Searching memory in time and space

Dougherty, M.R., Chang, W., Rudoler, J.H., Katerman, B.S., Halpern, D., Bruska, J.P., Diamond, N., Kahana, M.J.
University of Pennsylvania

Abstract

We investigated memory encoding and retrieval during a hybrid spatial-episodic memory task in which subjects delivered items to landmarks in a virtual environment and later recalled the delivered items. Conditional probabilities of recall transitions revealed spatial and temporal organization of studied items. We asked whether neural signatures of successful encoding and retrieval uncovered in traditional word list paradigms also appear when learning occurs within a spatiotemporal context under conditions that interpose a navigation task between each item encoding event. We found that increased theta (+T) and decreased alpha/beta (-A) accompanied successful encoding events, with the addition of increased gamma (+G) for retrieval events. Logistic-regression classifiers trained on these spectral features reliably predicted encoding and retrieval success in "hold out" sessions. Both univariate and multivariate analyses of EEG data revealed a similar spectral $T^+A^-G^+$ of successful encoding and retrieval.

Introduction

We experience events as occurring within a particular place and at a particular time, and episodic memory refers to our ability to remember such events. After remembering an event, we tend to remember events that occurred in nearby places (Miller, Neufang, et al. (2013)) and at nearby times (Kahana (1996)), indicating that the spatiotemporal context of our experiences provides an organizational framework that supports later retrieval. A vast body of work analyzing electrophysiological data in both animal and human studies has shown clear neural correlates of successful episodic memory encoding and retrieval (Burke, Ramayya, & Kahana, 2015; Johnson & Knight, 2015; Herweg, Solomon, & Kahana, 2020; Paller & Wagner, 2002), but this work has relied almost exclusively on experiments involving lists of discrete items. Such experiments lack important features of everyday life,

The authors gratefully acknowledge support fromthe U.S. Army Medical Research and Development Command (USAMRDC) through the Medical Technology Enterprise Consortium (MTEC) project 20-06-MOM. Correspondence concerning this should be addressed to Michael J. Kahana (kahana@psych.upenn.edu).

where experience flows continuously and events occur through our interaction with our environment.

The present paper seeks to address this gap in our knowledge by investigating the electroencephalographic (EEG) correlates of encoding and retrieval in a more naturalistic memory task in which subjects play the role of a courier, navigating through a virtual town and delivering items to target locations (businesses). On each delivery day of this task, subjects complete a route composed of 15 deliveries at unique locations. Then, after navigation to a 16th business, a blank screen prompts subjects to recall all of the delivered items from this delivery day. Before describing our key hypotheses, we first summarize some key EEG correlates of successful encoding and retrieval as identified in list-memory experiments.

Numerous prior EEG and intracranial EEG studies have documented a distinct pattern of spectral changes that accompany both successful encoding and subsequent episodic recollection. In an early intracranial EEG study, Sederberg, Kahana, Howard, Donner, and Madsen (2003) documented what we refer to as the $T^+A^-G^+$ of successful memory. Analyzing individual electrodes, they found that reliable increases in 4 Hz theta activity (T^+), decreases in $10-20$ Hz alpha and to some degree beta band activity (A^-), and increased >40 Hz Gamma activity (G^+) accompanied the encoding of subsequently recalled items. Subsequent studies have replicated the alpha decreases and gamma increases in large patient samples (Burke et al., 2014), but the theta increases have been somewhat elusive (Herweg, Solomon, & Kahana, 2020) except for studies that specifically separate narrowband and broadband components (e.g., Rudoler, Herweg, & Kahana, in press). Studies of episodic recollection have observed similar spectral patterns of retrieval success. For example, Burke, Merkow, Jacobs, Kahana, and Zaghloul (2015) report theta increases in frontal regions, and widespread alpha decreases and gamma increases associated with spontaneous recall of previously studied items. Similar spectral signatures of successful memory appear in a range of tasks (Greenberg, Burke, Haque, Kahana, & Zaghloul, 2015; van Vugt, Schulze-Bonhage, Litt, Brandt, & Kahana, 2010).

Scalp EEG studies of word list experiments have uncovered similar spectral signatures of successful encoding and recall. Sederberg et al. (2006) identified gamma power increases and low frequency power decreases as markers of successful encoding. Notably, they were able to recognize a dissociation between these markers as a function of study timing, with gamma power increases associated with early list item encoding and low frequency power decreases associated with late list item encoding. Long, Burke, and Kahana (2014) built upon this research in another word list free recall paradigm, demonstrating a more detailed $T^+A^-G^+$ pattern associated with successful item encoding. Recently, these analyses have been extended to retrieval in scalp EEG paradigms, indicating similar $T^+A^-G^+$ patterns to those seen at encoding and in intracranial EEG (Katerman, Li, Pazdera, Keane, and Kahana (2022)).

Here we ask whether similar $T^+A^-G^+$ patterns appear during successful memory encoding and retrieval in our Courier task, where learning occurs dynamically as subjects navigate through a rich synthetic environment. Ideally, patterns of brain activity mediating successful learning and recall in a word list task would generalize to more naturalistic settings. However, several distinctive aspects of word list memory tasks may caution this conclusion. First, strategic processes including rehearsal (repeating items in ones mind)

and the creation of verbal or imagery-based mediators to link items can lead to successful subsequent recall. These strategic processes would surely have physiological correlates, and would in turn appear as biomarkers of successful learning. In our task, however, following each learning event, subjects faced the challenge of successfully navigating through a visually complex environment to find the next store. This inter-item task relied upon a combination of subjects' memory for store location and their ability to efficiently search the environment for each store. Under these conditions, one might expect quite a difference in effective learning strategies, potentially leading to an absence of some or all of the spectral $T^+A^-G^+$ of encoding. Further, the engagement of spatial coding in our task might also lead to specific strategies, such as remembering an item in relation to its spatial location within the environment. Such spatial encoding processes could also lead to distinct neural sequelae. Finally, the literature on the enactment effect (Nyberg, 1993) indicates superior memory for enacted events compared to memory for items. Encoding of delivery items is intrinsically tied to the action of delivering them to a given business in this task, as compared to static memory of word presentations in a word list task.

The present study sought to answer three questions. First, to what degree subjects in the Courier task would organize their recalls on the basis of the temporal and spatial relations among the studied items. Previous findings in related tasks suggest that we would expect to see strong temporal clustering and at least a modest degree of spatial clustering (Miller, Weidemann, and Kahana (2012); Herweg, Sharan, et al. (2020)). Observing similar effects in our task would validate the hypothesis that our paradigm encouraged subjects to encode items in relation to their spatial context. This leads us to ask whether $T^+A^-G^+$ patterns previously identified as indicators of successful memory both in non-spatial tasks and under conditions of discrete item presentations generalize to our more naturalistic study. We expect subtle yet pronounced differences in $T^+A^-G^+$ patterns at encoding given two factors: prior literature correlating high and low frequency activity with encoding list dynamics, and the introduction of navigation between encoding events in our Courier task. Third, we asked whether machine-learning approaches that have successfully decoded variable performance in tasks with discrete, sequentially encoded items could also decode performance in a task in which subjects experience items as they navigate ad lib through a complex synthetic environment.

Methods

Subjects

30 young adults (ages 18–27, 13 Female) recruited among the students and staff at the University of Pennsylvania and neighboring institutions each contributed between three and eight sessions of navigation and recall data. Data from this experiment may be downloaded from our public repository `http://memory.psych.upenn.edu/data`.

