Supporting Information

Scott et al. 10.1073/pnas.1209685109

SI Methods

Subjects and Apparatus. Subjects were two adult male rhesus monkeys (*Macaca mulatta*). One monkey (F) was naïve before this study, whereas the other (S) had participated earlier in an unrelated, auditory experiment (1). Experiments took place within a double-walled, sound-attenuating booth (IAC) while monkeys were seated in a primate chair fitted with a metal contact bar. A sipper tube was positioned for delivery of liquid reward (typically water) via a computer-controlled solenoid (Crist Instrument). During training the animals were free to move their heads, but after acquisition of the task their heads were fixed during testing by attaching a surgically implanted titanium head-holder to the primate chair. The data in this report were collected during daily sessions when only behavioral testing was performed, although both animals participated in intermingled physiological recording sessions under the identical task conditions.

The behavioral task was controlled by National Institute of Mental Health Cortex software (Laboratory of Neuropsychology, National Institute of Mental Health; http://dally.nimh.nih.gov), which triggered sound playback via a custom-built interface with a second computer running SIGNAL software (Engineering Design; http://www.engdes.com/). The output of the SIGNAL buffers was flattened across frequency (Rane RPM 26v parametric equalizer), attenuated (Agilent HP 355C and 355D), amplified (NAD Electronics), and delivered via a loudspeaker (Ohm Acoustics) located 1 m directly in front of the animal's head. Sound level was calibrated with a Brüel and Kjær 2237 sound-level meter using A-weighting. Task-relevant events were collected on a CED 1401 acquisition system controlled by Spike2 software (Cambridge Electronic Design). Data were exported to MATLAB (MathWorks) for analysis, and statistics were computed by the MATLAB Statistics Toolbox.

Stimuli. The final set of 21 sounds used in formal testing is illustrated in Fig. S1. All sounds were recorded at 16-bit resolution at a sampling rate of 32 kHz, except for the conspecific monkey vocalizations (Mvocs), for which the sampling rate was 24 kHz. The rhesus vocalizations were from the collection of Dr. Marc Hauser (Harvard University, Cambridge, MA), so the individual callers were unfamiliar to our subjects. All stimuli were equalized in rms amplitude to have approximately equal loudness and were presented at 60 or 70 dB sound pressure level. During training and acquisition of the task, different stimulus sets of varying size were used, but the set in Fig. S1 was introduced after the delayed matching-to-sample (DMS) rule was acquired and was used for all sessions described in this report.

Training and Acquisition of the DMS Rule. Training on the DMS task was begun after a habituation and a pretraining period. The habituation period lasted approximately 1–2 wk, during which the animal learned to sit in a monkey chair and obtain liquid by releasing a touch bar. During pretraining each trial consisted of the presentation of one sound that was repeated with effectively zero delay. The animal obtained a reward for releasing the touch-bar only after the repetition of the sound. A short delay between sounds was then introduced and gradually lengthened until the sounds were separated by an interstimulus interval (ISI) of 800 ms. DMS training started when a single nonmatch sound was introduced between the sample and match presentations on ~50% of trials, requiring the animal to withhold response to the second sound on those trials and to delay release until the third.

The progress of one animal's training (monkey F) is charted in Fig. S2; black points mark actual performance measured by a discrimination index (DI; *Analysis of Behavioral Performance*, below) for each day's training session, overlaid by a running average (black line); the gray line marks the statistical threshold for above-chance performance. Across sessions, the parameters of the task were changed gradually until they matched those in the final DMS task described in the main text (Fig. 1). Training on DMS started when one nonmatch sound was introduced, after which performance was near chance (Fig. S2, sessions 55–110). To provide the animals with an additional cue, the intensities of the nonmatch stimuli were attenuated 20 dB relative to the sample and match stimuli (starting at session 106). The third trial type with two nonmatch sounds was introduced at session 130 (gray arrow), and, after performance recovered (session 134), the attenuation for nonmatch sounds was gradually eliminated (Fig. S2, sessions 144–176; changes are marked by vertical gray lines). Training was considered complete when the monkey could generalize the DMS rule across stimulus sets, as evidenced by consistent performance well above chance (beginning around session 205). Stimulus sets used during training consisted of monkey vocalizations, tones, and band-pass noise (BPN), but no stimuli were identical to those in the final testing set (Fig. S1). A longer trial type (three nonmatches) was tried with this animal, but performance was poor (44% correct over all trial types), so the maximum number of nonmatch stimuli in a trial was fixed at two for both monkeys.

