Neural Correlates of Auditory Short-Term Memory in Rostral Superior Temporal Cortex

Brian H. Scott,1,2* Mortimer Mishkin,1 and Pingbo Yin1,2
1Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA
2Neural Systems Laboratory, Institute for Systems Research, University of Maryland, College Park, MD 20742, USA

Summary

Background: Auditory short-term memory (STM) in the monkey is less robust than visual STM and may depend on a retained sensory trace, which is likely to reside in the higher-order cortical areas of the auditory ventral stream. Results: We recorded from the rostral superior temporal cortex as monkeys performed serial auditory delayed match-to-sample (DMS). A subset of neurons exhibited modulations of their firing rate during the delay between sounds, during the sensory response, or during both. This distributed population carried a predominantly sensory signal modulated by the mnemonic context of the stimulus. Excitatory and suppressive effects on match responses were dissociable in their timing and in their resistance to sounds intervening between the sample and match. Conclusions: Like the monkeys' behavioral performance, these neuronal effects differ from those reported in the same species during visual DMS, suggesting different neural mechanisms for retaining dynamic sounds and static images in STM.

Introduction

Auditory perception and language depend on linking sounds through time [1, 2]. In vision and touch, short-term memory (STM) is thought to rely on the same regions of secondary sensory and association cortex that support perception [3], such as the inferotemporal (IT) visual cortex [4]. The rostral superior temporal cortex (rSTC), including the rostral supratemporal plane and superior temporal gyrus, occupies a position in the auditory processing hierarchy similar to that of IT in the visual processing hierarchy [5, 6] and may play an analogous functional role. Neurons in rSTC show long response latency and a preference for complex stimuli [7, 8]; ablation of rSTC disrupts auditory pattern discrimination and delayed match-to-sample (DMS) performance [9, 10], and rSTC affords a bridge to the prefrontal cortex (PFC) [11], known to function in concert with IT during visual DMS [12] and implicated in auditory DMS as well [13–16].

Despite these commonalities between the visual and auditory systems, recent behavioral studies indicate that auditory DMS performance in the monkey is less robust than visual DMS performance and is likely to depend on a retained sensory trace [17, 18]. To test the hypothesis that the rSTC supports this trace, we recorded neurons throughout rSTC while rhesus monkeys performed auditory DMS (Figure 1). A substantial population of neurons exhibited sustained modulation of their firing rate during the delay interval, as well as task-related modulation of their sensory responses, as observed in IT during visual DMS [20–23]. Our findings confirm the engagement of these areas during auditory DMS and suggest that the disparity between modalities evident in behavior [17] is rooted in concomitant neurophysiological differences.

Results

Three monkeys (F, S, and K) were trained to perform auditory serial DMS (Figure 1A). Sequences of two to four sounds (~300 ms in duration) were presented at an interstimulus interval (ISI) of ~1 s. Monkeys released a touch bar to indicate the repetition of the first sound (sample) as a match and withheld response to any intervening nonmatch sounds. Stimuli were drawn from a set of 21 exemplars including both synthetic and natural sounds. Behavioral performance declined markedly as the number of nonmatch stimuli in the trial increased [17, 18]. Performance of monkeys F and S was quite similar, but monkey K could not be trained to criterion with >1 nonmatch stimulus (data from this animal are included where appropriate).

Recording sites spanned the rostral auditory cortical areas, including the auditory core (rostral [R] and rostrotemporal [RT]), the adjacent medial and lateral belt, the rostral parabellum, and the tissue extending rostrally to the dorsal temporal pole (Figure 1B; Table S1 available online). Auditory responses were obtained at 36% of 640 sites, yielding 280 responsive units (37% of 749 units tested; 85 from monkey F [all in left hemisphere], 148 from monkey S [117 in right hemisphere and 31 in left hemisphere], and 47 from monkey K [all in left hemisphere]). The median number of effective stimuli was six, and responses were predominantly excitatory (80%). Of the auditory units, 13% also responded at the time of reward delivery, but this epoch of the trial is excluded in later analysis.