Courier hybrid spatial-episodic memory task

Virtual Town. The virtual town consisted of 17 businesses and 15 non-identified buildings (see Figure 1C for an overhead map of the town). The town also contained a park and smaller props such as trees, benches, trash cans, and mailboxes. All businesses were visually distinct, with various unique features (including large banners displaying the

business' names) identifiable from a distance (see Figure 1A for an example street view). Town layout remained consistent across subjects, however, target location banners varied randomly across subjects (e.g. the location of the Craft Store for one subject could be the location of the Bakery for another). For a given subject, the locations of businesses remained consistent across all sessions. The environment was created and displayed using an in-house C# library that interacted with Unity, a world renowned game engine. We created the 3-D models used in the virtual environment using the Town Constructor 3 Asset Pack on the Unity Asset Store.

Game Play. Subjects played the role of a courier in a hybrid spatial-episodic memory task, riding a bicycle and delivering parcels to businesses located within a virtual town. Each experimental session consisted of a series of delivery days (trials) during which subjects navigated to deliver items and subsequently recalled these items (See Figure 1A). Before starting the first delivery day of the first session, subjects participated in a "town learning" task. During this time, subjects navigated between each of the 17 target locations according to on-screen instructions generated in pseudo-random order without delivering items. Subjects navigated using the joystick and buttons on a video game controller. In the town learning phase and in each delivery day, an arrow pointing in the direction of the target store appeared on the street after 12 seconds of navigating to ensure the experiment ran within a reasonable window of time.

Each delivery day consisted of a navigation phase followed by two recall tasks (see Figure 1A). For the navigation phase of each delivery day, 16 of the 17 total businesses were chosen as targets in a random sequence to minimize any relationship between spatial and temporal proximity. Subjects navigated to each location sequentially according to instructions presented on screen (see Figure 1A). Upon arrival at the first 15 businesses, subjects waited for 1000ms $\pm$ a 250 ms jitter, after which a word appeared on screen for 1600ms seconds while it was dictated aloud. These words, aka delivered items, never repeated within a session and were semantically related to their target store to aid recall performance. Immediately following word presentation, the next on-screen navigation instruction appeared (e.g. "please find the pet store"). At the 16th location, rather than word presentation, subjects saw a black screen and heard a beep, indicating the beginning of the free recall phase. During free recall, subjects had 90 seconds to recall as many items from this delivery day as they could in any order. Following completion of the free recall task, subjects completed a cued recall task. One business at a time appeared on screen for six seconds, during which subjects attempted to recall the item they delivered at this location. Vocal responses were recorded and annotated offline. Each session consisted of ten delivery days, with a two minute cumulative free recall phase following each block of five delivery days (see Figure 1B). After the first block of five delivery days and the first cumulative free recall, subjects had an approximately ten minute-long break. The first session was a special case, in which the first block of trials prior to the break consisted of the town learning phase, one practice delivery day, two test delivery days, and cumulative free recall.

Data collection and pre-processing

We recorded EEG with a 128-channel BioSemi Active Two system with a 2048 Hz sampling rate. In order to clean and pre-process the data, we first applied a 0.1 Hz high-pass filter to remove the baseline drift of the EEG signals over the course of each session. We

then partitioned the recording into two by splitting at the end of the mid-session break. For each section of the data, we identified bad channels as those with extremely high or low ($|z| > 3$) log-variance, or those with an extremely high ($z > 3$) Hurst exponent relative to other channels. We identified bad channels separately for each section, as we reapplied gel to address impedance issues during mid-session breaks.[1] We then dropped bad channels from their respective partitions and applied a common average reference scheme. We then reconstructed the original channels from the cleaned components, interpolated bad channels using spherical splines, and applied two fourth-order Butterworth notch filters, one with a 58-62 Hz stop-band and one with a 118-122 Hz stop-band, to remove electrical line noise.

Behavioral Analyses

As previously mentioned, in addition to immediate free recall, subjects performed cued and cumulative free recall tasks multiple times within a session. Although these auxiliary measures of memory can inform our understanding of encoding and retrieval processes, they do not provide a direct assay of initial learning, as immediate free recall precedes cued recall within each delivery day and precedes cumulative free recall across delivery days. Performance on immediate free recall likely exerts a powerful influence on subsequent recall tests (Kuhl, Johnson, & Chun, 2013), and as such, we limit the present analyses to immediate free recall data from each delivery day. Analyses of cued and cumulative free recall phases can be found in *Supplementary Figures*.

Clustering Analyses. We replicated the same behavioral measures used in Herweg, Sharan, et al. (2020) to assess recall transitions. We calculated the distance between two retrieval events for a given transition (i.e., lag in presentation order between two items for temporal, and euclidean distance between delivery locations of two items for spatial), then noted the actual and possible distances of transitions from a given recall to calculate conditional response probabilities (Kahana, 1996). We also calculated percentile rankings of each possible transition ordered according to their distance from the just recalled item, where 0% corresponded to the shortest distance and 100% to the farthest. The temporal and spatial clustering scores for a given transition were then assigned by subtracting the actual transition's assigned percentile from one, resulting in clustering scores ranging from zero to one. Each subject's overall clustering scores were calculated by averaging the scores of all recalls across all sessions. High temporal/spatial clustering scores indicate recall of items close in time/space.

We performed a permutation test to assess the significance of the clustering scores. For each trial, we shuffled the order of recalled items and recomputed clustering scores according to the above method. We then averaged scores across subjects and repeated the process 1000 times to attain a null distribution of clustering scores. By calculating the proportion of the null distribution with higher clustering scores than the observed average, we were able to determine whether average clustering scores across all subjects were above chance. To test the correlation between the two scores, we conducted a permutation test by calculating the Pearson correlation between the two clustering scores for every permutation.

[1] Recordings during break periods were frequently noisy as a result of these adjustments and participant movement, and were therefore excluded when calculating variances and Hurst exponents.

Excess Path Length Analyses. To find the shortest navigation path between businesses, we used the Breadth-First Search (BFS) algorithm, a well-known algorithm for calculating the shortest path between any two nodes in a graph. In order to apply the BFS algorithm, we first vectorized the town layout into a binary format (e.g., 0 for road and 1 for non-road objects), where each point in the vectorized array became a "node". For a given navigation period (i.e. navigation from business A to B), we computed the shortest route and calculated the shortest total distance to navigate on this path by summing the euclidean distances of its nodes. We then calculated the excess path length by subtracting the shortest navigation distance from the actual navigation distance.

Univariate Analyses

Spectral decomposition. We applied a Morlet Wavelet transform (five cycles) to compute spectral power at each electrode for eight log-transformed frequencies ranging from 3-128 Hz. We then log-transformed the resulting signal and z-scored the log-transformed signal at each frequency and electrode pair to estimate deviation from mean power.

Epoch construction. We categorized encoding and retrieval events into two groups for comparison. Subsequent memory of encoding events during immediate free recall determined categorization of encoding events. During encoding, words appeared on screen for 1600ms. For encoding analyses, each wavelet was calculated between 300ms and 1300ms following presentation onset, with a 833.33ms mirrored buffer applied to both sides of the data. For retrieval analyses, we identified 1000ms preceding correct recall events (i.e. first instance of each item's recall from the just-completed delivery day) during immediate free recall, excluding events for which subjects were not silent during these 1000ms.