Analysis of Behavioral Performance. Performance on the DMS task was measured by a metric, based on signal detection theory, that incorporates both the accuracy and latency of the behavioral response. Such a technique has proven advantageous in analyzing performance on difficult discriminations because it exploits reaction time (RT) as a measure of confidence, in addition to the subjects' binary decision (2, 3), and provides multiple probability values from which a receiver operating characteristic (ROC) may be derived (4). The bar-release latency within the response window (Fig. 1) was measured relative to the onset of the preceding sound and counted as a hit if that sound was the match or a false alarm (FA) if that sound was a nonmatch. (Release to the sample was considered an aborted trial and discarded.) The distributions of RT from hits and FAs were obtained for each session (Fig. S3*A*). The cumulative probabilities of hits and FAs were then calculated at 50-ms intervals across the 1,200 ms of the response window (Fig. S3*B*). The cumulative hit and FA probabilities, plotted against one another, define a curve in ROC space (Fig. S3*C*), and the area under the curve (ROC value) is used as a measure of task performance, the DI. Perfect performance would yield a DI value of 1, whereas a random response would yield a value of approximately 0.5. To derive a threshold for above-chance performance, the matrix of hit and FA labels was randomly shuffled with respect to the corresponding RTs and the DI computed from the shuffled data. This was repeated 100 times, and the threshold was defined as 2 SDs above the mean of the shuffled DIs.

Fig. S3*D* presents a schematic diagram of the three trial types (zero, one, or two nonmatch stimuli), which were randomly interleaved in the task, and the positions in the sequence at which sample, match, and nonmatch stimuli may appear. The DI measure includes hits and FAs from stimulus positions 2 and 3, and these are combined unless stated otherwise; position 4 was ex-

cluded, because the stimulus at this position was always a match, and therefore no FA is possible.

Performance was also assessed by the FA rate, calculated as FA/(FA + CR), the ratio of the number of false alarms to the sum of false alarms and correct rejections; like DI, FA rate was computed separately at each stimulus position through the trial. The hit rate was calculated as Hits/(Hits + Misses). Percent correct was also measured, both for overall performance and for each trial type. Variability in performance was calculated as the SD across sessions.

SI Results

Performance by Stimulus. The analysis of performance by sound type relied on categorical distinctions imposed by the experimenters, leaving open the possibility that the animals may have developed and used their own categorical distinctions among the stimuli. To examine this possibility and to verify that performance was consistent within the experimenter-defined categories (e.g., was not due to a single outlier stimulus driving the difference between categories), performance was analyzed by individual stimuli. All 21 stimuli were rank-ordered by performance for each monkey and subjected to hierarchical clustering (Matlab Statistics Toolbox; Fig. S4). The color-coded bar graphs and corresponding dendrograms confirm that performance was better for the "temporally simple" stimuli [pure tone (PT), BPN, and frequency-modulated sweeps (FM), coded in warm colors] than for the "temporally complex" stimuli (coded in cool colors).

As mentioned in the main text, performance for rippled noise stimuli [temporally orthogonal ripple complexes (TORCs), in green in Fig. S4] was good for one animal and poor for the other, but consistently so in each case for all three stimuli within that category. Monkey F showed better performance for a rhesus monkey's coo (sound 15) than for the scream or bark (sounds 13 and 14), but this difference was not evident for monkey S. The bars in Fig. S4*A* suggest there were "steps" in the performance function for monkey F, and this was borne out by the cluster analysis. On the basis of this monkey's performance, and using an arbitrary threshold, we could divide the 21 stimuli into three clusters: three temporally simple, high-frequency stimuli (Fig. S4*C*, heavy dark gray branches), the seven other temporally simple stimuli (Fig. S4*C*, black branches), and 11 temporally complex stimuli, including TORCs (Fig. S4*C*, light gray branches). The more continuous decline evident in the bar graph for monkey S (Fig. S4*B*) yielded only two clusters: one containing two noisy environmental sounds for which performance was particularly poor (Fig. S4*D*, 19 and 21, light gray branches) and the remaining 19 sounds (Fig. S4*D*, black branches). The analysis by stimulus is consistent with the analysis by category and does not indicate that the category effects were driven by one stimulus that skews the average performance. Instead, the hierarchical clustering confirmed the distinction between temporally simple and temporally complex stimuli proposed to explain the results of performance sorted by category (Fig. 3).

