Supplemental Data

Explicit Encoding of Multimodal Percepts
by Single Neurons in the Human Brain
R. Quian Quiroga, A. Kraskov, C. Koch, and I. Fried

Supplemental Experimental Procedures

Recordings and paradigm

The data comes from 16 sessions in 7 patients with pharmacologically intractable epilepsy (5 right handed, 5 male, 21 to 54 years old). Extensive non-invasive monitoring did not yield concordant data corresponding to a single resectable epileptogenic focus. Therefore, they were implanted with chronic depth electrodes for 7-10 days to determine the seizure focus for possible surgical resection [1]. All studies conformed to the guidelines of the Medical Institutional Review Board at UCLA and the Institutional Review Board at Caltech. The electrode locations were based exclusively on clinical criteria and were verified by MRI or by CT co-registered to preoperative MRI. Each electrode probe had a total of 9 micro-wires at its end, 8 active recording channels and one reference. The differential signal from the micro-wires was amplified using a 64-channel Neuralynx system (Tucson, Arizona), filtered between 1-9000 Hz and sampled at 28 kHz. Each recording session lasted about 30 minutes.

Subjects sat in bed, facing a laptop computer where pictures, text or sounds were presented. In order to enforce attention to the stimuli, subjects had to respond whether the picture, sound or text corresponded to a person or not by pressing the ‘Y’ and ‘N’ keys, respectively. To maximize the probability of getting neuronal responses, the stimuli used were chosen from previous ‘screening sessions’ where a set of about 100 different pictures of people well known to the subjects, landmarks, objects and animals were shown for 1 second, 6 times each in pseudo-random order [2]. The pictures used in the screening sessions were partially chosen according to the subject’s preferences. After a fast analysis of the data, it was determined which of the presented pictures elicited responses in at least one unit. All the stimuli eliciting responses in the screening sessions (i.e. a given person, landmark, object or animal) were included in the subsequent ‘invariance sessions’ reported here. In the invariance
sessions, an average of 16.7 (s.d.: 3.1; range: 11-20) individuals or objects were presented. For each of them, three different pictures as well as their names written in the laptop screen and spoken by a computer synthesized voice were presented in pseudorandom order, 6 times each. The pictures and written names covered about 1.5° of visual angle and were shown for 1 second. The spoken names were generated with commercial available software (AT&T Labs Inc., Text-to-Speech) and were played by the laptop at normal speed and at a volume that was clearly audible to the subjects. The duration of the sounds was variable according to the particular name that was said. After the experiments, the subjects confirmed that they knew the names of these persons, landmarks, objects and animals.

Spike detection and sorting

Spike detection and sorting was performed using ‘Wave_clus’ [3, 4]. After bandpass filtering the data between 300-3000 Hz, spikes were detected using an automatic amplitude threshold. A total of 64 datapoints, corresponding to approximately 2.3 ms, were stored for each detected spike. All spike shapes were aligned to the maximum at datapoint 20 and to avoid possible misalignments due to insufficient sampling rate, this was done after up-sampling the original shapes by a factor 2, using interpolation with cubic splines.

After detection, spike sorting was in two steps: i) the wavelet transform was used to extract relevant features of the spike shapes, which were the inputs to the clustering algorithm, and ii) clustering; i.e. assigning spikes with similar shapes to the same neuron, was done using super-paramagnetic clustering, a stochastic method that does not assume any particular distribution of the clusters [5]. Super-paramagnetic clustering groups the data into clusters as a function of a single parameter, the temperature, which can be changed by the user if the automatic clustering is not satisfactory [6]. Finally, the clusters were classified into single- or multi-units based on: 1) the spike shape and its variance; 2) the ratio between the spike peak value and the noise level; 3) the ISI distribution of each cluster; and 4) the presence of a refractory period for single-units; i.e., < 1% spikes within less than 3 ms ISI. The code for spike detection and sorting is available at www.le.ac.uk/neuroengineering.