We compared successful retrieval epochs to matched "deliberation" periods. We replicated the methodology of Solomon et al. (2019) to identify deliberation periods for each successful retrieval event. First, for each successful retrieval event, we searched for a 1000ms interval during immediate free recall in a separate delivery day of the same session during which no vocalization took place. We prioritized deliberation periods that occurred within the same 1000ms of a recall window as the retrieval event, and if no match existed, we extended the interval within 2000ms surrounding vocalization onset of the retrieval event. We excluded all successful retrieval events without matched deliberation periods from analysis. As a result, each successful retrieval event was matched with one deliberation interval from a different list. Each wavelet for successful retrieval events was calculated between 900ms and 50ms prior to vocalization onset, with a 833.33ms mirrored buffer applied to both sides of the data. In accordance with this method, we calculated each wavelet for matched deliberation events between 100ms and 950ms of the 1000ms window with a 833.33ms mirrored buffer applied to both sides of the data.

Regions of interest. We collapsed spectral power estimates across subsets of electrodes to generate average spectral patterns at eight frequencies $\times$ eight regions of interest (ROIs). We selected ROIs based on previous studies using similar EEG caps (Long & Kahana, 2017; Weidemann, Mollison, & Kahana, 2009). Figure 5A provides a diagram of the ROIs utilized in neural analyses.

Multivariate analyses

Classifier Training. We trained L2-penalized logistic regression classifiers based on item encoding events, and separately on item retrieval events during free recall. For each epoch corresponding to an encoded item, we trained 1024 electrophysiological features (channel x frequency) against binary labels reflecting memory success. For encoding, this label indicates successful or failed retrieval during subsequent free recall. For retrieval, this label distinguishes correctly recalled words from periods of silent deliberation. The training data were re-sampled to ensure balanced class weighting. To select the L2 penalty parameter, we performed a nested leave-one-session-out cross validation procedure, selecting the penalty term for the outer fold which achieved the highest average accuracy score across the inner folds.

To ensure proper model fitting, we excluded subjects who did not record at least three sessions of training data, which slightly reduced our sample size for this analysis to 29 subjects. Using less data would increase the risk of over-fitting the penalty parameter, especially in the inner folds of our nested cross validation procedure.

Evaluating classifiers. We evaluated the accuracy of classifiers by computing the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) function, which computes the proportion of true positives and false positives at varying decision thresholds. We evaluated the statistical significance of our observed AUC values for each subject via permutation testing. The expected AUC for random guessing is 0.5, but we cannot compare the observed within-subject AUC values to this theoretical null. Since AUC scores are highly correlated across CV folds, it is not possible to compute an unbiased estimate of the variance of the model's within-subject performance. Instead, we generated an empirical null distribution of AUC values by randomly shuffling the true target labels and recomputing the classifier's performance. We consider subject classifiers that achieve an AUC greater than 95% of this null distribution to have above-chance predictive accuracy. To determine if predictive accuracy across subjects is significantly greater than chance, we computed a one-sample t-test on the distribution of observed AUC values against the expected mean of 0.5. This across-subject test is valid because subjects are independent.

To estimate the contribution of each frequency $\times$ ROI pair in the input space to classification performance, we constructed forward models based on the learned weights of the significant individual classifiers (Haufe et al., 2014).

Results

Unlike typical word list recall tasks, in this task, inter-stimulus intervals varied randomly between items as a function of distance between successive delivery locations and subject's ability to efficiently navigate between locations. In theory, these task demands would serve as a source of distraction, akin to continual distractor paradigms, making it somewhat more difficult to engage in strategic rehearsal than in standard list recall paradigms. (Parkinson (1972)).

Recall dynamics in immediate free recall task

As shown in Figure 2A, subjects performed at high levels in immediate free recall, with an average recall accuracy of 71.5% ($SD = 0.12$) and slightly higher recall for primacy (serial

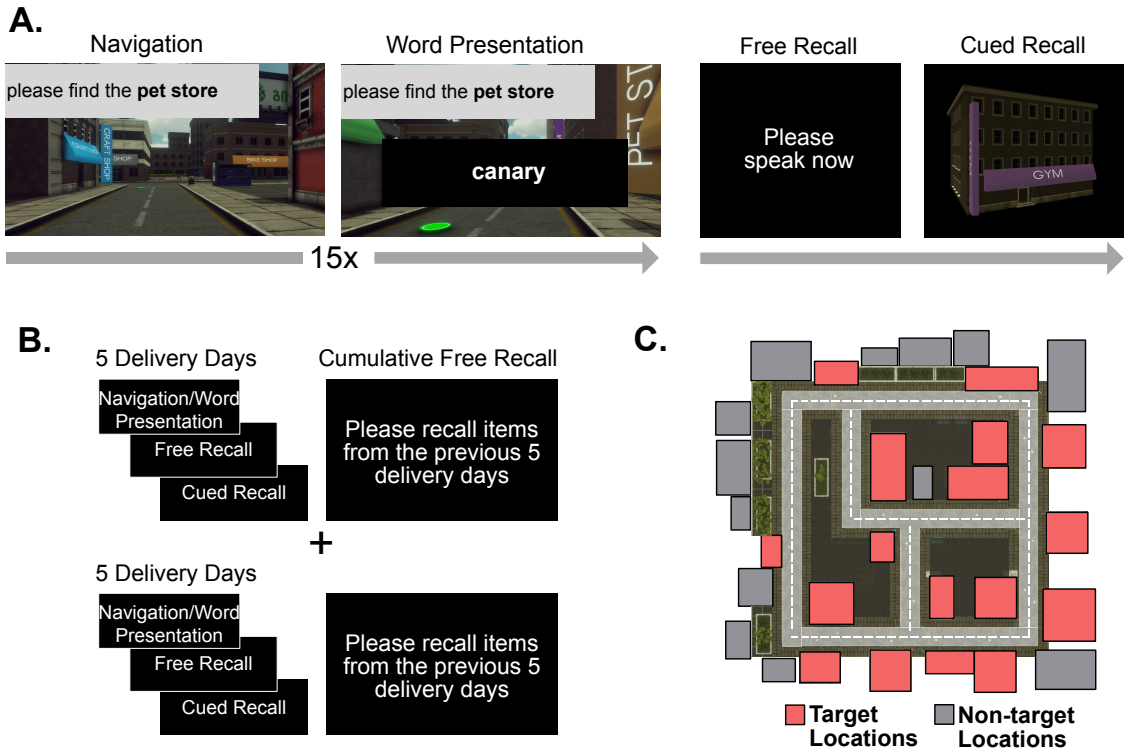

Figure 1. **Task Design**. **A.** A single trial within the Courier task. Subjects navigate to a specified business to deliver an item. Upon arrival, a word appears on screen indicating the item that was delivered. Subjects then navigate to the following business. After completing 15 deliveries, subjects complete free recall and cued recall of the delivered items. **B.** Structure of one session of the Courier task. Subjects complete 5 delivery days, followed by cumulative free recall of all items delivered across preceding delivery days. After a short break, subjects complete an additional five delivery days and cumulative free recall of items from these five delivery days. **C.** Top-down view of the virtual town. The town consists of 17 target locations, i.e. delivery locations, and 13 non-target locations.