Modulation of Delay-Period Activity

In about one-third of the units, a sustained modulation of firing rate during at least one of the delay epochs in the trial was observed (98 out of 280, 35%). Activity was measured over the last 600 ms of each delay and compared to the 600 ms pretrial baseline (Wilcoxon rank-sum test, p < 0.008, correcting for multiple comparisons). As shown in Figure 2, this modulation could take the form of delay suppression (DS) or delay enhancement (DE), which occurred in roughly equal proportion (50 out of 280, 18%, and 48 out of 280, 17%, respectively). DE diminished across the three epochs within the trial, but the sensory responses evoked by match and nonmatch stimuli did not differ in magnitude relative to the sample (Figure 2C). By contrast, DS was sustained across all three epochs of the trial, and the responses evoked by match and nonmatch sounds were suppressed relative to those for the sample (Figure 2D).

Firing rate during the first delay was dependent on the identity of the preceding sample stimulus in 10 units (3.6%; Kruskall-Wallis test using sample identity 1–21 as the single factor, p < 0.008, correcting for multiple comparisons).
No unit showed a significant effect of sample identity in the second delay. For comparison to previous studies that used smaller stimulus sets (e.g., [24]), stimuli were ranked by the magnitude of the sample response, and trials were grouped between the top half and bottom half of the stimuli. By this analysis, firing rate during the first delay carried information about sample identity in 16 units (16 out of 280 = 6%), including 8 DE units (8 out of 42 = 19%) and 1 DS unit (1 out of 36 = 2.8%; Kruskall-Wallis test, p < 0.008). Even among the subpopulation showing elevated activity during the delay, that activity was selective for the prior sample sound in <20% of units.

Among units also recorded during passive presentation, 15% (20 out of 133) showed “delay” modulation during the interstimulus interval (see Supplemental Experimental Procedures), a proportion lower than that observed during behavior ($\chi^2$ test, p = 0.004). Tested separately, this distinction was stronger for DE (7% passive, 15% behaving, p = 0.016) than for DS (8%, 13%, p = 0.17). However, passive effects were observed in only a small minority of those units that showed DE (12%) and DS (24%) during behavior, suggesting that these phenomena were largely specific to the DMS task. As task engagement has been shown to induce both phasic and tonic shifts in firing rate in auditory cortex [24–28], delay modulation (particularly suppression) may reflect a passive process that is strengthened or recruited during DMS performance.

Modulation of the Match Response

The response of a single unit could be influenced not only by stimulus selectivity but also by the context in which that sound appeared in the DMS task. The unit in Figure 3 exhibited match suppression (MS), a reduction in response magnitude for match presentations, relative to those for the same sounds presented as samples (Figures 3A–3D and S1). This effect persisted through at least two intervening nonmatch stimuli (Figures 3B–3D), spanning a total interval of >3 s. This MS appears to result from stimulus-specific repetition suppression of the sample sound because no such effect was seen when the same sounds were presented as a nonmatch (Figures 3G and 3H). To isolate the effect of repetition, we compared responses to match and nonmatch presentations within each trial position (Figures 3E and 3F), revealing a significant effect only at position 2.

An overall match suppression was evident in the averaged population response and persisted through the full trial duration (Figures 4A–4C; all p < 10^{-8}, Wilcoxon sign-rank [WSR] test on firing rates from 25–200 ms). Match responses were also suppressed relative to the nonmatch response at position 2 (Figure 4D; p = 0.001, WSR), but not position 3 (p = 0.14). Suppressive effects were not entirely stimulus specific across the full population of neurons, as revealed by a generalized suppression of nonmatch responses relative to the preceding sample at positions 2 (Figure 4E) and 3 (both p < 0.0004, WSR). The magnitude of the match/sample suppression at position 2 was greater than that of either the match/nonmatch or the sample/nonmatch comparison (p < 0.001, p < 0.0009, respectively, WSR). Nonmatch suppression may stem from partial adaptation to shared features between sample and...
nonmatch sounds, which we have previously shown to predict matching errors during DMS [18].

