Title:
Neural Mechanisms of Rhythmic Masking Release in Monkey Primary Auditory Cortex: Implications for Models of Auditory Scene Analysis

Abbreviated Title:
Neural Basis of Rhythmic Masking Release in A1

Authors:
Yonatan I. Fishman¹, Christophe Micheyl³, and Mitchell Steinschneider¹,²

Author Addresses:
Departments of Neurology¹ and Neuroscience², Albert Einstein College of Medicine, Kennedy Center, Room 322, 1410 Pelham Parkway, Bronx, NY 10461
³Dept. of Psychology, University of Minnesota, Minneapolis, MN

Corresponding Author:
Yonatan I. Fishman, Department of Neurology, Albert Einstein College of Medicine, Kennedy Center, Room 322, 1410 Pelham Parkway, Bronx, NY 10461
Tel: 718-430-4110; Fax: 718-430-8588; Email: yonatan.fishman@einstein.yu.edu

Keywords:
multiunit activity, simultaneous suppression, spectral integration, auditory stream segregation

Author Contributions:
Yonatan Fishman and Mitchell Steinschneider designed the experiments and collected the data. Yonatan Fishman and Christophe Micheyl analyzed the data. Yonatan Fishman, Christophe Micheyl, and Mitchell Steinschneider wrote the paper.
Abstract

The ability to detect and track relevant acoustic signals embedded in a background of other sounds is crucial for hearing in complex acoustic environments. This ability is exemplified by a perceptual phenomenon known as ‘rhythmic masking release’ (RMR). To demonstrate RMR, a sequence of tones forming a target rhythm is intermingled with physically identical ‘Distracter’ sounds that perceptually mask the rhythm. The rhythm can be ‘released from masking’ by adding ‘Flanker’ tones in adjacent frequency channels that are synchronous with the Distracters. RMR represents a special case of auditory stream segregation, whereby the target rhythm is perceptually segregated from the background of Distracters when they are accompanied by the synchronous Flankers. The neural basis of RMR is unknown. Previous studies suggest the involvement of primary auditory cortex (A1) in the perceptual organization of sound patterns. Here, we recorded neural responses to RMR sequences in A1 of awake monkeys in order to identify neural correlates and potential mechanisms of RMR. We also tested whether two current models of stream segregation, when applied to these responses, can account for the perceptual organization of RMR sequences. Results suggest a key role for suppression of Distracter-evoked responses by the simultaneous Flankers in the perceptual restoration of the target rhythm in RMR. Furthermore, predictions of stream segregation models paralleled the psychoacoustics of RMR in humans. These findings reinforce the view that pre-attentive or ‘primitive’ aspects of auditory scene analysis may be explained by relatively basic neural mechanisms at the cortical level.
Introduction

Hearing in natural environments requires the ability to detect and track ecologically relevant sounds, such as conspecific vocalizations, embedded among other, less relevant sounds (e.g., Bee and Micheyl 2008; Fay 2008). This ability, in turn, requires that animals perceptually segregate or integrate sounds emanating from multiple sources into auditory ‘streams’, a process known as ‘auditory scene analysis’ (Bregman 1990).

The neural mechanisms of auditory scene analysis are an active research topic (Carlyon 2004; Gutschalk et al. 2005; Snyder and Alain 2007; Nelken and Bar-Yosef 2008; Fishman and Steinschneider 2010; Shamma and Micheyl 2010). Importantly, many aspects of auditory scene analysis, including auditory stream segregation, are phylogenetically widespread (Fay 2008), and are thought to rely on ‘primitive’ mechanisms that do not require learning or attention (Bregman 1990). These considerations motivate the investigation of neural correlates of auditory scene analysis in experimental animal models.