Figure S1 shows the output of the Wave_clus code from a single channel recording with several units. The algorithm automatically detected 3 clusters. The first cluster (in blue)
corresponds to a multiunit; i.e.: spikes from several neurons that cannot be separated due to low signal to noise ratio. The other 2 clusters (in red and green) correspond to 2 single neurons. There were 267 detected events not assigned to any cluster (cluster 0, in black). The neuron in red fired to one of the researchers performing experiments at UCLA, as shown in Figure 3. The neuron in green fired to pictures of the actor Michael Douglas (see Figure S6) and the multiunit in blue fired to pictures of a daughter of the patient (not shown).

Figure S1: Output of Wave_clus. The upper subplot corresponds to a 60 sec segment of the band pass filtered continuous data and the automatic threshold used for spike detection (in red). The leftmost middle panel shows the values of the first two wavelet coefficients chosen by the algorithm for all the spikes. The remaining middle panels show the corresponding spike shapes of the sorted units, including the number of events in each cluster. The lowest panels disclose their inter-spike interval distribution (ISI). The leftmost lower plot shows the number of members in each cluster as a function of the clustering “temperature”.
As in the example of Figure S1, the type of neurons described here have usually a nearly silent baseline activity and are, therefore, very difficult to detect without presenting the exact stimuli eliciting responses – as determined with the screening sessions - followed by optimal spike sorting to dissociate the firing of sparse firing neurons from that of nearby neurons with high firing rates recorded from the same electrode.

**Responsiveness criterion**

The response to a picture was defined as the median number of spikes across trials in the first second after stimulus onset. Similarly, the baseline for each picture was the median number of spikes in the second before stimulus onset. A unit was considered responsive if the activity to at least 1 stimulus fulfilled three criteria: 1) the median number of spikes was larger than the average baseline (for all stimuli) plus 5 standard deviations; 2) the median number of spikes was at least 2; and 3) a t-test comparing the baseline and response period for the particular stimulus showed a significant difference with p<0.05. The same criterion of responsiveness was used for the text presentations. Since the sound stimuli elicited later responses, in this case we used the same criterion but with the response interval going from 500ms to 1500ms after stimulus onset.

We should remark that the above definition is not meant to be a formal test but rather a heuristic criteria, as proposed in previous studies [2, 7, 8]. The core of this definition is given by the criterion that the median number of spikes should be at least the average number of spikes for the baseline plus 5 standard deviations. We took the median, instead of the mean, to avoid contaminations of outliers (e.g. spontaneous bursting activity in some of the trials, as for picture 38, 41 and 29 in Figure S2). For comparison, we also calculated the number of responses using a threshold of 4 s.d. and 6 s.d., and obtained similar results (see below). Note, however, that based on a visual inspection of the raster plots-, the 4 s.d. threshold in Figure S2 seems to give more false positives (e.g. pictures 3, 31 and 35). Also based on visual inspection, the 6 s.d. threshold seemed to miss responses and therefore, the 5 s.d. threshold was heuristically chosen as a good compromise giving a low number of misses (as obtained with high thresholds) and false positives (as obtained with low thresholds). The criterion that the
median number of spikes should be at least 2 was added due to the fact that some neurons had very low baseline firing and even a threshold 5 s.d. above the baseline firing could still be low (as in Figure S2). This eliminates false positives such as the response to picture 35 in Figure S2. Finally, the T-test criterion was added to avoid cases with large firing in some of the trials starting in the baseline period (e.g. pictures 31, 3, 41, 29). Note that given the very low number of trials for each picture we do not give any meaning to the p-value given by the T-test (which would also require testing the presence of a Gaussian distribution) and we rather use it to heuristically eliminate some false positive responses.

A comparison of results with 3 other criteria for responsiveness is given in Figure S4 and Table S2.

