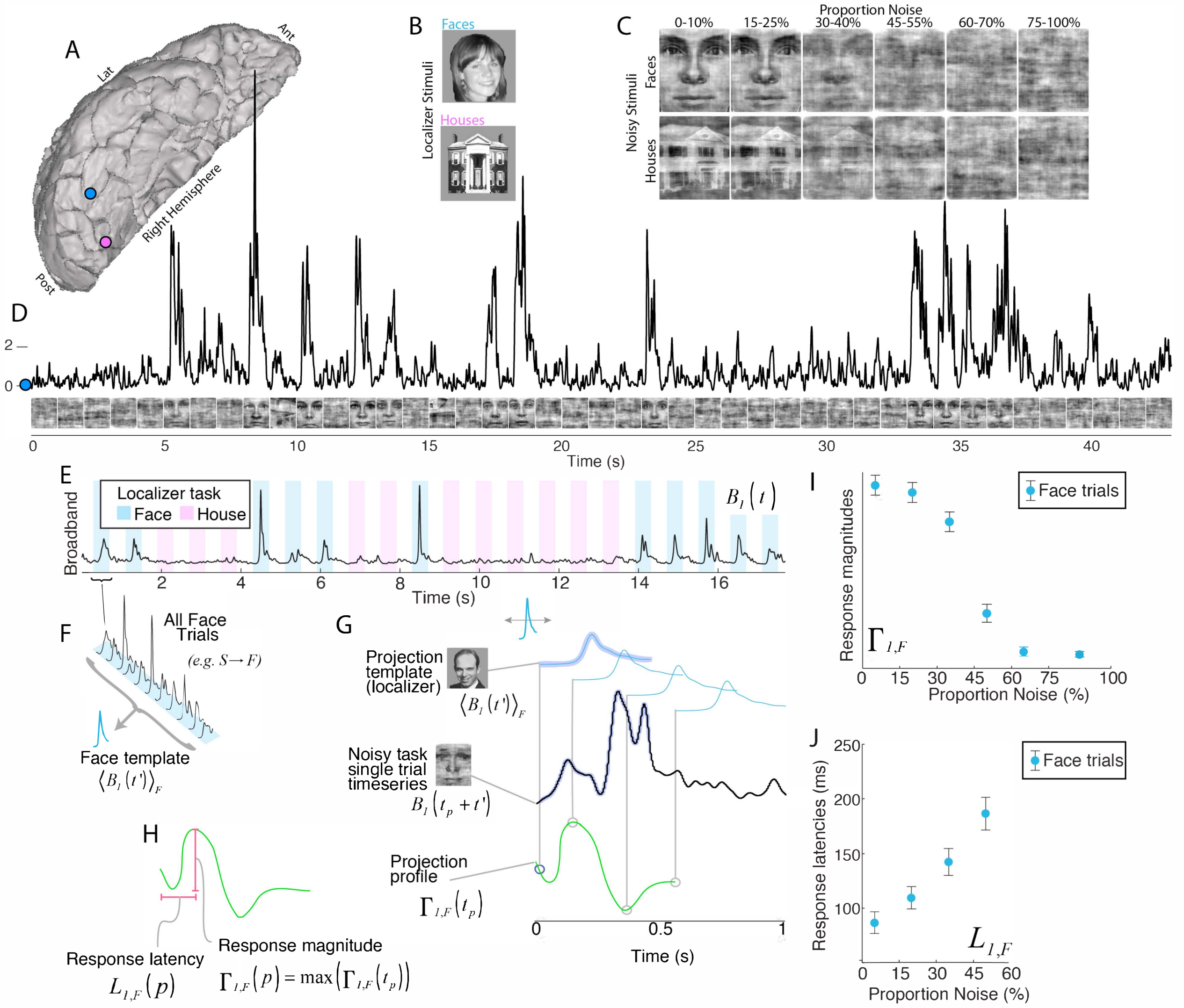

Figure 1: Tasks and template projection. **(A)** Ventral temporal ECoG was recorded from fusiform (blue) and lingual (pink) gyral electrodes (subject 1). **(B)** *Localizer task:* simple pictures of whole-field faces and houses were displayed in random order for 400ms each, with 400ms blank screen in-between. **(C)** *Noisy-task:* Phase-scrambled close-up pictures of faces and houses shown for 1s each. Stimuli ranged from 0-100% noise in 5% increments; subjects pressed a key when they believed a face was shown. **(D)** Broadband spectral change in the electrical potential, a reflection of averaged neuronal population activity, is shown above the corresponding stimuli from the noisy task (fusiform site in (A)). **E-J: Template projection to generate single trial response magnitudes and latencies.** **(E)** Broadband from a fusiform electrode in the localizer task. **(F)** Template face response generated by averaging across all face trials in the localizer task. **(G)** Templates were projected into the noisy task single trial, using a sliding dot product / covariance function over a 300ms interval of each noisy-task single trial to obtain a projection profile (green trace). **(H)** The "response magnitude" is the maximum of this projection profile, and the latency associated with this maximum, is the "response latency". **(I&J)** Single-trial response magnitudes and latencies can be plotted as a function of noise. Latencies are only plotted up to the 45-60% noise level, because the template projection is not stable for response traces of ~0 magnitude. *Figure modified from original form in [1].*

E. *Signal processing:*

Lateral frontoparietal electrode grids were discarded from analysis, and only inferotemporal strip electrodes were further considered. Electrodes with significant artifact or epileptiform activity were rejected. The electrical potential was then re-referenced with respect to the common average, and line noise was rejected with a notch filter.

F. *Decoupling the cortical spectrum to isolate broadband spectral change:*