A classic demonstration of the perceptual organization of sound sequences is provided by the ‘auditory stream segregation’ paradigm. In this paradigm, low- and high-frequency tones (denoted A and B, respectively) are presented in an alternating pattern, ‘…ABAB…’. Depending on the frequency separation between the tones and on the alternation rate, the sequence is heard either as a single coherent ‘stream’ of tones alternating in pitch, or as two separate streams- one comprised of ‘A’ tones, and the other of ‘B’ tones (Bregman 1990). Previous neurophysiological studies suggest that auditory stream segregation can be explained by relatively basic neural response properties, such as frequency selectivity, forward suppression, and adaptation, which are present at, or below, the level of primary auditory cortex (A1) (for reviews, see: Micheyl et al. 2007; Fishman and Steinschneider 2010). Earlier investigations led to the ‘population-separation’ model of stream segregation. According to this model, sounds are
perceptually organized into separate auditory streams if they excite distinct neural populations in
auditory cortex, and into a single stream if they excite largely overlapping populations;
frequency selectivity, forward suppression, and adaptation promote functional separation of
neural populations and, therefore, stream segregation (Fishman et al. 2001; Micheyl et al. 2007;
Fishman and Steinschneider 2010). More recent studies have emphasized that in order to be
perceived as separate streams, sounds must not only excite separate neural populations, but do
so in a temporally incoherent (or anti-correlated) fashion. These findings have led to the
‘temporal-coherence’ model of auditory streaming (Elhilali et al. 2009; Shamma and Micheyl
2010).

An important test of both the population-separation and temporal-coherence models of
auditory streaming is to examine whether they can account for the perceptual organization of
sound sequences in which the same sound element (e.g., a frequency component) can be
assigned perceptually to two different streams, depending upon its acoustic context. Such
acoustic sequences are exemplified by the Rhythmic Masking Release (RMR) paradigm, which
is schematically illustrated in Figure 1. In the RMR paradigm, a sequence of tones, ‘Targets’
(Ts), is presented in a rhythmic pattern (Fig. 1A). The target rhythm is perceptually masked by
the introduction of ‘Distracter’ tones (Ds) that are physically identical to, and temporally
intermingled with, the Targets (Fig. 1B). However, the target rhythm can be perceptually
restored (‘released from masking’) when ‘Flanker’ tones (Fs), occupying adjacent frequency
channels, are added synchronously to the Distracters (Fig. 1C). This ‘rhythmic masking release’
causes the Targets to be ‘heard out’ and perceptually segregated from the Distracters, thereby
facilitating identification of the target rhythm (Turgeon et al. 2002, 2005). Thus, physically
identical Target and Distracter tones are perceived as belonging to the same or to separate
streams depending on the absence or presence of the Flankers. Importantly, the RMR effect is
most pronounced when Flankers and Distracters are synchronous and declines with increasing
asynchrony between them, effectively disappearing when their onset asynchrony exceeds about 20 ms (Figure 1D; Turgeon et al. 2002, 2005). Onset synchrony and asynchrony are powerful cues that the auditory system uses to group and segregate, respectively, acoustic components in auditory scene analysis (Darwin and Carlyon 1995). Thus, RMR has been explained in terms of perceptual ‘capture’ of the Distracters by the Flankers, i.e., perceptual grouping of the synchronous Distracter and Flanker tones, which are then segregated from the Target tones (Turgeon et al. 2002, 2005). RMR can be viewed as a particular case of auditory stream segregation, with the two perceptual streams corresponding to the sequence of pure-tone Targets and to the sequence of Distracter-Flanker complexes (see Figure 1). As RMR is a perceptual phenomenon that combines two key principles of auditory perceptual organization: sequential segregation and simultaneous grouping (Bregman 1990), the RMR stimulus paradigm provides an interesting context in which to evaluate neural models of stream segregation.