Interference Across Trials. Our DMS task used a small stimulus set. One effect of such a set is that the same sound will appear in different contexts on different trials, sometimes as a match requiring a response and at other times as a nonmatch to which the response must be withheld. To examine whether stimulus context biased responses on subsequent trials, we calculated FA rate for every nonmatch presentation as a function of the number of trials elapsed since that particular nonmatch stimulus had appeared as a match that led to a rewarded response. The resulting curves (Fig. S5) suggest that there was indeed a cross-trial effect. When the match from the previous trial was presented as a nonmatch on the following trial, the FA rate was nearly 38% for both animals, but this rate declined rapidly with intervening trials. To estimate a threshold for a significant deviation from chance, the computation was repeated 1,000 times with the FA rate randomly shuffled with respect to the match-recency measure; the dashed line marks the mean plus 2 SDs of the shuffled data points. For monkey S, the FA rate dropped to within the expected range after one trial, but for monkey F the FA rate was anomalously high for up to three trials. The reverse effect (an increase in miss errors following a correct rejection of the same stimulus) could not be detected, presumably because of the animals' bias toward positive responses. Given the monkey's generally poor auditory memory, this cross-trial interference is presumably due, not to the monkey recollecting that its response to this particular sound was recently rewarded, but rather to a temporarily strengthened conditioned response to this sound (i.e., a habit; see footnote in Introduction), thereby increasing the probability that it would respond to it again.

1. Yin P, Mishkin M, Sutter M, Fritz JB (2008) Early stages of melody processing: Stimulus-sequence and task-dependent neuronal activity in monkey auditory cortical fields A1 and R. *J Neurophysiol* 100:3009–3029.
2. Carterette EC, Friedman MP, Cosmides R (1965) Reaction-time distributions in the detection of weak signals in noise. *J Acoust Soc Am* 38:531–542.
3. Emmerich D, Gray J, Watson C, Tanis D (1972) Response latency, confidence, and ROCs in auditory signal detection. *Atten Percept Psychophys* 11:65–72.
4. Yin P, Fritz JB, Shamma SA (2010) Do ferrets perceive relative pitch? *J Acoust Soc Am* 127:1673–1680.