The decoupling process to extract the timecourse of broadband spectral change is described in full detail and illustrated in previous manuscripts [4, 5]. Briefly summarized: From each electrode, discrete samples of power spectral density (PSD) were calculated, and normalized. Motifs in change of the normalized log-PSD were decomposed using a principal component approach. Eigenvectors of a singular value decomposition of these normalized log-PSD measures were calculated. The first eigenvector was applied to a normalized continuous time-frequency log-power approximations (dynamic spectra, calculated using complex Morlet wavelets), and this quantity was then z-scored and exponentiated, and then 1 was subtracted (setting the mean at 0) to obtain the "Broadband timecourse," which has been shown to reflect a power law in the cortical PSD [6]. This was performed independently for the "Localizer" and "Noisy" tasks.

G. Template projection technique:

Stimulus triggered averaged raw potential and broadband template: We obtain In each electrode n, stimulus-triggered averages of the raw voltage trace were obtained for the common-averaged electric potential for the face ($S \rightarrow F$, face) and house ($S \rightarrow H$, house) stimuli independently (τ_{k_s} denotes the k^{th} of N_S total instances of stimulus type S in the training set):

$$\left\langle V_n(t')\right\rangle_S^0 = \frac{1}{N_S}\sum_{k_s=1}^{N_S} V_n\left(\tau_{k_s}+t'\right).$$

This quantity is only calculated on the peri-stimulus interval defined by t'. It is then re-centered by subtracting the average potential peri-stimulus baseline to obtain $\left\langle V_n(t')\right\rangle_S$, which is commonly called the 'event-related potential' (ERP). We perform the same averaging over for the broadband signal to obtain $\left\langle B_n(t')\right\rangle_S$, which we refer to as the 'event-related broadband' (ERBB). Note that templates were obtained only from the localizer task, and for the classification of the localizer task data, templates were obtained separately for each cross-fold.