Figure S2: Responses of the neuron shown in Figure 1 for 20 pictures. Conventions are the same as for Figure 1. For simplicity, the text and sound responses are not shown. The bottom
plot shows the median number of spikes for each picture. The red bars mark the responses to the pictures of Oprah Winfrey and the horizontal lines mark three different thresholds for defining responses (mean plus 4 s.d., 5 s.d. and 6 s.d. above the baseline activity). The median response for picture 40 (11 spikes) was trimmed for better visualization.

**Visual invariance: ROC analysis**

Single cell recordings studies in monkey IT have shown some degree of visual invariance (for reviews see [9-11]). Invariance in these cases is usually established by showing significant single neuron responses to presentations of stimuli at different sizes and at different positions. It has to be noted that such analysis does not strictly test for invariance, since it is evaluated whether a neuron has a significant response to different versions of a given stimulus, but not whether the neural response is the same. This for example can be evaluated using a decoding framework, with which we have shown that from the firing of invariant single cell responses in the human MTL it was possible to predict the identity of the person or object shown to the subject, but not the particular picture shown [7]. Following the terminology used by previous studies, here by visual invariance we mean a neuron responding to the pictures of a given person or object, disregarding whether the responses are statistically the same.

As previously described [2], to test for visual invariance –i.e.: a preferred firing to the pictures of a given person or object-, whenever a unit had a response to a given picture we used a Receiver Operator Characteristics (ROC) analysis to test whether the unit fired preferentially – but not necessarily uniquely - to the 3 pictures of the particular individual or object.

To illustrate the ROC procedure, let us consider the responses of the neuron in Figure S2. The hit rate (y-axis) is defined as the number of responses to the individual (Oprah Winfrey) divided by the total number of pictures of this individual (3 pictures). The false positive rate (x-axis) is defined as the number of responses to the other pictures divided by the total number of other pictures. The ROC curve (Figure S3) is obtained by gradually lowering the threshold of the responses (the median number of spikes in the bottom plot of Figure S2). Starting with a very high threshold (no hits, no false positives; red asterisk in Figure S3), if a unit responds exclusively to images of a particular person or object (e.g. Oprah Winfrey), the ROC curve will show a steep increase when lowering the threshold (a hit rate of 1 and no false
Note in Figure S3 that the second highest threshold (orange asterisk) gives 1 out of 3 hits and no false positives, the next lower threshold gives one extra hit and one false positive (light blue asterisk) and the smallest threshold gives both a hit rate and false positive rate of 1. If, in contrast to the case of Figure S3, a unit responds to a random selection of pictures, it will have a similar number of hits and false positives and the ROC curve will fall along the diagonal. For a highly visual invariant unit, the area under the ROC curve will be close to 1, while in the latter case it will be about 0.5. For Figure S3 the area under the ROC curve was 0.96. To evaluate the statistical significance, we created 99 surrogate curves, testing the null hypothesis that the unit responded preferentially to 3 randomly chosen pictures. In other words, each surrogate curve was created by randomly choosing 3 pictures of the whole dataset (in the case of Fig. S3 possibly including pictures of Oprah Winfrey). A unit was considered invariant to a certain individual or object if its area was larger than the area of any of the 99 surrogates (i.e. with a confidence of p<0.01).

**Figure S3:** ROC curve for the neuron in Figure S2 (in blue) and for 99 surrogates (in light grey) testing for preferred responses compared to random combinations of 3 pictures. Note that not all individual surrogate curves are visible due to superposition. The area under the ROC curve was of 0.96 and the maximum area of the 99 surrogates was of 0.75.
Statistical analysis

Statistical differences in the response characteristics for the different MTL areas were evaluated using a Fisher Exact Test implemented in R (the R project for statistical computing: http://www.r-project.org/), based on [12, 13]. In particular, we tested for the null hypotheses that in all four MTL areas there were the same relative number of: 1) responsive units, 2) units with invariant responses to pictures, 3) units with responses to spoken names, 4) units with responses to written names and 5) units with visual invariance and responding to the spoken and written names (triple invariance). Very similar results were obtained using a Chi-Square analysis, however, it is usually argued that this test may give unreliable p-values for cases with zero entries (but see [14], pp: 504-505), as for the sound, text and triple invariance responses.