Timing of Match Response Modulation
The proportion of units showing a significant difference in firing rate between stimulus contexts was calculated in a sliding 100 ms window and overlaid on each panel of Figure 4. The match effect (Figures 4A–4C) showed a biphasic time course, which appears to reflect the sum of two underlying processes (Figure 4F): a transient effect peaking at ~100 ms and a steady buildup during and beyond the stimulus presentation that is also evident in the match/nonmatch comparison (Figure 4D). The early component could reflect recognition of the match, although it was also seen to a lesser degree in the sample/nonmatch comparison (Figure 4E), suggesting it may result from shared features between the sample and nonmatch sounds that were not sufficient to trigger a behavioral match response. The latter component could reflect an accumulating decision process, preparation of the motor response, or anticipation of reward. To control for motor and reward effects, we compared activity between nonmatch presentations that did or did not lead to an erroneous response (Supplemental Experimental Procedures and Figure S2) and confirmed that bar release and reward anticipation had no effect during the stimulus period in the vast majority of neurons.

Averaging across the population obscures the heterogeneity of response modulations in individual units. In AA trials, modulation of the match response relative to the sample was observed in 19% of units (53 out of 280), but these effects were not universally suppressive: 12% showed MS (33 out of 280; Figures 3 and S1), but 7% exhibited the opposite effect, match enhancement (ME; 20 out of 280; Figure S3). Averaged responses of these subpopulations are presented in Figure 5 (for the proportion of units showing effects in the match/nonmatch and sample/nonmatch comparisons, see Figure S4). Whereas MS was evident throughout the first 200 ms of the response (peaking at ~100 ms), ME peaked later in the response (~180 ms after stimulus onset; compare Figures 5A and 5E). After correction for the onset latency of each unit, ME effects lagged MS by a mean of ~50 ms (p = 0.003, Wilcoxon rank-sum test). A contingency analysis (Table S2) revealed a tendency for ME and DE, or MS and DS, to co-occur within the same units at both trial positions (binomial test, p < 0.003).

Delay and Match Effects Diminish Selectively across the Trial
To control for differences in statistical power and anticipatory effects across trial positions, a subset of trials from monkeys F and S (n = 233 units) was reanalyzed as described above. The proportion of units exhibiting significant DS (9%) was unchanged between delay 1 and delay 2 ($\chi^2$ test, p = 1; Figure 6A), but the proportion of units showing DE declined from 13% to 6% ($\chi^2$ test, p = 0.008). Similarly, the proportion of units showing MS was equivalent at positions 2 and 3 (11% and 9%; $\chi^2$, p = 0.43), whereas ME was observed in 5% of units at position 2 but was nearly absent at position 3 (1% of units; $\chi^2$, p = 0.03). Thus, whereas suppressive effects persisted across the duration of the trial, excitatory effects were apparently reset by the intervening nonmatch stimulus. Despite changes in prevalence of the effects, the average magnitude of MS and ME was equivalent across trial lengths (Figure S5). Coincident with this shift in the physiological phenomena associated with the DMS task, the behavioral accuracy of the animals declined sharply after the first nonmatch (Figure 6B), indicating that DMS performance was related, not to the degree of suppression, but to the degree of enhancement in the delay activity and match response.

Time Course of Stimulus Encoding and Retention
The proportion of units showing a significant difference in firing rate across all correct trials, the horizontal line marks baseline firing rate, before averaging across units (shading indicates 1 SEM across trials; black bars indicate 1 SEM across units).
Relative Strength of Sensory and Mnemonic Signals

To capture the relative weight of sensory and mnemonic influences in the second epoch of the trial, the ANOVA model was expanded to include three factors at position 2. The first was the identity of the preceding sample (an integer from 1 to 21), which seldom showed a significant effect. The second was the match/nonmatch condition at position 2 (a value of zero or one), which was taken to represent mnemonic information within the match/nonmatch factor and explained variance of 19%.