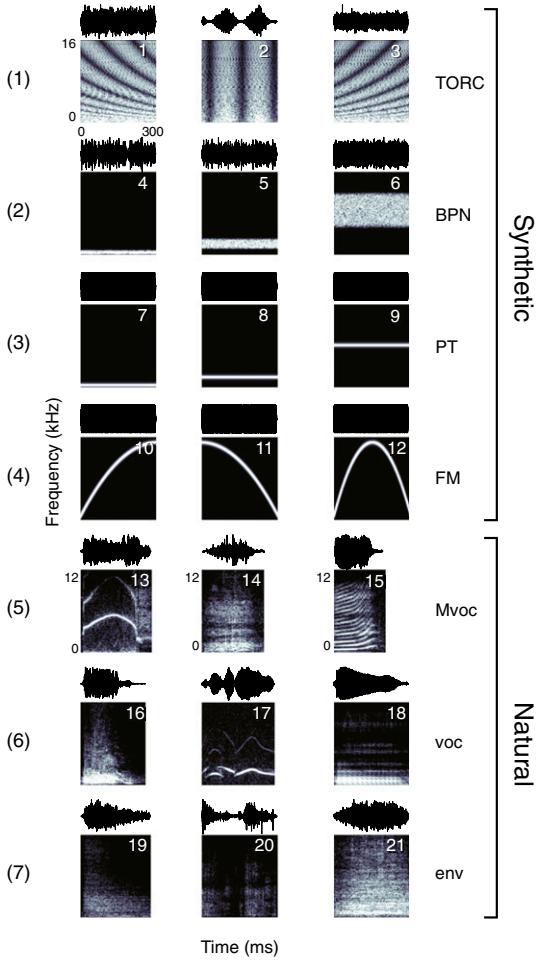

Fig. S1. Set of 21 sounds used during formal DMS testing. The sounds are illustrated as both amplitude–time waveforms (*Upper*) and as frequency–time spectrograms (*Lower*). The set included three exemplars (columns) for each of seven categories (rows): 1, TORCs; 2, 1/3-octave BPN at center frequencies of 512, 2,048, and 8,192 Hz; 3, PTs at the same frequencies; 4, FM, upward, downward, and bidirectional, between 0.25 and 16 kHz; 5, Mvoc, archscream, bark, and coo; 6, other species' vocalizations (voc), dog bark, bird song, and female human vowel /a/; and 7, environmental sounds (env), cage door closing, click of water solenoid opening, and metallic noise. All synthetic sounds were 300 ms in duration, whereas the natural sounds varied in duration, with the Mvocs tending to be shorter than the other categories (stimulus 13, 282 ms; 14, 246 ms; 15, 195 ms; 16, 257 ms; 17, 288 ms; 18, 300 ms; 19, 280 ms; 20 and 21, 300 ms each). The frequency axis on all spectrograms spans 0–16 kHz (linear scale), except those of the three Mvocs, for which the axis spans 0–12 kHz.

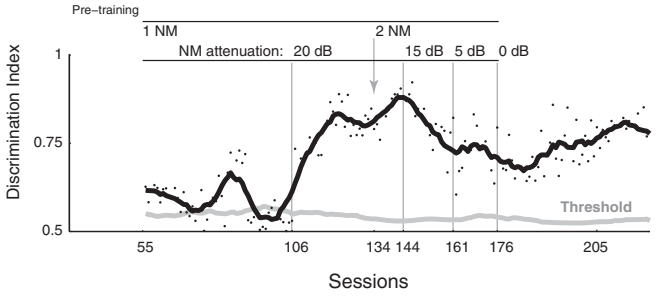

Fig. S2. Acquisition of the DMS task for monkey F. Each dot indicates the DI in one training session, overlaid by a 10-session running average (thick black line); thick gray line marks the threshold for above-chance performance as computed from the shuffled DI within each session (*SI Methods*). Text across the top marks the changes in selected task parameters as training progressed. Upper line indicates introduction of the nonmatch stimulus (1 NM) at trial 55, and up to two nonmatches (2 NM) at session 134 (gray arrow). Vertical gray lines mark the introduction (at session 106) of attenuation of the nonmatch stimuli relative to the intensity of the match and sample stimuli, followed by gradual elimination of the attenuation at session 176. The initial attenuation led to marked improvement in performance, which remained above chance thereafter.