As shown in Table S1 there were significant differences across the MTL areas for the number of sound, text and triple invariance responses in parahippocampal cortex.

<table>
<thead>
<tr>
<th></th>
<th>All areas</th>
<th>Hipp.</th>
<th>Entorhinal cortex</th>
<th>Amyg.</th>
<th>Parahipp. Cortex.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of units</td>
<td>750</td>
<td>255</td>
<td>172</td>
<td>216</td>
<td>107</td>
<td>-</td>
</tr>
<tr>
<td>Responsive units</td>
<td>79</td>
<td>21</td>
<td>17</td>
<td>22</td>
<td>19</td>
<td>n.s.</td>
</tr>
<tr>
<td>Visual Invariance</td>
<td>53 (67%)</td>
<td>18 (86%)</td>
<td>12 (70%)</td>
<td>13 (59%)</td>
<td>10 (53%)</td>
<td>n.s.</td>
</tr>
<tr>
<td>Sound responses</td>
<td>25 (32%)</td>
<td>10 (48%)</td>
<td>8 (47%)</td>
<td>7 (32%)</td>
<td>0 (0%)</td>
<td>p&lt;0.01</td>
</tr>
<tr>
<td>Text responses</td>
<td>24 (30%)</td>
<td>10 (48%)</td>
<td>10 (59%)</td>
<td>4 (18%)</td>
<td>0 (0%)</td>
<td>p&lt;10^-4</td>
</tr>
<tr>
<td>Triple invariance</td>
<td>17 (21%)</td>
<td>8 (38%)</td>
<td>6 (35%)</td>
<td>3 (14%)</td>
<td>0 (0%)</td>
<td>p&lt;0.01</td>
</tr>
</tbody>
</table>

Table S1: Number of recorded units; responsive units; units with visual invariance; units with significant responses to sound and text presentations and units with triple invariance (visual invariance, plus responses to sound and text) for all MTL areas we recorded from. Numbers in brackets are percentage number of units, relative to the number of responsive units for each
area. The rightmost column gives the p-values of Fisher Exact Test evaluating statistical
differences between the MTL areas.

Population results: dependence of the responsiveness criterion

To verify that these results were not the effect of the particular criterion for defining
responses, we also used 3 other responsiveness criteria: i) same as above but using 4 s.d.; ii)
same as above but using 6 s.d. and iii) a U-Test criterion, as defined in [15]. For the latter
definition, the 100 ms response period was divided into 19 half-overlapping 100 ms bins and
for each bin we compared the spike count for the 6 presentations of each stimulus to similar
baseline intervals for all pictures by means of a two-tailed Mann-Whitney U-Test, using the
Simer procedure to correct for multiple comparisons and applying a conservative significance
threshold of p=0.001 to reduce false positives [15].

As shown in Figure S4, results were similar for all 4 criteria of responsiveness. In
particular, the relative number of neurons with sound, text and triple invariance responses was
larger in the hippocampus and entorhinal cortex, followed by the ones in the amygdala, and
was smallest for the neurons in the parahippocampal cortex. Moreover, the statistical results
described in the main text were very similar for all responsiveness criteria (see Table S2), with
the only exception that with the U-Test the relative number of cells with triple invariance was
not significantly different for the different MTL areas (although there was a close to significant
trend with p=0.07).
Figure S4: Relative number of neurons with visual invariance, sound responses, text responses and triple invariance obtained with the different responsiveness criteria (see text).
Latency of MTL responses