As no study so far has examined neural responses to stimulus sequences eliciting RMR, the neural basis of the phenomenon remains unknown. Moreover, it is unclear whether existing physiological models of auditory stream segregation can account for RMR. One hypothesis is that the perceptual segregation of the Target tones from the Distracter-Flanker complexes is paralleled, at the neural level, by the excitation of two functionally separate populations: one population responding selectively to the Target tones, and the other responding selectively to the Distracter-Flanker complexes. The question addressed in the present study, is whether such neural populations could be found in A1. The existence of “Target-selective” neurons in A1 is not trivial, because frequency-selective neurons in A1 that are strongly activated by the Target tones are likely to respond also to the (same-frequency) Distracter tones. Therefore, it was unclear a priori whether neural response patterns in A1 would parallel the perceptual segregation of the Target and Distracter tones.
We considered three possible scenarios. In the first scenario, A1 neurons that are tuned to the Targets would be insensitive to the Flankers, so that they would respond equally strongly to the Distracters regardless of whether the Flankers are present or absent; consequently, the temporal response pattern of these neurons would not parallel the psychophysical findings. In the second scenario, A1 neurons that respond strongly to the Targets would also respond equally strongly to the Distracters in the absence of the Flankers; however, owing to the suppression of the Distracter responses by the synchronous Flankers (Shamma and Symmes 1985; Sutter and Loftus 2003; Sadagopan and Wang 2011), these neurons would respond less strongly to the Distracters when these are accompanied by the synchronous Flankers. This suppression of Distracter responses by the synchronous Flankers would effectively create “Target-selective” neural responses. These “Target-selective” responses could support the selective detection of the Target rhythm when Flankers are added to the Distracters, i.e., the main RMR effect (Turgeon et al. 2002, 2005). In the third scenario, neurons that respond strongly to the Target tones would respond equally strongly to the Distracter tones when the Flankers are absent; however, due to spectral-integration (Schwarz and Tomlinson 1990; Kanwal and Rauschecker 2007; Sadagopan and Wang 2009), when Flankers are added to the Distracters these neurons would show larger responses to the Distracters than to the Targets. The enhanced responses to the Distracter-Flanker complexes could potentially support listeners’ ability to attend selectively to the rhythm carried by the Distracter-Flanker complexes. It is important to note that the three above-described outcomes are not mutually exclusive, and in fact, considering the variety of neural response interactions observed in auditory cortex in previous studies, a priori it is quite conceivable that different sites in A1 would exhibit different response patterns.

To determine which of these scenarios is realized in A1, we recorded neural responses to sound sequences similar to those illustrated in Figure 1 in A1 of awake monkeys. In
accordance with the rationale described above, we tested the hypothesis that A1 contains two separate neural populations which respond selectively to the Targets and to the Distracter-Flanker complexes, respectively. Furthermore, we predicted that “Target-selective” responses would be generated by simultaneous suppression of responses to Distracters by the simultaneous Flankers. Consistent with psychoacoustic findings (Turgeon et al. 2002, 2005), we predicted that suppression of Distracter responses by the Flankers in A1 would be independent of the frequency separation between the Flankers and Distracters (at least across the ~1 octave range tested in psychoacoustic RMR experiments), would be maximal when Flankers are presented synchronously with the Distracters, and would diminish or disappear when Flankers are delayed relative to the Distracters by 20 ms or more. Finally, we examined whether the population-separation and temporal-coherence models of stream segregation, when applied to A1 responses, could account for the perceptual organization of RMR sequences in human psychoacoustic studies.

Materials and Methods

Four adult male macaque monkeys (*Macaca fascicularis*) were studied using previously described methods (Fishman et al. 2001; Steinschneider et al. 2003). All experimental procedures were reviewed and approved by the AAALAC-accredited Animal Institute of Albert Einstein College of Medicine and were conducted in accordance with institutional and federal guidelines governing the experimental use of primates. Animals were housed in our AAALAC-accredited Animal Institute under daily supervision of laboratory and veterinary staff. To minimize the number of monkeys used, other auditory experiments were conducted in the same animals during each recording session. Animals were acclimated to the recording environment and trained while sitting in custom-fitted primate chairs prior to surgery.
Stimuli

Stimuli were generated and delivered at a sample rate of 48.8 kHz by a PC-based system using an RX8 module (*Tucker Davis Technologies, Inc.*). Frequency response functions (FRFs) based on pure tone responses characterized the spectral tuning of the cortical sites. Pure tones used to generate the FRFs ranged from 0.15 to 18.0 kHz, were 200 ms in duration (including 10 ms linear rise/fall ramps), and were presented with a stimulus onset-to-onset interval of 658 ms. Resolution of FRFs was 0.25 octaves or finer across the 0.15 to 18.0 kHz frequency range tested.