H. Discrete projection of templates into localizer task:

Training feature points were obtained by back-projecting $\left\langle V_n(t')\right\rangle_S$ and $\left\langle B_n(t')\right\rangle_S$ into the localizer task (separately into testing and training cross-folds for classification of the localizer task data) to obtain a set of training feature points, $\Gamma^V_{n,S}(q)$ and $\Gamma^B_{n,S}(q)$ for each event q at time τ_q:

$$\Gamma^V_{n,S}(q) = \sum_{t'=-199}^{400}\left\langle V_n(t')\right\rangle_S\left(V_n\left(\tau_q+t'\right)-\overline{V_n^b\left(\tau_q\right)}\right) \quad ,$$

where $\overline{V_n^b\left(\tau_q\right)}$ represents an "instantaneous" baseline surrounding time τ_q: $\overline{V_n^b\left(\tau_q\right)} = \sum_{t=-199}^{50} V_n\left(t+\tau_q\right)$. $\Gamma^B_{n,S}(q)$ were obtained in the same fashion. The training event types q were face picture stimulus onset ($q \rightarrow F$), house picture stimulus onset ($q \rightarrow H$), or randomly chosen points during the inter-stimulus interval (ISI, $q \rightarrow o$).

$\left\langle B_n(t')\right\rangle_S$ and $\left\langle V_n(t')\right\rangle_S$ were back-projected into the localizer task data to obtain a set of localizer feature points, $\Gamma_{n,S}(q)$ for stimulus presentations at time τ_q : $\Gamma_{n,S}(q) = \sum_{t'=1}^{600}\left\langle B_n(t')\right\rangle_S\left(B_n\left(\tau_q+t'\right)-\overline{B_n^b\left(\tau_q\right)}\right)$, where $\overline{B_n^b\left(\tau_q\right)}$ represents an "instantaneous" baseline: $\overline{B_n^b\left(\tau_q\right)} = \sum_{t=1}^{100} B_n\left(t+\tau_q\right)$. The event types were face picture stimulus onset or house picture stimulus onset.

I. Projection of templates into continuous data stream (localizer task)

To quantify how well the averaged raw potential $\left\langle V_n(t')\right\rangle_S$ (training cross-fold) is represented in the voltage time series of the testing data at time *t*, it is directly forward-projected onto the continuous time series (testing cross-fold) at each millisecond:

$$\Gamma^V_{n,S}(t) = \sum_{t'=-199}^{400}\left\langle V_n(t')\right\rangle_S\left(V_n\left(t+t'\right)-\overline{V_n^\tau(t)}\right), \text{ where}$$

$\overline{V_n^\tau(t)}$ was obtained in the same fashion as above. The same projection is performed for the broadband template $\left\langle B_n(t')\right\rangle_S$, to obtain $\Gamma^B_{n,S}(t)$.

J. Projection of templates into noisy task:

In order to quantify the single trial response in the noisy task, the broadband projection templates (generated from the localizer task) were applied to each stimulus presentation, scanning through a variable delay with respect to each stimulus onset:

$$\Gamma_{n,S}\left(t_p\right) = \sum_{t'=1}^{600}\left\langle B_n(t')\right\rangle_S\left(B_n\left(t_p+t'\right)-\overline{B_n^b\left(t_p\right)}\right) \quad , \text{ where } t_p$$

ranges over the 0-300ms interval, and $\overline{B_n^b\left(t_p\right)}$ is the instantaneous baseline obtained for each timepoint. For each such trial, the "Projection magnitude" is $\Gamma_{n,S}(p) = \max\left(\Gamma_{n,S}\left(t_p\right)\right)$, and the time of this maximum value – the "Projection latency" – is denoted $L_{n,S}(p)$.