Next, we studied the onset latency and duration of the responses to picture, text and sound stimuli. For this, the instantaneous firing rate was calculated by convolving the normalized spike trains with a Gaussian window of 60 ms width and normalizing, for each neuron, to the maximum response (to either picture, text or sound). To avoid overemphasizing the firing pattern of neurons with more than one response, for each neuron only the person/object eliciting the largest activation was considered. Then, for each response we estimated the latency onset as the point where the instantaneous firing rate crossed 4 s.d. above the baseline for at least 100ms. The duration was estimated as the difference between onset and offset, where the offset was defined as the point where the instantaneous firing rate crossed the 4 s.d. threshold downwards for at least 100 ms. Similar results were obtained using a threshold of 3 s.d. or of 5 s.d. and when using a Gaussian window of 100 ms, 30 ms and of 20 ms width. Statistical differences between the picture, text and sound presentations -both for the onset latency and the duration of the responses- were evaluated with a one-way analysis of variance.
(ANOVA). In separate tests, the same analysis was implemented to evaluate significant differences across the MTL areas for the responses to pictures, text and sound.

There were significant differences in the onset latencies of the 3 type of stimuli (ANOVA, $p<10^{-15}$). The earliest responses were to pictures (298ms; s.d.: 120ms), followed by the text (380ms; s.d.: 118ms) and sound stimuli (517ms; s.d.: 176ms). Post-hoc comparisons showed that all pair-wise differences were statistically significant (T-test, $p<10^{-3}$ for all cases). For the duration of the responses, there were no significant differences between the 3 type of stimuli (ANOVA, $p=0.21$; mean duration: 813ms, s.d.: 555ms).

The differences in the onsets for the picture, text and sound responses could be attributed to the different times needed to recognize a given person or object when seeing an image, reading the name or hearing the name pronounced. In particular, a visual stimulus becomes instantly available and consequently the stimulus onset is clearly defined. In contrast, the appearance of auditory stimuli takes time, which is also variable according to the particular name said. For example, it may take longer to recognize the name “Tower of Pisa” compared to “Oprah”. Another problem is that in many cases part of the pronounced name can be redundant, as for example, we may need to hear “Arnold” to realize it refers to “Arnold Schwarzenegger”. The ‘effective’ onset and duration of each sound stimuli cannot be objectively determined since it varies with each subject and with the other stimuli used (e.g. one needs to hear the second name to distinguish “Arnold Schwarzenegger” from “Arnold Palmer”, the golf player).

To evaluate regional effects, in separate tests we compared the onset latency and duration across the different MTL areas, for the picture, text and sound presentations. For the pictures, the onset latencies were statistically different (ANOVA, $p<0.01$), with the earliest responses occurring at the parahippocampal cortex (245ms; s.d.: 113ms), followed by the responses in the amygdala (298ms; s.d.: 124ms) and the ones in the hippocampus (311ms; s.d.: 98ms) and entorhinal cortex (334ms; s.d.: 171ms). Post-hoc analyses showed that these differences were due to the significantly earlier responses in the parahippocampal cortex in comparison to the ones in the amygdala (T-test, $p<0.05$) and the ones in the hippocampus and entorhinal cortex (T-test, $p<0.005$ in both cases), which is in agreement with the hierarchical structure of the MTL described in Figure 4a (see also [15]). Since for the text and sound presentations there were no responses in the parahippocampal cortex, we carried out the same
ANOVA analysis for the other three MTL areas but none of these differences were significant. There were also no statistically significant differences in the duration of the responses across the different MTL areas neither for the picture, sound or text presentations.

Figure S5 shows the latency for the different MTL areas obtained using Gaussian windows of 100 ms, 60 ms (as reported above), 30 ms and 20 ms width. Note that the pattern of latencies described above was consistent for all the Gaussian windows used. However, given the typically sparse firing of MTL neurons, for the smaller windows the width was too small to get enough counts and therefore the instantaneous firing rate curves for each picture were noisier and latency differences were not significant, as with the larger windows.