positions 1-3 vs. 4-12, $t_{(29)}$=5.65, $p < 0.001$) and recency (serial positions 13-15 vs. 4-12, $t_{(29)}$=4.77, $p < 0.001$) items. Recall initiation showed a marked effect of serial position (Figure 2B), with subjects most often beginning recall with items from the end (serial positions 13-15 vs 4-12, $t_{(29)}$=8.17, $p < 0.001$), and beginning (serial positions 1-3 vs 4-12, $t_{(29)}$=3.52, $p < 0.001$) of a list. Validating the results of previous experiments (Miller, Lazarus, Polyn, & Kahana, 2013), subjects demonstrated strong tendencies to temporally and spatially organize memories (Figure 2C,D). We replicated temporal contiguity ($|\text{lag}| = 1$ vs $|\text{lag}| \geq 4$, $t_{(29)}$=5.66, $p < 0.001$) and forward asymmetry (lag $+1$ vs -1, $t_{(29)}$=4.55, $p < 0.001$) effects, well-known phenomena in traditional word list experiments (Healey, Long, & Kahana, 2019), in our hybrid spatial-episodic memory task. In addition, subjects demonstrated a strong tendency to cluster recalls according to their spatial proximity at encoding, validated through a repeated measures ANOVA ($F(4, 116) = 27.44, p < 0.001$; very near vs. very far $t_{(29)}$=5.07, $p < 0.001$; Figure 2D). Permutation tests of average clustering scores

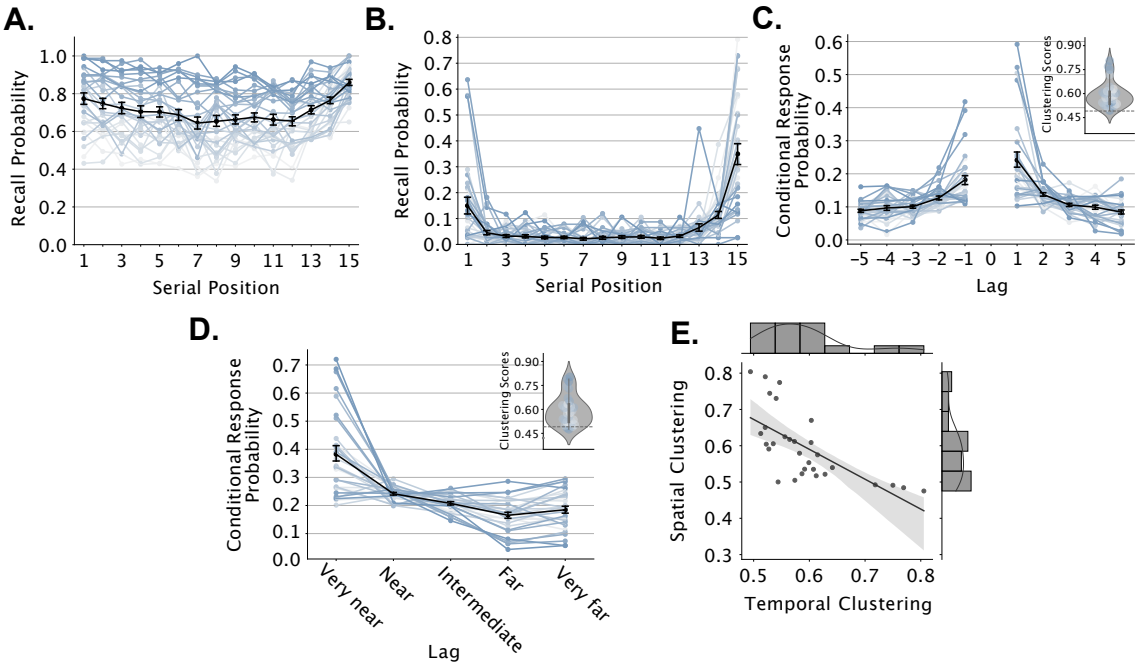

Figure 2. **Immediate Free Recall**. **A.** Probability of correctly recalling an item as a function of its serial position. **B.** Probability of initiating a recall sequence with an item at each serial position. **C.** Probability of recalling an item as a function of its lag from the just-recalled item (the difference between its predecessor's serial position and its serial position). Graph in the top right corner shows subjects' average temporal clustering scores. **D.** Probability of recalling an item a given spatial distance away from the just-recalled item. Possible distances at a given recall transition are binned into five distance groups according to quantile splits of the euclidian distance between target locations. Graph in the top right corner shows subjects' average spatial clustering scores. The dotted lines in panels C and D indicate chance levels of clustering, determined according to aforementioned permutation tests. **E.** Correlation between temporal and spatial clustering scores. Each dot represents a subject, with histograms of temporal and spatial clustering scores on the top and right axes, respectively. Bolded curves in panels A,B,D, and E show averaged data across all subjects, plotted individually as colored curves. The darker the color, the higher the subject's recall performance. Error bars in panels A through D indicate 95% confidence intervals, calculated using 1000 iterations of bootstrapped resampling of the data.

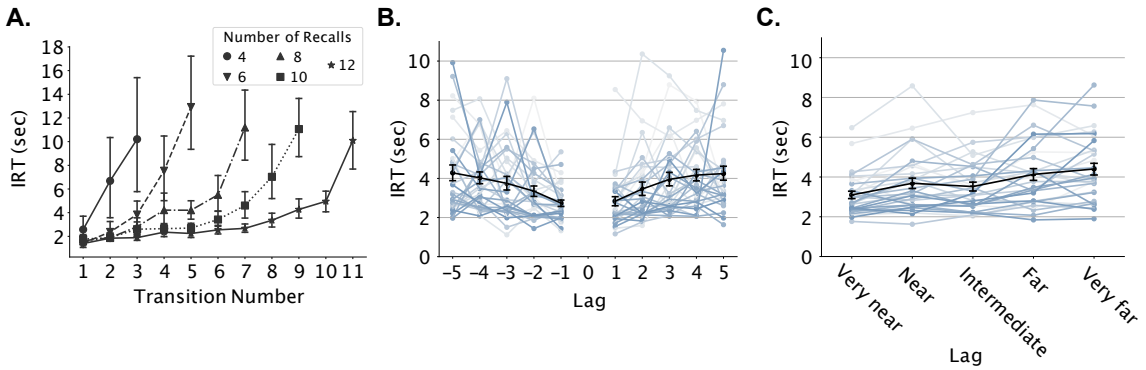

Figure 3. **Inter Response Time**. **A.** IRT curves for correct recalls by transition number, shown separately for different number of correct recalls. **B.** Conditional response latency curves as a function of temporal lag. **C.** Conditional response latency curves as a function of distance bin. Error bars in panels A through C indicate 95% confidence intervals, calculated using 1000 iterations of bootstrapped resampling of the data. Bolded curves in panels B and C show averaged data across all subjects, plotted individually as colored curves. The darker the color, the higher that subject's recall performance.

further validated such findings (temporal clustering $p < 0.001$, spatial clustering $p < 0.001$, see *Methods*). Average clustering scores across all subjects were significantly negatively correlated ($r_{(29)}$=−0.68, $p < 0.001$, see *Methods*), indicating a subject-level preference for clustering strategy that builds upon findings in previous spatiotemporal episodic memory tasks (Herweg, Sharan, et al., 2020).

We additionally analysed inter response times (IRTs) of recalls during immediate free recall. As shown in Figure 3A, IRT curves showed exponential growth patterns (i.e., subjects took longer to initiate correct recall transitions as memory was exhausted), replicating recall characteristics found in traditional word list memory experiments (Murdock & Okada, 1970). Figure 3B and C show conditional response latency curves for temporal and spatial distances, respectively. Subjects demonstrated quicker inter-response times for successive correct recalls that were encoded close in time ($|\text{lag}| = 1$ vs $|\text{lag}| \geq 4$, $t_{(29)}$=−6.47, $p < 0.001$) and space ($F(4, 116) = 10.25, p < 0.001$; very near vs. very far, $t_{(29)}$=−5.02, $p < 0.001$).