K. Generation of a feature space

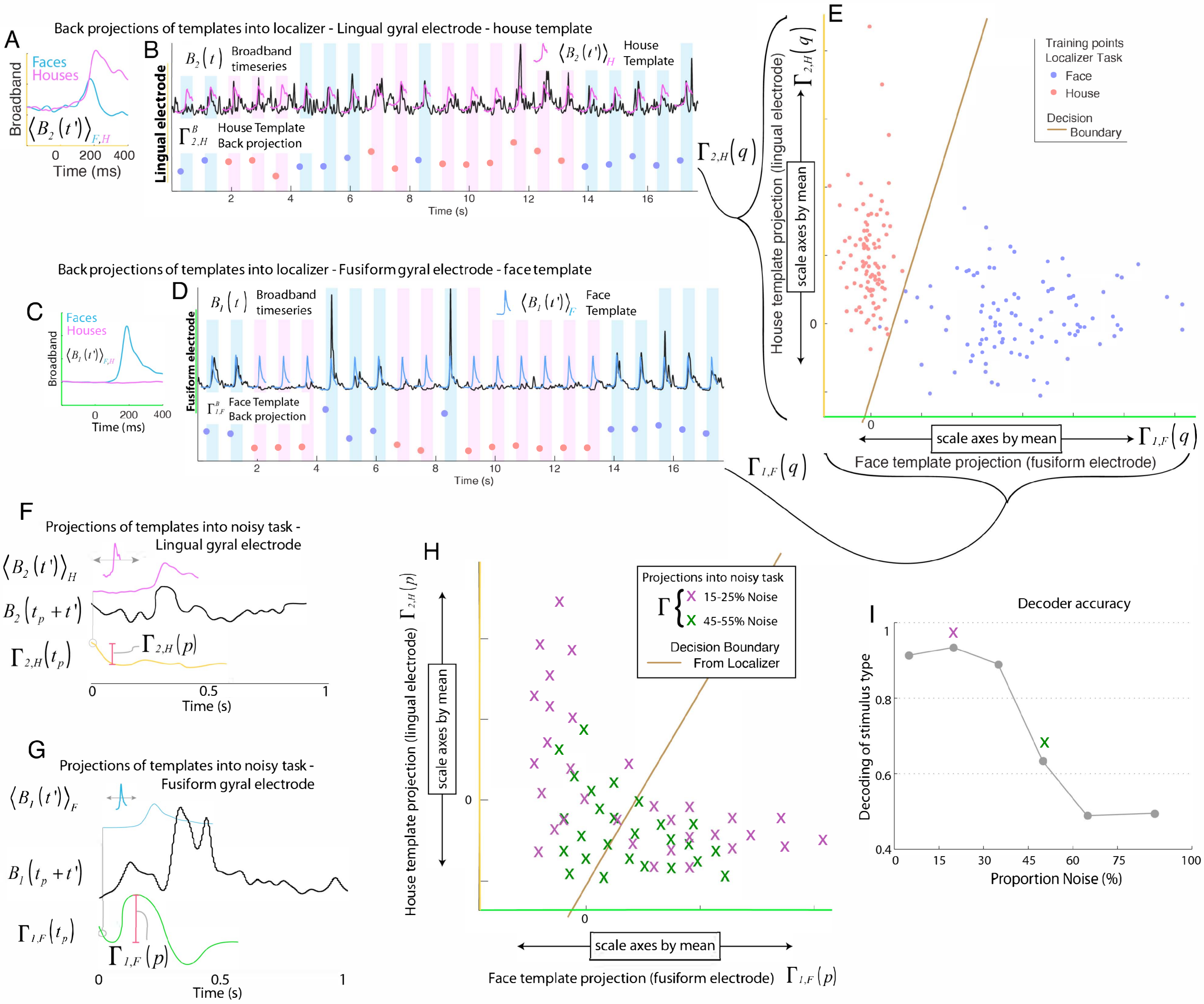

Figure 2: Decoder is built upon localizer task, and applied to noisy task. (A) House-image selective template (pink) from a lingual gyral electrode is generated as illustrated in Figure 2. **(B)** The template from (A) is back-projected into the broadband timeseries from that electrode at the time of each image presentation (note: there is no projection profile, as the projection is aligned to the time of stimulus onset, as described in the text), where trials can be of type face or house. **(C&D)** As in (A&B), but for a face template in a fusiform site. **(E)** An example 2-D feature space built from back-projections into the localizer task. After scaling each feature by its' mean, a decoder (Fisher linear discriminant based classifier) is built within this feature space to distinguish face image presentations from house image presentations. **(F&G)** As illustrated in Figure 2, localizer templates are projected into single trials of the noisy task to obtain response magnitudes. **(H)** After scaling each response magnitude feature by its' mean, the decoder built on back-projections of the localizer task (E) was applied to predict whether a face or house image had been seen. **(I)** The output of the decoder can be compared with stimulus type, and subject choice, at different levels of image noise. *Figure modified from original form in [1].*

The full feature space for classification, consisting of the projections of the stimulus triggered broadband across all electrodes (n), for face and house templates independently, is the combination of $\Gamma_{n,F}$ and $\Gamma_{n,H}$. For brevity, we can combine the notation to denote each feature as Γ_m, where m represents a unique combination of one electrode n , and F or H . Each feature was scaled with division by its mean, $\Gamma_m \to \Gamma_m / \overline{\Gamma_m}$, for the localizer and noisy tasks independently.