RMR stimulus parameters were similar to those used in psychoacoustic studies (Turgeon et al. 2002) and in the audio demonstration of RMR provided at the website of Albert Bregman ([http://webpages.mcgill.ca/staff/Group2/abregm1/web/](http://webpages.mcgill.ca/staff/Group2/abregm1/web/)). Stimuli consisted of three continuously looped sequences of tones: Targets Alone, Targets & Distracters, and Targets & Distracters + Flankers presented in separate blocks (Fig. 1). Tones comprising RMR sequences were 50 ms in duration (including 5 ms linear on/off ramps). The frequency of Targets was set equal to the best frequency (BF) of the recording sites (defined below). In the Targets Alone sequences, Target tones were presented with an onset-to-onset interval of 246 ms. In the Targets & Distracters sequences, Distracter tones were introduced and offset from the Target tones by 123 ms. Distracters had a frequency equal to that of the Targets. Thus, the Targets & Distracters sequences were comprised of BF tones presented with an onset-to-onset interval of 123 ms (a tone presentation rate which is twice that of the Targets Alone sequences). In the Targets & Distracters + Flankers sequences, two Flanker tones, one higher and one lower in frequency than the Distracters, were added in sine phase to each Distracter tone. To test the effects of frequency separation (ΔF) on responses to the Distracter tones, Flankers were separated in frequency from Distracters by 0.25, 0.5, and 1.0 octaves. As RMR can also be demonstrated using Distracter-
Flanker complexes comprised of more than a single pair of Flankers (Turgeon et al. 2005), it is important to test whether the physiological findings could be generalized to cases wherein more than two Flankers are presented. Thus, the effect of including additional pairs of Flankers was examined in a limited number of electrode penetrations in the fourth animal. For these electrode penetrations, one pair of Flankers was placed at 0.25 octaves, two pairs of Flankers at 0.25 and 0.50 octaves, and three pairs of Flankers at 0.25, 0.50, and 0.75 octaves, above and below the frequency of the Distracter tones. The condition wherein the onsets of the Flankers were synchronous with those of Distracters (onset asynchrony of 0 ms) was tested in all electrode penetrations. To test effects of onset asynchrony on responses to the Distracters, onsets of Flankers were delayed relative to that of Distracters by 20 and 40 ms. For these Flanker delays, psychoacoustic studies show a reduced or absent RMR effect in human listeners (Turgeon et al. 2002, 2005; see Fig. 1D). Effects of onset asynchrony were tested with a constant ΔF between Flankers and Distracters of 0.25 octaves. As RMR stimuli were comprised of continuous, uninterrupted sequences of tones, there was no ‘first’ tone of the sequences.

All stimuli were presented via a free-field speaker (Microsatellite; *Gallo, Inc.*) located 60 degrees off the midline in the field contralateral to the recording sites and located 1 meter away from the animal’s head (*Crist Instruments, Inc.*). Sound intensity was measured with a sound level meter (type 2236; *Bruel and Kjaer, Inc.*) positioned at the location of the animal’s ear. The frequency response of the speaker was essentially flat (within +/- 5 dB SPL) over the frequency range tested. Pure tone stimuli and individual tones comprising RMR stimuli (Targets, Distracters, and Flankers) were presented at 60 dB SPL. Thus, the complex sounds formed by the addition of the two Flankers to the Distracter tones had an overall level of ~65 dB SPL.
Surgical Procedure