As confirmation that subjects acquired spatial knowledge throughout the experiment, we investigated subjects' learning of the spatial environment across sessions. To do so, we performed an analysis of excess path length, where short excess path lengths indicate town layout learning via increased path efficiency (see *Methods*). Figure 4B shows excess path length binned across sessions. As sessions progressed, excess path decreased, indicating increased knowledge of the spatial environment with time. We assessed the significance of the spatial learning experience with a repeated measures ANOVA, followed by a post-hoc test comparing excess path lengths at session bins one and three ($F(3, 87) = 20.07, p < 0.001$; $t_{(29)}$=4.52, $p < 0.001$). In addition, subjects with a shorter average excess path length demonstrated higher degrees of spatial clustering ($r_{(29)}$=−0.37, $p = 0.04$). These results demonstrate the importance of spatial context in the organization of episodic memories, and the validity of spatial clustering scores as indicators of this organization.

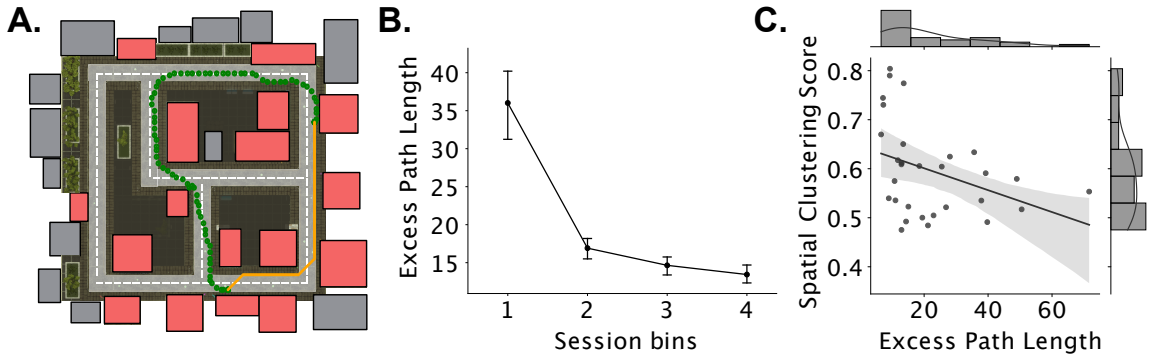

Figure 4. **Environment Learning**. **A.** Sample trace of actual navigation and shortest navigation paths. Green dots represent the actual navigation path an example subject made. The orange line represents the shortest path for that given navigation. See *Methods* for details on calculating the shortest path. **B.** Excess path length analysis. Excess path length is the difference between the shortest path distance and actual navigation path distance. Since subjects vary in their number of completed sessions, sessions are grouped into four bins within subjects. Error bars indicate 95% confidence intervals, calculated using 1000 iterations of bootstrapped resampling of the data. **C.** Correlation between excess path lengths and spatial clustering scores, with histograms of excess path lengths and spatial clustering scores on the top and right axes, respectively. Each dot represents one subject's data.

Spectral correlates of successful memory encoding and retrieval

Next, we sought to characterize EEG correlates of successful memory encoding and retrieval in our hybrid spatial-episodic memory task. Previous work in traditional word list retrieval studies has identified increased theta power (+T), decreased alpha/beta power (-A), and increased gamma power (+G) as markers of successful memory retrieval (Katerman et al. (2022)). Studies of memory encoding have revealed similar findings, with alpha/beta decreases being particularly robust (Long et al. (2014); Sederberg et al. (2006)). Here, we first transformed multivariate EEG time series into spectral power estimates between 3 and 128 Hz frequencies (see *Methods*). We extracted mean power estimates during 1000ms item encoding epochs (between 300ms and 1300ms of the 1600ms total item presentation), and during an 850ms window preceding correct retrievals (900ms to 50ms prior to vocal onset, see *Methods*).

We first consider the spectral correlates of successful memory encoding, comparing spectral estimates for the encoding of subsequently remembered and forgotten items. To do so, we first conducted an independent-samples t-test across events separately for each subject. This yielded t-scores for each combination of electrode and frequency. Following this, we averaged the resulting t-scores across electrodes within an ROI to control for residual noise from any particular electrode. Figure 5B averages these mean t-scores across subjects to determine a variance-normalized effect size for each ROI $\times$ frequency pair. Here we see striking decreases in alpha/beta power during the encoding of subsequently remembered items. We also see increases in low-theta power during successful memory encoding. A clear pattern did not emerge in the gamma band. To assess the reliability of these effects

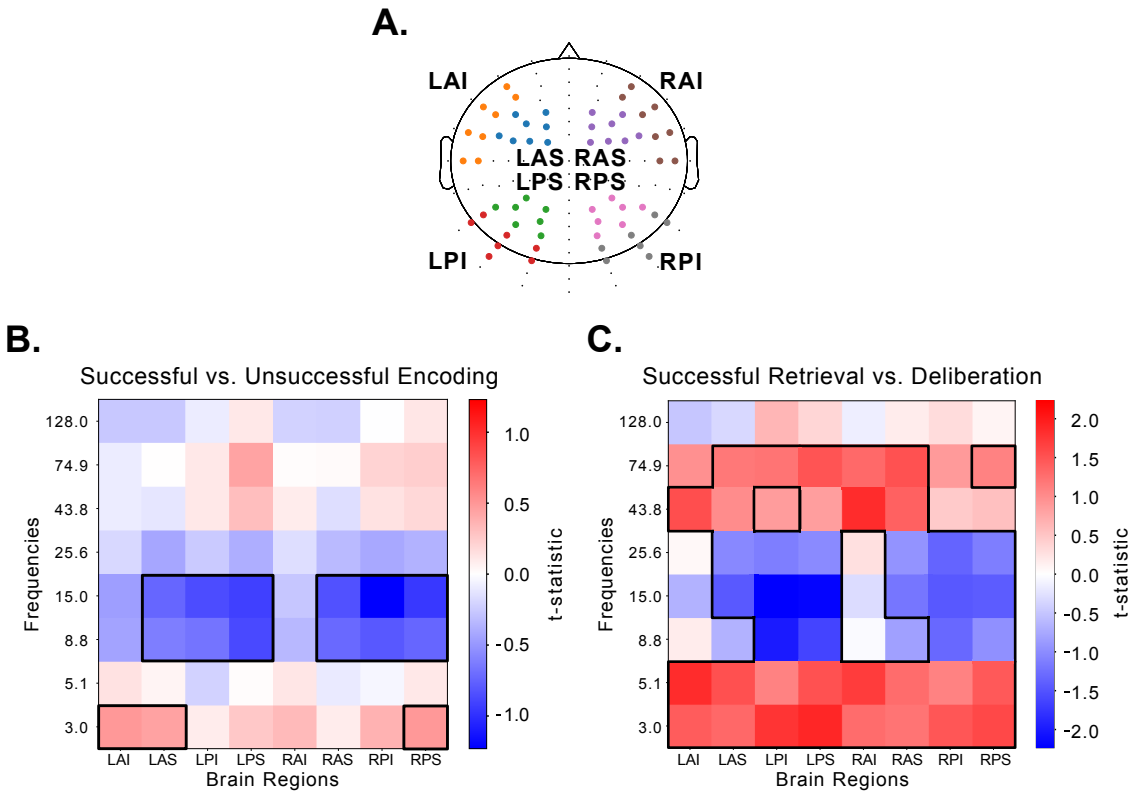

Figure 5. **Spectral markers of successful encoding and retrieval. A**. Diagram of regions of interest on the scalp. L: left, R: right, A: anterior, P: posterior, I: inferior, S: superior. **B**. Differences in spectral power during encoding for subsequently remembered versus subsequently forgotten items. **C.** Differences in spectral power during encoding for intervals 900ms to 50ms preceding correct recalls versus matched deliberation periods. Colors in each frequency × ROI pair correspond to the mean of independent t-statistics for electrodes within a region of interest, averaged across subjects. Outlined frequency × ROI pairs indicate $p < 0.05$ of FDR corrected one-sample t-statistics of the average t-statistics within each frequency × ROI pair across subjects.