Anatomical Distribution of Memory Effects

Our recording sites spanned cortical areas across four hierarchical levels, from core and belt regions to parabelt and STGr (Figure 1B). To quantify whether the prevalence of memory effects differed across levels, the population was split into two positions (1–21), which was nested within the match/nonmatch factor and taken to represent purely sensory information. The unit in Figure 7D showed strong ME, particularly in the latter half of the sensory response. As revealed by the ANOVA model (Figure 7E), sensory selectivity of the response reached its maximum 100 ms after sound onset and persisted for 100 ms after sound offset; by contrast, the influence of match/nonmatch status was maximal between 200 ms and 300 ms. Although this unit showed clear sensory and mnemonic selectivity, the population as a whole conveyed primarily sensory information (Figure 7F), with relatively little influence of the abstract match/nonmatch distinction. At position 2, 41 out of 280 units (15%) showed an effect of the match/nonmatch factor (criterion: >1 significant time bin between 0 ms and 300 ms), and among those units, the mean variance explained was 5.4%. By contrast, stimulus identity was a significant factor in 103 out of 280 units (37%), with a mean explained variance of 19%.
Short-Term Memory in Auditory Cortex

groups: group 1 (n = 101), comprising rostral core and belt, and group 2 (n = 167), comprising parabelt, rostrotemporal-polar (RTp), dorsal temporal pole, and upper bank of the superior temporal sulcus (STS) (Table S1). The proportion of units exhibiting match or delay effects did not differ between the two populations (χ² test, p = 0.4 for DS/DE, p = 0.1 for MS/ME).

By the ANOVA analysis (Figure 7E), match/nonmatch status significantly affected firing rate in 16% of group 1 units and 14% of group 2 units.

Discussion

Delay-Period Effects

Firing rate during the memory delay was modulated in 35% of units (Figure 2), but delay activity was seldom selective for the preceding stimulus. Similar results have been reported in the caudal belt [24], dorsal temporal pole [29], and, recently, in primary auditory cortex (A1) [30]. Our serial DMS paradigm revealed that whereas DS persisted throughout the trial, DE was not robust to interference and diminished in tandem with behavioral accuracy (Figure 6). In this regard, DE seems more closely tied to the sensory trace, whereas DS may represent a more general attentional effect that may be necessary, but not sufficient, to support STM.

DS is seldom mentioned in the visual DMS literature, but it is common in primate studies of auditory STM [13, 24, 29, 30]. Inhibitory projections from PFC to the STGr [31] provide a possible substrate for the robust DS observed in our unit data (Figure 2C) and in human fMRI [32–34]. Whereas visual STM in humans has been reported to rely on active maintenance of information in early sensory cortex [35], auditory STM elicits delay-period suppression [34] that may protect the STM trace from interfering sounds [33]. Of particular relevance to the present study, Linke et al. [33] describe DS that was strongest in subjects who relied on passive, echoic memory, as opposed to active rehearsal. We believe nonhuman primates are limited to this type of auditory memory [17], as they lack the “phonological loop” [36] necessary for rehearsal.

Match Suppression and Enhancement

If the sensory trace is not evident as sustained stimulus-specific delay activity, a tenable alternative is a subthreshold mechanism such as synaptic plasticity [37], which would affect subsequent responses. Responses to match stimuli were modulated relative to responses to sample stimuli in 19% of units, with roughly two-thirds exhibiting MS and one-third exhibiting ME. The ME that we observed, which appeared 80–180 ms after sound onset, has not been described previously in auditory cortex; ME > 300 ms after sound onset has been reported in A1 and Tgd [29, 30] but likely represents response selection and/or feedback from PFC [13]. By contrast, short-latency MS has been reported in A1 (23% of units; [30]), caudal auditory belt (22%; [24]), and Tgd (9%; [29]). Collectively, these data argue against a specialization for STM at the temporal pole and are in favor of a more-distributed representation that includes core and belt. Consistent with this, the rostral STG (“rSTG”) lesion of Fritz et al. [10] comprised the higher-level areas we designated as group 2 (Table S1), yet those animals did not show a deficit in auditory DMS at a 5 s delay.
Comparisons to Visual DMS
The prevalence of match and delay effects we observed is similar to that reported in some studies of visual DMS in IT cortex, which described excitatory and inhibitory delay activity that carried little information about the preceding stimulus [20, 22] and a relatively weak influence of match/nonmatch status on sensory responses [22]. Those recordings covered a broad area of IT cortex, as did ours in the rSTC, using tasks that required only sensory memory for simple colors or patterns.