Many of these features will not be particularly informative about when and how the brain is processing these visual stimuli. Therefore, features were downselected by independently assessing the squared cross-correlation between face vs house from the localizer task, and rejecting those which fell beneath the pre-defined threshold $r_m^2 < 0.10$.

The squared cross-correlation to compare face and house stimuli is: $r_m^2 = \frac{\left(\overline{\Gamma_m(q=F)} - \overline{\Gamma_m(q=H)}\right)^2}{\sigma_m^2} \frac{N_F * N_H}{N_{FH}^2}$, where σ_m is the standard deviation of the joint distribution for face and house stimuli $\Gamma_m(q=F\,\&\,H)$, N_F is the number of face presentation events, N_H is the number of house events, and $N_{FH} = N_F + N_H$.

L. *Classification:*

For the sake of simplicity, Fisher linear discriminant analysis (LDA) was used for classification. From the localizer task, the classifier is trained based upon the means and covariances of the full joint distribution and the sub-distributions $\Gamma_m(q \to F)$, $\Gamma_m(q \to H)$, and $\Gamma_m(q \to o)$ (i.e. as if they are normally distributed).

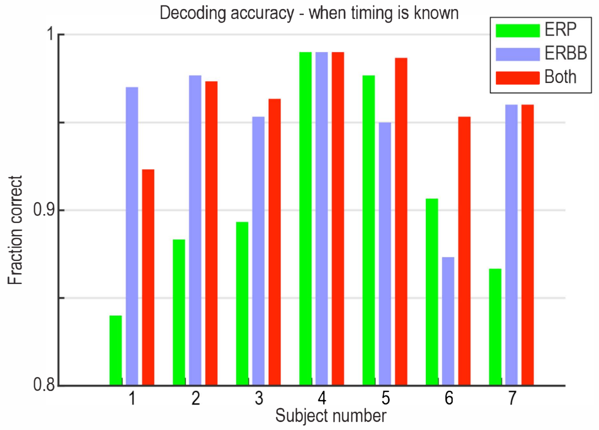

Figure 3: Classification accuracy for 7 patients when the onset of a stimulus is known, using ERP, ERBB, or both template types. In some subjects, 100% accuracy was reached. As was the case in the initial manuscript, all accuracies were above 90% when both raw potential and broadband templates were used. *Figure modified from original form in [2].*

Given the feature space of the training distribution (localizer task), single trials (p , from times t_p) from the noisy or localizer task can be assigned a posterior probability of belonging to the face- or house- distribution: $\Pr\{\Gamma_m(p)|q \to F\}$ or $\Pr\{\Gamma_m(p)|q \to H\}$. The higher posterior probability is the one that is chosen.