![Figure S5: Latency of responses for the different MTL areas calculated using Gaussian windows of different widths.](image-url)
More examples of multimodal responses

Altogether, we found 5 units responding to the researchers performing experiments at UCLA. This first one is shown in Figure 3. Another neuron in the hippocampus that fired to a UCLA researcher is shown in Figure S7. A third neuron in the hippocampus (Figure S8) was selectively activated by presentations of different pictures of four researchers carrying out recordings at UCLA, as well as their written and spoken names. Two more units, each responding to two other experimenters, are shown in Figures S9 and S10.

A sparsely firing neuron in the amygdala was selectively activated by presentations of the Argentinean soccer player Diego Maradona (Figure S11) - even responding to a picture where his face is not seen but that the patient recognized as Maradona when asked after the experimental session; a neuron in the entorhinal cortex (Figure S12) responded strongly to ‘Luke Skywalker’ and to a picture of ‘Yoda’, both characters in the film ‘Star Wars’; an amygdala neuron (Figure S13) responded to snakes and spiders and a neuron in the entorhinal cortex (Figure S14) responded to presentations of four landmark buildings: the White House, the World Trade Center, the Sydney Opera House and the Bahai Temple in India. As for the neurons firing to the UCLA experimenters, the persons or objects to which these last 3 neurons fired are clearly related. Additional selective responses to pictures, sound and text presentations are shown in Figures S15 to S18.

References

Figure S6: A single unit in the amygdala that fired to pictures of the actor Michael Douglas. This unit was recorded from the same microwire as the unit shown in Figure 3. Conventions for this and the following figures are the same as for Figure 1 of the main text.
Figure S7: A single unit in the hippocampus selectively activated by presentations of one of the experimenters at UCLA. In this case two different computer synthesized voices—one male and one female—were tried.
Figure S8: A single unit in the hippocampus firing to pictures, sound and text presentations of four experimenters engaged in the clinical recordings at UCLA. This unit also responded to a picture of the actor Al Pacino and to a lesser degree to another UCLA experimenter (stimulus 49-52, 95). There were no significant responses to any of the 60 stimuli not shown in the figure. The researchers were unknown to the patient a day or two before the recording took place, which demonstrates that MTL neurons can form and represent recently formed associations.
Figure S9: A multi-unit in the entorhinal cortex that responded to two of the UCLA experimenters familiar to the patient. Stimulus 5 corresponds to another UCLA researcher for whom only 1 picture was shown.
Figure S10: A multi-unit in the amygdala responding to pictures of two of the experimentalists working with the patient.
Figure S11: A single unit in the amygdala that was nearly silent during baseline and responded with up to 50 Hz to presentations of the soccer player Diego Maradona. Note that stimulus 15 is an iconic picture of Maradona where his face is not seen.
Figure S12: A single unit in the entorhinal cortex responding selectively to pictures of Luke Skywalker, as well as to his written and spoken name. For the sound responses, in this case two different computer synthesized voices - one male and one female - were tried. Both voices elicited significant responses. Note that this unit also had a significant response to Yoda – only one single picture of Yoda was presented-, another well know character of the movie ‘Star Wars’.
Figure S13: A single unit in the amygdala firing to picture, sound and text presentations of spiders and snakes. This unit did not respond to any of the other 85 stimuli shown to the patient.
Figure S14: A single unit in the entorhinal cortex that fired to pictures, sound and text presentations of four different landmarks: the White House, the World Trade Center, the Sydney Opera House and the Bahai Temple in India.
Figure S15: A sparsely firing single unit in the hippocampus that responded selectively to presentations of the actress Halle Berry.
Figure S16: A single unit in the hippocampus that fired both to presentations of the actor Jack Nicholson and the cartoon character ‘Mr. Incredible’.
Figure S17: A single unit in the entorhinal cortex selectively activated by pictures, sound and text presentations of Robert Plant, singer of the band ‘Led Zeppelin’, which was known to the patient.
Figure S18: A single unit in the amygdala selectively activated by pictures of the actress Jennifer Lopez. Note that like 75% of responsive neurons in the amygdala, this neuron did not fire to either the text or the sound presentations.