across subjects, we conducted a one-sample t-test comparing the resulting average ROI × frequency t-scores to zero. Each ROI × frequency pair that exceeded a $p < 0.05$ FDR corrected criterion appears outlined in black. All ROIs except for anterior-inferior regions exhibited statistically reliable alpha/beta band decreases. Statistically significant theta increases appeared in left anterior inferior, left anterior superior, and right parietal superior regions. Figure 6A provides a visualization of subject-level power changes at encoding, averaged across all electrodes.

We compared average power estimates for successful recall versus matched deliberation events (see *Methods*) according to the same method. As shown in Figure 5C, successful retrieval events demonstrate similar trends to encoding, with alpha/beta power decreases and low-theta power increases. However, the effects were much more pronounced and widespread at retrieval, with low and high-theta increases across all ROIs as well as gamma

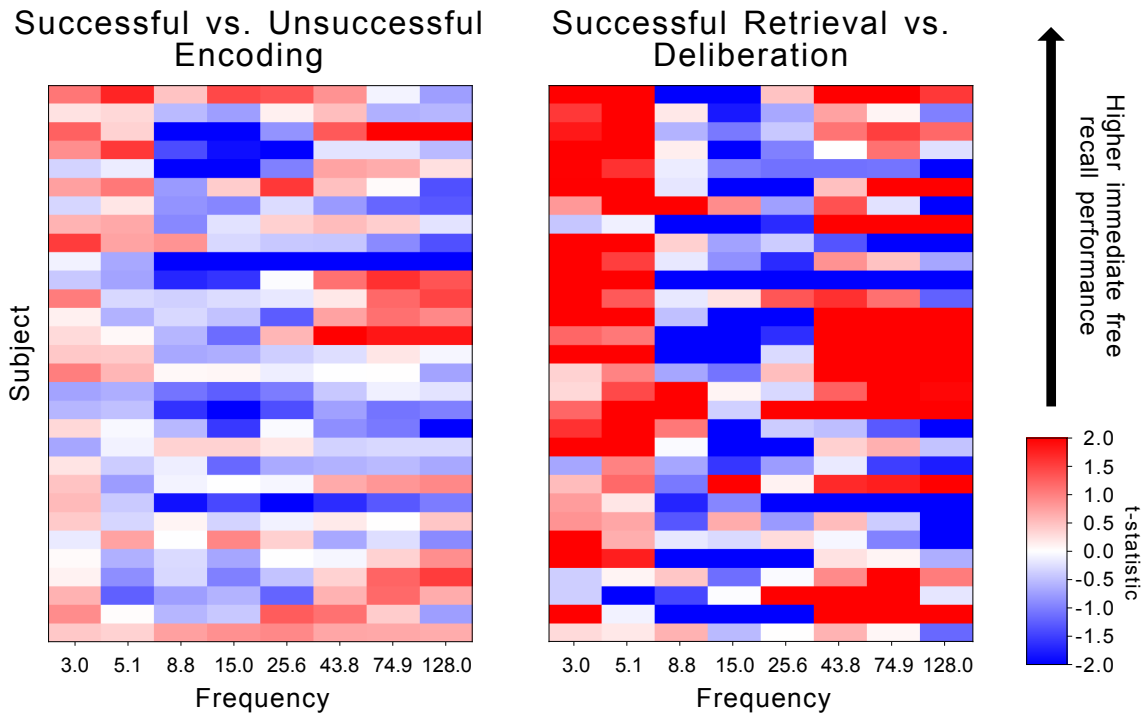

Figure 6. **Participant-specific spectral markers of successful episodic memory in encoding and retrieval.** Each row shows results from one subject, sorted by recall performance. Subject-specific independent t-statistics for the successful vs. unsuccessful memory encoding contrast (left) and successful retrieval vs. deliberation contrast (right) are collapsed across all electrodes. Power increases and decreases are shown in red and in blue, respectively.

power increases across various ROIs in the right and left hemispheres. Figure 6B provides a visualization of subject-level power changes at retrieval, averaged across all electrodes.

These neural results replicate those previously documented in word list studies, with the exception of gamma power at encoding. In scalp EEG, one would expect to see broadband gamma power increases at encoding similar to those seen at retrieval. However, Sederberg et al. (2006) have indicated the possibility that these gamma band increases at encoding are primarily driven by list position. High Frequency Activity (HFA) is particularly pronounced at the beginning of lists and trails off as lists progress, lending to the idea that HFA at encoding is more indicative of attentional orientation than a marker of successful memory. Our Courier task is unique in that subjects must navigate to a business at the beginning of the encoding period prior to presentation of the first to-be remembered item, which takes 16s on average. If HFA is in fact indicative of an initial attentional orientation, we would not expect to see HFA at encoding in our Courier task.