For the case of spontaneous classification of the continuous datastream, classification is performed at each ms, and local maxima of $\Pr\{\Gamma_m(t)|q \to F\}$ and $\Pr\{\Gamma_m(t)|q \to H\}$ are chosen to predict the timing and type of visual stimulus shown.

III. Results

A. *Cross-validated classification of localizer task data, when timing is pre-defined*

Initially, we classified the localizer data, pre-defining times of stimulus presentation, using 3-fold cross-validation (Figure 3). The averages by template type were 0.91 (ERP), 0.95 (ERBB), and 0.96 (Both ERP&ERBB).

B. *Cross-validated spontaneous classification of both object type and stimulus timing from localizer task data*

When we classified the continuous datastream we could predict the timing and type of visual stimuli shown (Figure 4). Average accuracies and errors for each template type were:

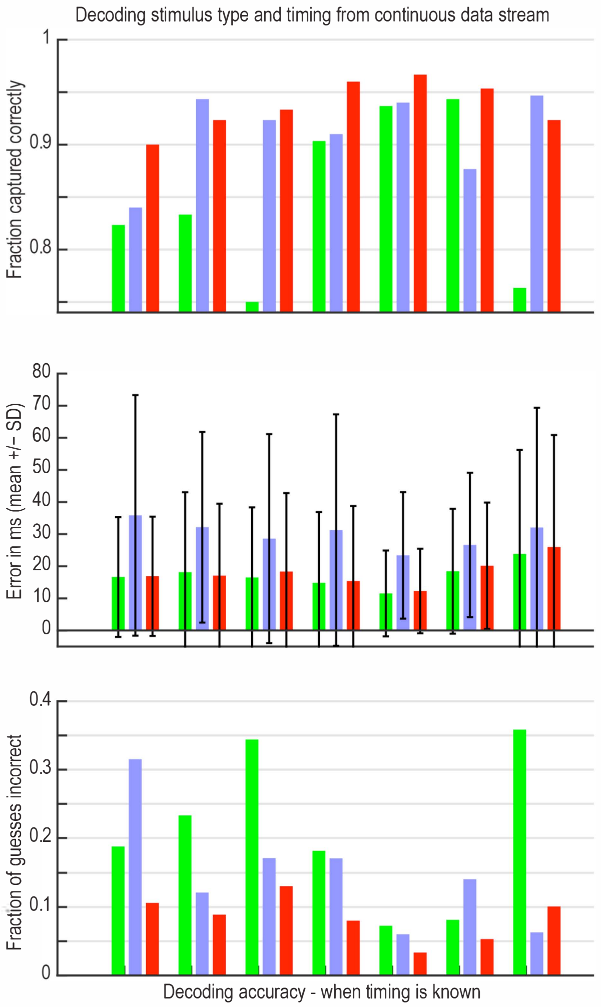

Figure 4: Classification accuracy for decoding stimulus class and onset in a continuous data stream. When both features were used (red bars), approximately 94% of all stimuli were captured correctly in every subject, with 15–20 ms error. At least 90% of stimuli were captured in all cases when both template types were used. An average of 8% of predictions using both features were incorrect (i.e., predicted stimuli at the wrong time, or as the wrong class. *Figure modified from original form in [2].*

- ERP: 85% captured, 17ms error, 0.21 false prediction rate.
- ERBB: 91% captured, 30ms error, 0.15 false prediction rate.
- Both ERP&ERBB: 94% captured, 18ms error, 0.08 false prediction rate.

C. *Decoding of noisy-task response magnitudes using a classifier built from the localizer task*

Decoder performance paralleled the patients' performance, and robustly predicts the stimulus type up to the perceptual threshold, before falling to chance (Figure 5).