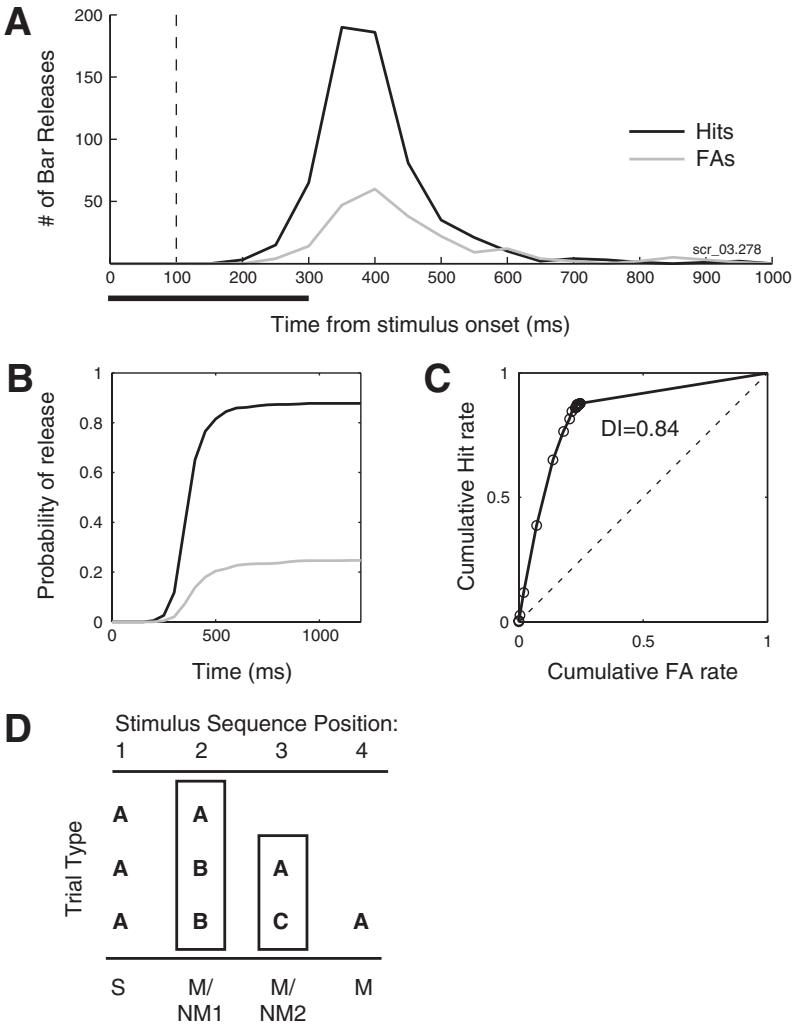

Fig. S3. The DI exploits both RT differences as well as hit and FA probabilities, as illustrated here. (*A*) Distributions of bar-release latencies after sound onset for hit and FA responses (black and gray lines, respectively), for one testing session (n = 931 trials). The horizontal bar indicates the duration of the stimulus (although some stimuli were shorter than 300 ms; the results were the same when the metric analyzed was response latency relative to stimulus offset). Vertical dashed line marks the start of the window for a correct response; releases within 100 ms of sound onset were considered early responses and counted as incorrect. (*B*) Cumulative probability of a hit or FA response over the course of the response window (from same session as *A*). (*C*) ROC plot of hit probability against FA probability, computed at 50-ms intervals from 0 to 1,200 ms. The DI was measured as the area under the curve. Chance performance would trace a curve along the diagonal, yielding a DI of ~0.5. Data are from the same session as *A* and *B*. (*D*) Schematic diagram of stimulus sequence positions (numbered across the top) and the stimuli that may appear within the three trial types (AA, ABA, and ABCA, corresponding to zero, one, or two nonmatch stimuli). The stimulus at position 1 is always the sample (S); stimuli at positions 2 and 3 may be a match (M) or a nonmatch (NM1, NM2); position 4 is always a match. The DI metric in *C* was computed from responses at stimulus positions 2 and 3.