Delay activity and match effects were observed to a greater degree by Miller et al. [21, 38], who recorded from a restricted IT region in or near the perirhinal cortex, which is strongly associated with visual recognition memory [39]. Our DMS paradigm is modeled after theirs, which required the animals to overcome multiple nonmatch items in a series of complex images. Responses to match stimuli were more suppressed than responses to nonmatch stimuli in rSTC and in IT [21], indicating that MS is stimulus specific. In parallel with our findings, MS appeared at the same latency as the response, suggesting that it originated at or before the level of IT [21]. However, ME in IT cortex appeared at the same latency as MS and survived intervening nonmatch stimuli [23], unlike the ME we observed in auditory cortex, which occurred ~50 ms later than MS and did not survive intervening distractors. The time lag suggests that the ME we observed could have arisen via a top-down signal; Plakke et al. [13] recently found that the population response in lateral PFC shows ME within the first 100 ms after cue onset [13], a latency short enough to potentially drive ME in the rSTC. Alternatively, the lag may reflect temporal integration of the dynamic auditory signal within the rSTC itself, as required for recognition of sounds that evolve over time, but not for recognition of static images.

Adaptation and Context Effects
The latency of the MS effect in rSTC suggests it is a local or bottom-up process, possibly an outgrowth of adaptive processes evident in A1. Although the duration of forward masking or enhancement in A1 would be insufficient to span the 1 s delay in our task [40, 41], context effects lasting ≥1 s have been reported in A1 of the awake primate [42-44]. The time course of adaptation has not been systematically studied in the fields downstream from A1, but evidence from human electrophysiology suggests that the decay of the activation trace is slower in auditory association cortex than in A1 [45], consistent with the long-lasting MS we observed.

Conclusions
Despite ethological evidence for long-term learning and storage of sounds by monkeys (e.g., [46]), their auditory memory falls short of visual and tactile memory when tested by DMS [9, 10], a discrepancy across modalities that may extend to humans as well [47]. Visual memory and tactile memory appear to tap the same cortical system [4], and tactile objects may be encoded as visual images or shapes regardless of the modality of
input. Auditory "objects," by contrast, are more likely to refer to transient events that unfold over time, complicating their storage and retrieval.

Miller and Desimone [23] proposed two parallel mechanisms for visual STM in the temporal lobe: MS, representing a passive memory trace, and ME, representing an actively retained memory, in a distinct population of neurons. In rSTC, ME was neither widespread nor robust to interference, bolstering prior behavioral evidence implying that monkeys may depend primarily upon the passive sensory trace [17, 18]. Alternatively, the active mechanism in audition may emerge in PFC, although whether the ME recently described in lateral PFC [13] is robust to interference remains unknown.

Experimental Procedures

All procedures accorded with the NIH Guide for the Care and Use of Laboratory Animals and were approved by the Animal Care and Use Committee of the National Institute of Mental Health (NIMH). Subjects were three adult male rhesus monkeys (Macaca mulatta). Details of the task, stimuli, training, and behavioral performance were published previously [17, 18]. Detailed methods are available in the Supplemental Experimental Procedures. Briefly, animals sat in a primate chair within a sound-attenuating booth. Head position was fixed, and a sipper tube was positioned for delivery of water reward. The trial sequence is shown in Figure 1A. The standard stimulus set included three exemplars from each of seven categories: modulated noise, band-pass noises, pure tones, frequency-modulated sweeps, rhesus monkey vocalizations, other species’ vocalizations, and environmental sounds. Synthetic sounds were 300 ms in duration, whereas the duration of the natural sounds varied slightly (195–282 ms). Stimuli were presented at 60–70 dB sound pressure level via a loudspeaker (Ohm Acoustics) located 1 m directly in front of the animal.

An MRI-compatible recording chamber was implanted to allow a vertical approach to the rSTC (Figure 1B). Electrode tracks were guided by alignment to an MRI acquired after implantation of the chamber. Most sites (81%) yielded one or two simultaneously recorded units; the 280 units in this report derive from 114 sites yielding one unit, 57 sites yielding two units, 16 sites yielding three units, and one site that yielded four separable units. Spike sorting was verified offline by principal components analysis (Spike2,CED), and spike and event times were exported to MATLAB (MathWorks) for analysis.