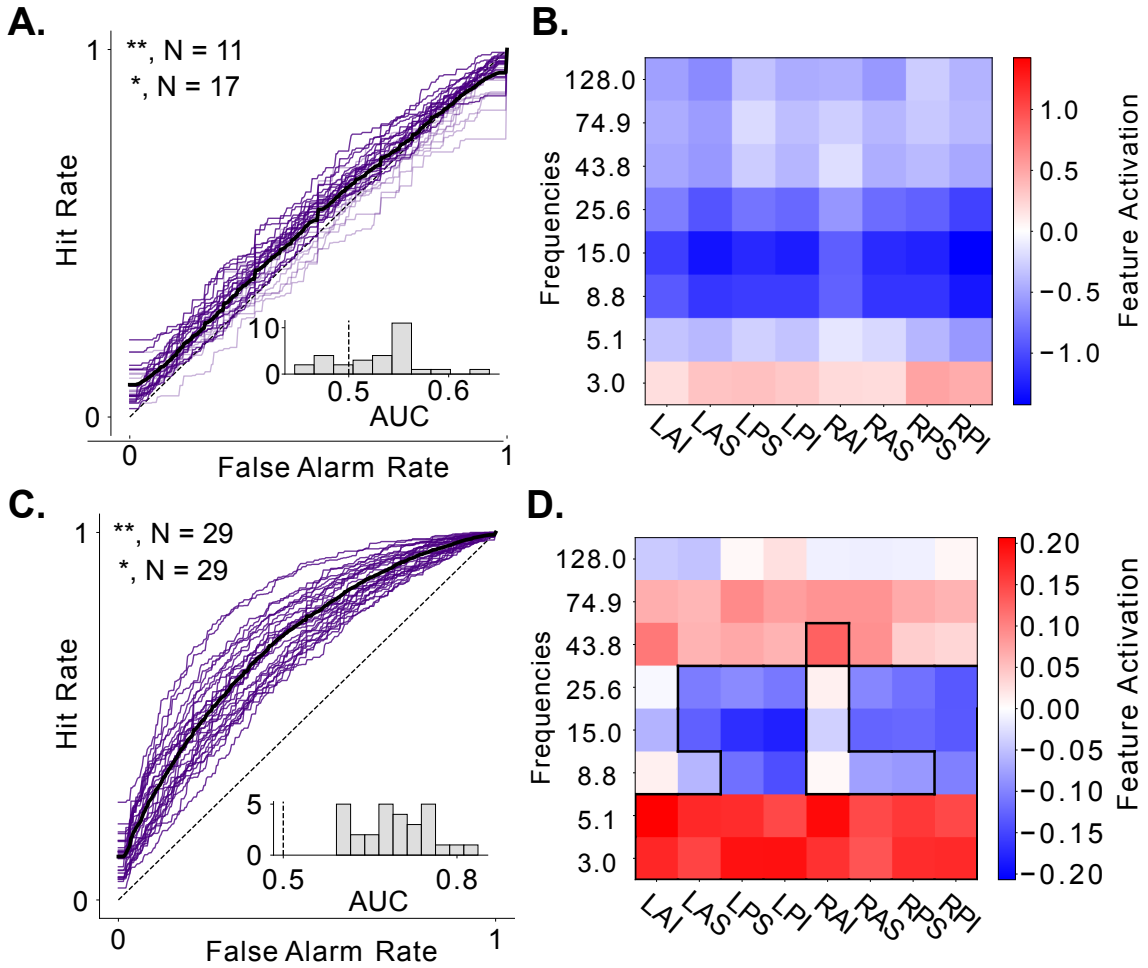

Figure 7. **Decoding mnemonic success.** L2 Logistic Regression classifiers were trained to predict memory success at both encoding and retrieval. **A**. Receiver Operating Characteristic (ROC) curves depicting classifier performance for every subject. Encoding classifiers trained on all events predicting subsequently recalled words. The number of classifiers with significant performance (by permutation test) is shown in the upper left region of the ROC plots. **, $p < 0.01$. *, $p < 0.05$. The distribution of observed AUCs is shown in the lower right region of the ROC plots. **B**. Classifier feature activations. Colors in each frequency × ROI pair correspond to the mean activation value across participants. Some alpha band features were significant in a t-test against zero, but did not survive FDR correction. **C.** Same as in panel A, but for retrieval classifiers trained on all events predicting correct recall responses. **D.** Feature activations for retrieval classifiers. Features with significant activations (t-test against zero, FDR corrected) are bordered in black.

Multivariate analysis of mnemonic success

We leveraged multivariate pattern analyses to quantify event-level memory states and test the generalizability of the spectral contrasts. Given the strong, consistent univariate biomarkers of memory success, we predicted that multivariate classifiers would be able to reliably decode memory states from EEG and would utilize features similar to the ones shown in Figure 5.

We trained participant-specific multivariate classifiers to decode successful encoding and retrieval epochs on the basis of spectral features across all electrodes. We labeled encoding epochs on the basis of subsequent correct recall and we labeled retrieval epochs based on whether they preceded a successful recall or silent deliberation (failed retrieval). We used a leave-one-session-out cross-validation scheme, evaluating classifier performance as the area under the curve (AUC) generated by the Receiver Operating Characteristic (ROC) function, over the predicted probabilities of events in the hold-out sessions.

Figure 7 A and C summarize classifier performance across all subjects. The encoding classifiers reliably predicted memory success for data from 17 of the 29 participants (all p's < 0.05, permutation tests). At the group level (mean AUC $= 0.533$, $SE = 0.007$), the distribution of the observed AUCs was significantly greater than the theoretical null of 0.5 ($t(29) = 4.53$, $p < 0.001$). The retrieval classifiers reliably discriminated between correct recalls and failed retrieval for all 29 subjects ($p < 0.05$, permutation test) and at the group level (mean AUC $= 0.698$, $SE = 0.010$, $t(29) = 18.99$, $p < 0.001$).

Panels B and D of Figure 7 show the feature activation maps derived from the subject-specific forward-models, with the activation patterns aggregated over electrodes in different ROIs and across subjects. In predicting whether a word will likely be subsequently recalled, the encoding classifiers consistently relied on alpha band decreases similar to those shown in Figure 5D, but this effect was not significant after FDR correction. The feature weights for the retrieval classifiers similarly echoed findings from our univariate contrasts for successful and unsuccessful retrieval, showing that similar spectral markers drive the predictions of these models.

Conclusions

Recent decades have witnessed enormous strides in our understanding of the neural basis of learning and memory. Using electrophysiological recordings and functional neuroimaging, cognitive neuroscientists have uncovered brain signals that predict whether an experienced item will later be remembered, and whether a participant is about to remember a previously experienced event. To precisely control the conditions of encoding and retrieval, researchers have predominantly studied brain signals while participants learn and recall discrete lists of memoranda such as words or pictures – conditions that differ dramatically from recalling experienced events in our daily lives.

The present study sought to examine the EEG correlates of memory encoding and retrieval in a task that embodies far greater complexity, and perhaps greater ecological validity, as compared with word list learning. In our Courier task, participants experienced items as they navigated through a complex virtual environment. Playing the role of a courier, they faced multiple "delivery days" during which they navigated to a sequence of locations (businesses), and upon arrival at each business, they learned the identity of the

item that they delivered. Participants later recalled the delivered items, either freely or in response to specific retrieval cues. The spatial nature of our task allowed participants to associate their experiences with a rich spatiotemporal context. By having participants experience objects only after navigating to the target store, our task created a contextual separation between learning events, more closely mimicking the way people learn in their daily lives.

Participants exhibited striking temporal and spatial clustering, recalling items in association with those experienced at nearby times and in nearby locations (see Figure 2). Temporal and spatial clustering manifested in IRTs as well, with shorter IRTs for successive recalls presented close in time and space (see Figure 3). These results lend credence to the use of naturalistic tasks to capture additional dimensions of memory organization that evade the simplicity of traditional word list experiments. As the town layout did not change across sessions, participants learned to navigate more efficiently as they gained experience playing the game. Their improved spatial knowledge led to an increased tendency to exhibit spatial clustering of recalled objects (see Figure 4).

Spectral analysis of EEG signals revealed pronounced alpha/beta band decreases and theta band increases during successful encoding and retrieval, in addition to gamma band increases at retrieval. These results, along with those derived from multivariate classifier-based analyses of our data, instantiate key elements of the spectral $T^+A^-G^+$ pattern previously documented in word list experiments. Whereas retrieval data illustrated all three components of $T^+A^-G^+$ (Figure 5C), encoding data did not yield a significant gamma increase for subsequently recalled items (Figure 5B). Previous work, however, has linked the gamma increase during successful encoding to the pronounced primacy effect seen in most word list experiments. As a result, previous authors have hypothesized that encoding gamma increases may be indicative of an attentional boost given to early list items (Sederberg et al., 2006; Serruya, Sederberg, & Kahana, 2014). The missing (or muted) gamma effect in this task aligns with this hypothesis, as one would expect the requirement to find the next store prior to delivering an item to reduce primacy (Howard & Kahana, 1999; Kahana, 2012). Taken together, our spectral analyses suggest a conservation of memory-related neural processes between artificial list paradigms and more naturalistic experiences.