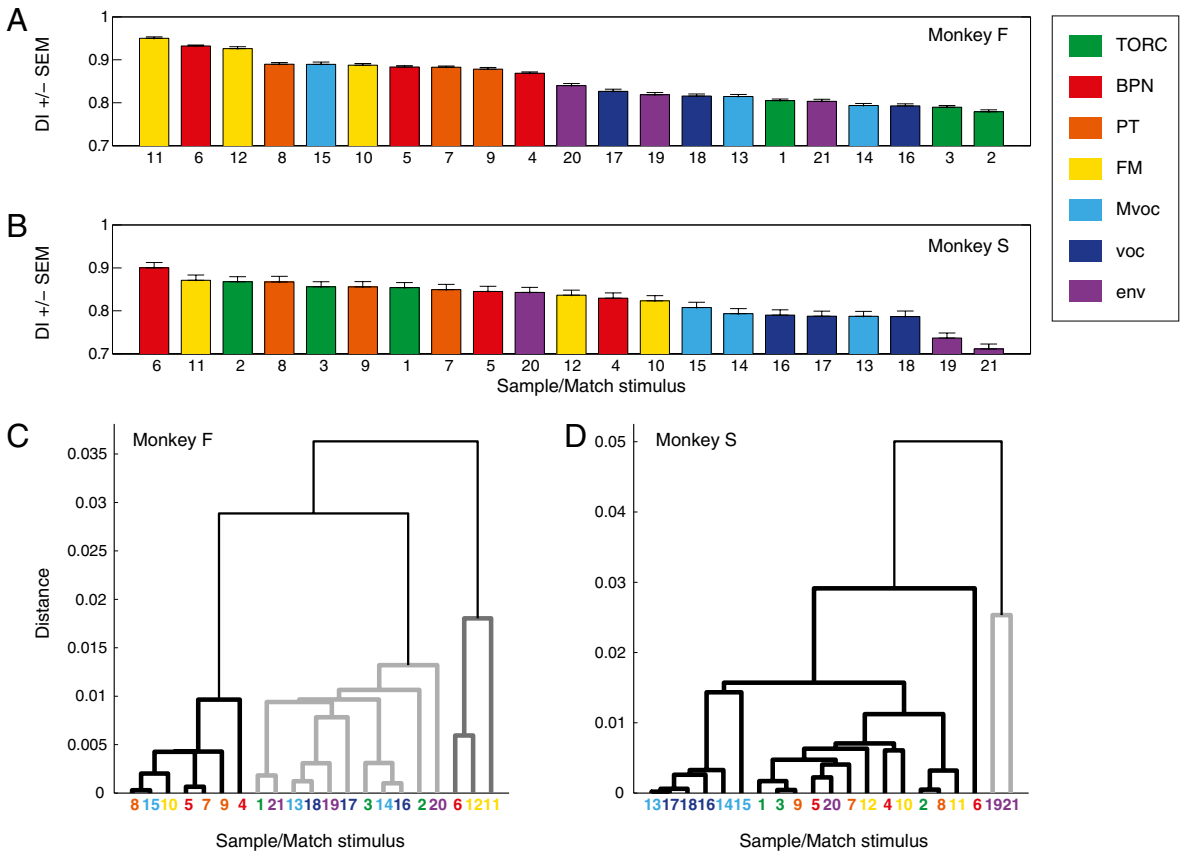

Fig. S4. Analysis of performance by individual stimuli. (*A* and *B*) Performance (DI + SEM) for each sample/match stimulus ranked from best to worst, for each animal. Color code indicates the experimenter-defined "categories" for each stimulus, as described in *SI Methods* and Fig. S1. "Hot" colors indicate temporally simple stimuli, and "cool" colors indicate temporally complex stimuli. (*C* and *D*) Dendrograms display the result of hierarchical clustering of performance for each sample/match stimulus. By an arbitrary threshold, performance for monkey F (*C*) was sorted into three clusters (black, medium gray, and light gray heavy lines) that correspond to the three "steps" or levels apparent in *A*: two groups of temporally simple stimuli for which performance was good, and a third group of complex stimuli for which performance was poor. The decline in performance across stimuli for monkey S was more gradual (*B*) and yielded only two clusters: one for two environmental sounds with very poor performance, and another cluster containing all other stimuli (*D*). Discussion in SI text.

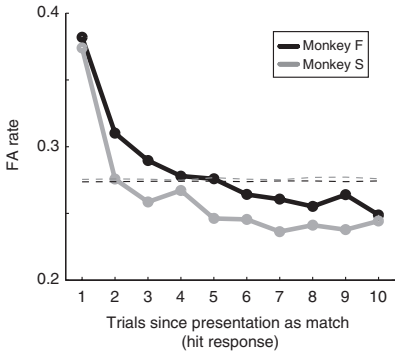

Fig. S5. Cross-trial interference. Curves plot FA rate for every nonmatch stimulus presentation as a function of the number of elapsed trials since that same stimulus appeared as a match (and a hit response to that match was rewarded). Dashed curves mark the threshold for a significant effect, 2 SDs from the mean of 1,000 shuffled simulations (*SI Results*). All points are based on ~6,500–8,000 nonmatch presentations for monkey F and 2,300–2,700 presentations for monkey S.