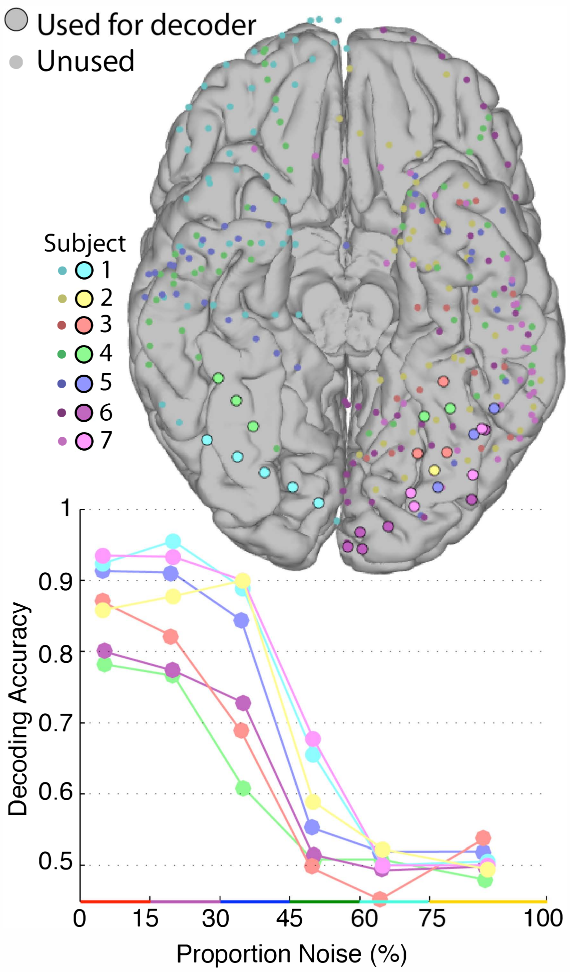

Figure 5: Decoding accuracy for the noisy task with a classifier built on the localizer task. Note that the 7 patients in this study were not the same as that shown in figures 3&4. *Figure modified from original form in [1].*

ACKNOWLEDGMENT

This manuscript is written as a long form abstract for the The 5th International Winter Conference on Brain-Computer Interface in High1 Resort, Korea, in January of 2017. It is excerpted from two existing works, titled "*Face percept formation in human ventral temporal cortex" [1]* and "*Spontaneous Decoding of the Timing and Content of Human Object Perception from Cortical Surface Recordings Reveals Complementary Information in the Event-Related Potential and Broadband Spectral Change" [2]*, from which it is entirely derivative. The original forms of all of these manuscripts lie under the copyrights of other journals they were published in, and the content is repurposed here by the original author. These original works had a number of other authors who deserve any credit for this content: Franco Pestilli, Gagan S Wig, Jeffrey G Ojemann, Rajesh PN Rao, and Gerwin Schalk. This research would not have been possible without the dedication of the patients and staff at Harborview Hospital in Seattle, WA.

REFERENCES

[1] Miller, K.J., et al., Face percept formation in human ventral temporal cortex. In Revision at Journal of Neuroscience, 2017, where it was submitted in February 2016.

[2] Miller, K.J., et al., Spontaneous Decoding of the Timing and Content of Human Object Perception from Cortical Surface Recordings Reveals Complementary Information in the Event-Related Potential and Broadband Spectral Change. PLoS computational biology, 2016. 12(1): p. e1004660.

[3] Heekeren, H.R., et al., A general mechanism for perceptual decision-making in the human brain. Nature, 2004. 431(7010): p. 859-62.

[4] Miller, K.J., et al., Broadband changes in the cortical surface potential track activation of functionally diverse neuronal populations. Neuroimage, 2014. 85 Pt 2: p. 711-20.

[5] Miller, K.J., et al., Decoupling the Cortical Power Spectrum Reveals Real-Time Representation of Individual Finger Movements in Humans. Journal of Neuroscience, 2009. 29(10): p. 3132.

[6] Miller, K.J., et al., Power-law scaling in the brain surface electric potential. PLoS computational biology, 2009. 5(12): p. e1000609.