After a unit was isolated, sounds were presented in pseudorandom order 8–10 times with an ISI of 2.5 s, as the animal sat passively. If a unit evinced an auditory-evoked response, then the animal was presented with the DMS paradigm. After completion of the recordings, all sites from each hemisphere were aligned to the left hemisphere of an averaged MRI template (Figure 1B; [19]), registered to a combined MRI and histology atlas [48].

Memory effects were investigated in 280 units that were responsive to at least one stimulus. To identify MS or ME (Figures 3, 4, and 5), responses from correct trials were segregated by stimulus context (sample, match, or nonmatch) and sequential position within the trial. In all statistical comparisons, responses were pooled across stimuli, and the number of trials per stimulus was equated between contexts. For each trial type, spike counts during sample and match presentations were compared by a Wilcoxon rank-sum test in a 100 ms sliding window moved in 20 ms steps. A unit was classified as showing an effect if two adjacent bins between 0 and 300 ms were significantly different between contexts (p < 0.01, Bonferroni corrected for overlap of time bins).

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures, five figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2014.10.004.

Acknowledgments

We thank H. Tak, M. Muñoz-Lopez, K. Moorhead, P. Sergo, A. Kloth, and H. Vinal for assistance with animal training and data collection and R.C. Saunders and M. Malloy for providing technical expertise. This research was supported by the Intramural Research Program of the NIMH/NINDS/DHHS.

Received: January 30, 2014
Revised: August 26, 2014
Accepted: October 2, 2014
Published: November 13, 2014

References

Figure 7. Firing Rate across Time May Be Influenced by Sensory and Mnemonic Factors within Individual Units, with a Predominance of Sensory Encoding at the Population Level

(A) Sliding-window ANOVA describes the rise and decay of sensory encoding through time. Firing rate (mean ± SEM across trials) of an example STGr unit (see inset in B) during both sample presentation and the ensuing delay period, overlaid on spike-time rasters (colors designate 21 different stimuli) is shown. Firing rate exceeds baseline + 3 SD at 100 ms latency and remains elevated for several hundred ms after sound offset. The discontinuity in the time axis is placed at 800 ms after onset of the sample and 800 ms before onset of the following sound to accommodate the variable delay durations. Gray horizontal line marks pretrial baseline firing rate.

(B) Single-factor ANOVA in a 100 ms sliding window revealed that the variance in firing rate during the delay was influenced by the preceding sample stimulus, even when mean firing rate during the delay (A) dropped below the prestimulus baseline (600–700 ms). Open circles in (B) mark bins with a significant F value (p < 0.05 after false discovery rate [FDR] correction).

(C) The mean variance explained (±SEM across units) by sample identity across the full population peaked between 100 ms and 200 ms after sound onset and faded to zero by ~300 ms after sound offset, well before the end of the delay interval. Sustained encoding like that seen for the unit in (B) was rarely observed. Inset in (C): units are sorted by “persistence,” i.e., the last time bin to evince significant stimulus encoding, relative to sound offset. Only ~50 units showed sustained selectivity after sound offset (dashed horizontal line at zero); arrow marks the example unit in (B).

(D) Firing rate of a unit in field RTP (see inset in E) for match and nonmatch presentations at position 2 (the same ME unit depicted in Figure S3, upper panels). Open circles mark centers of 100 ms bins with significantly different spike counts for match and nonmatch stimuli; open square marks mean bar-release time on match trials.

(E) Corresponding ANOVA result from the unit in (D) showing the proportion of variance in a 100 ms sliding window explained by three factors: the match/nonmatch status of the position 2 stimulus (black curve); nested within that, the identity of the position 2 stimulus (yellow curve); and the identity of the preceding sample stimulus (blue curve). Open circles mark bins with a significant F value (p < 0.05 after FDR correction).

(F) Mean explained variance (±SEM) for the population; same conventions as in (E).