Whereas prior research has largely studied memory for time and space in isolation, we sought to create a controlled memory paradigm to observe the joint effects of spatial and temporal organization during memory search. Our Courier task achieved this purpose, demonstrating that the order and timing of participants' recalls conforms to both the temporal and spatial organization of experienced events. Further, this task allowed us to ask whether neural features long-studied in artificial list learning tasks would generalize to a more naturalistic lab-based memory paradigm. Embracing study methods and materials beyond traditional paradigms creates opportunities to capture and understand mechanisms of memory that eluded prior research. Observing a spectral $T^+A^-G^+$ of spontaneous recall, and the theta and alpha components of encoding $T^+A^-G^+$, buttresses claims that electrophysiological studies of laboratory tasks provide a window into mechanisms of learning and recall used in our daily lives. Multivariate classifier predictions of successful memory states further support these claims. Overall, this study demonstrates the importance of spatial context in episodic memory organization, confirms the validity of naturalistic tasks in capturing neural markers of successful memory, and demonstrates predictable neural patterns

of memory in scalp EEG detectable by multivariate classifiers.

References

Burke, J. F., Long, N. M., Zaghloul, K. A., Sharan, A. D., Sperling, M. R., & Kahana, M. J. (2014). Human intracranial high-frequency activity maps episodic memory formation in space and time. *NeuroImage*, *85*, 834–843. doi: 10.1016/j.neuroimage.2013.06.067

Burke, J. F., Merkow, M., Jacobs, J., Kahana, M. J., & Zaghloul, K. (2015). Brain computer interface to enhance episodic memory in human participants. *Frontiers in Human Neuroscience*, *8*, 1055.

Burke, J. F., Ramayya, A. G., & Kahana, M. J. (2015). Human intracranial high-frequency activity during memory processing: Neural oscillations or stochastic volatility? *Current Opinion in Neurobiology*, *31*, 104-110. doi: https://doi.org/10.1016/j.conb.2014.09.003

Greenberg, J. A., Burke, J. F., Haque, R., Kahana, M. J., & Zaghloul, K. A. (2015). Decreases in theta and increases in high frequency activity underlie associative memory encoding. *NeuroImage*, *114*, 257–263. doi: 10.1016/j.neuroimage.2015.03.077

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. doi: 10.1016/j.neuroimage.2013.10.067

Healey, M. K., Long, N. M., & Kahana, M. J. (2019). Contiguity in episodic memory. *Psychonomic Bulletin & Review*, *26*(3), 699–720. doi: 10.3758/s13423-018-1537-3

Herweg, N. A., Sharan, A. D., Sperling, M. R., Brandt, A., Schulze-Bonhage, A., & Kahana, M. J. (2020). Reactivated spatial context guides episodic recall. *The Journal of Neuroscience*, *40*(10), 2119–2128. doi: 10.1523/jneurosci.1640-19.2019

Herweg, N. A., Solomon, E. A., & Kahana, M. J. (2020). Theta oscillations in human memory. *Trends in Cognitive Science*, *24*(3), 208-227. doi: https://doi.org/10.1016/j.tics.2019.12.006

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 923–941. doi: 10.1037/0278-7393.25.4.923

Johnson, E. L., & Knight, R. T. (2015). Intracranial recordings and human memory. *Current opinion in Neurobiology*, *31*, 18–25.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, *24*(1), 103–109. doi: 10.3758/BF03197276

Kahana, M. J. (2012). *Foundations of human memory.* New York, NY: Oxford University Press.

Katerman, B. S., Li, Y., Pazdera, J. K., Keane, C., & Kahana, M. J. (2022). EEG biomarkers of free recall. *NeuroImage*, *246*, 118748. doi: 10.1016/j.neuroimage.2021.118748

Kuhl, B. A., Johnson, M. K., & Chun, M. M. (2013). Dissociable neural mechanisms for goal-directed versus incidental memory reactivation. *The Journal of Neuroscience*, *33*(41), 16099–16109.

Long, N. M., Burke, J. F., & Kahana, M. J. (2014). Subsequent memory effect in intracranial and scalp EEG. *NeuroImage*, *84*, 488–494. doi: 10.1016/j.neuroimage.2013.08.052

Long, N. M., & Kahana, M. J. (2017). Modulation of task demands suggests that semantic processing interferes with the formation of episodic associations. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *43*(2), 167–176. doi: 10.1037/xlm0000300

Miller, J. F., Lazarus, E., Polyn, S. M., & Kahana, M. J. (2013). Spatial clustering during memory search. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*(3), 773–781.

Miller, J. F., Neufang, M., Solway, A., Brandt, A., Trippel, M., Mader, I., ... Schulze-Bonhage, A. (2013). Neural activity in human hippocampal formation reveals the spatial context of retrieved memories. *Science*, *342*(6162), 1111-1114. doi: 10.1126/science.1244056

Miller, J. F., Weidemann, C. T., & Kahana, M. J. (2012). Recall termination in free recall. *Memory & Cognition*, *40*, 540–550.

Murdock, B. B., & Okada, R. (1970). Interresponse times in single- trial free recall. *Journal of Verbal Learning and Verbal Behavior*, *86*, 263-267.

Nyberg, L. (1993). *The enactment effect: Studies of a memory phenomenon* (Doctoral Dissertaion). University of Umea.

Paller, K. A., & Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences*, *6*(2), 93-102. doi: 10.1016/S1364-6613(00)01845-3

Parkinson, S. R. (1972). Short-term memory while shadowing: Multiple-item recall of visually and of aurally presented letters. *Journal of Experimental Psychology*, *92*(2), 256-265. doi: https://doi.org/10.1037/h0032077

Rudoler, J. H., Herweg, N. A., & Kahana, M. J. (in press). Hippocampal theta and episodic memory. *Journal of Neuroscience*. doi: https://doi.org/10.1101/2022.03.13.484014

Sederberg, P. B., Gauthier, L. V., Terushkin, V., Miller, J. F., Barnathan, J. A., & Kahana, M. J. (2006). Oscillatory correlates of the primacy effect in episodic memory. *NeuroImage*, *32*(3), 1422–1431. doi: 10.1016/j.neuroimage.2006.04.223

Sederberg, P. B., Kahana, M. J., Howard, M. W., Donner, E. J., & Madsen, J. R. (2003). Theta and gamma oscillations during encoding predict subsequent recall. *Journal of Neuroscience*, *23*(34), 10809–10814. doi: 10.1523/JNEUROSCI.23-34-10809.2003

Serruya, M. D., Sederberg, P. B., & Kahana, M. J. (2014). Power shifts track serial position and modulate encoding in human episodic memory. *Cerebral Cortex*, *24*, 403–413. doi: 10.1093/cercor/bhs318

Solomon, E. A., Stein, J. M., Das, S., Gorniak, R., Sperling, M. R., Worrell, G., ... Kahana, M. J. (2019). Dynamic theta networks within the human medial temporal lobe support episodic encoding and retrieval. *Current Biology*, *29*(7), 1100-1111. doi: 10.1016/j.cub.2019.02.020

van Vugt, M. K., Schulze-Bonhage, A., Litt, B., Brandt, A., & Kahana, M. J. (2010). Hippocampal gamma oscillations increase with working memory load. *Journal of Neuroscience*, *30*(7), 2694–2699.

Weidemann, C. T., Mollison, M. V., & Kahana, M. J. (2009). Electrophysiological correlates of high-level perception during spatial navigation. *Psychonomic Bulletin & Review*, *16*(2), 313–319. doi: 10.3758/PBR.16.2.313