Under pentobarbital anesthesia and using aseptic techniques, holes were drilled bilaterally into the dorsal skull to accommodate matrices composed of 18-gauge stainless steel tubes glued together in parallel. Tubes served to guide electrodes toward A1 for repeated intracortical recordings. Matrices were stereotaxically positioned to target A1. They were oriented at a 20-30 degree anterior-posterior angle and with a slight medial-lateral tilt in order to direct electrode penetrations perpendicular to the superior surface of the superior temporal gyrus, thereby satisfying one of the major technical requirements of one-dimensional current source density (CSD) analysis (Vaughan and Arezzo, 1988). Matrices and Plexiglas bars, used for painless head fixation during the recordings, were embedded in a pedestal of dental acrylic secured to the skull with inverted bone screws. Peri- and post-operative antibiotic and anti-inflammatory medications were always administered. Recordings began after a 2-week post-operative recovery period.

Neurophysiological Recordings

Recordings were conducted in an electrically shielded, sound-attenuated chamber. Monkeys were monitored via video camera throughout each recording session. To confirm and maintain alertness, prior to each stimulus block, animals were observed by entry of the investigator into the recording chamber. In between stimulus blocks animals also received liquid reinforcements and preferred treats. However, the possibility of intermittent drowsiness cannot be excluded. Data acquisition began approximately 10 s following the beginning of RMR sequence presentation to allow neural responses to reach a "steady-state" and to exclude potential "build-up" effects associated with stream segregation (Micheyl et al. 2005).
Intracortical recordings were performed using linear-array multi-contact electrodes containing 16 contacts, evenly spaced at 150 μm intervals (U-Probe; Plexon, Inc.). Individual contacts were maintained at an impedance of about 200 kΩ. An epidural stainless-steel screw placed over the occipital cortex served as a reference electrode. Neural signals were band-pass filtered from 3 Hz to 3 kHz (roll-off 48 dB/octave), and digitized at 12.2 kHz using an RA16 PA Medusa 16-channel preamplifier connected via fiber-optic cables to an RX5 data acquisition system (Tucker-Davis Technologies, Inc.). Signals were averaged on-line by computer to yield auditory evoked potentials (AEPs). CSD analyses characterized the laminar pattern of net current sources and sinks within A1 generating the AEPs, and were used to identify the laminar location of concurrently recorded multiunit activity (MUA). CSD was calculated using a 3-point algorithm that approximates the second spatial derivative of voltage recorded at each recording contact (Freeman and Nicholson, 1975; Nicholson and Freeman, 1975).

To derive MUA, signals were simultaneously high-pass filtered at 500 Hz (roll-off 48 dB/octave), full-wave rectified, and then low-pass filtered at 520 Hz (roll-off 48 dB/octave) prior to digitization and averaging (see Super and Roelfsema, 2005 for a methodological review). MUA is a measure of the envelope of summed action potential activity of neuronal aggregates within a sphere of about 100 μm in diameter surrounding each recording contact (Vaughan and Arezzo, 1988; Brosch et al. 1997; Super and Roelfsema, 2005). Using an electrode impedance similar to that of the electrodes utilized in the present study, MUA and single-unit recordings in primary visual cortex were shown to yield similar results with regard to orientation tuning of closely spaced neurons in the vicinity of the electrode (Super and Roelfsema, 2005). While MUA displays greater response stability than single-unit activity (Nelken et al. 1994; Stark and Abeles, 2007), details of activity occurring in single neurons are lost.
Peristimulus time histograms (PSTHs) provided a complementary and more selective measure of action potential activity in neuronal ensembles. PSTHs were derived from unrectified high-pass-filtered data using a custom spike detection program implemented in MATLAB (MathWorks, Inc.). Trigger thresholds were set at 4 standard deviations above and below the mean baseline activity. As RMR sequences were presented in a continuous loop (with no designated pre-stimulus interval), the baseline used for PSTH analyses was defined as the mean activity occurring within the first 100 ms of response epochs evoked by Targets Alone sequences.

Positioning of electrodes was guided by on-line examination of click-evoked potentials. Pure tone stimuli were delivered when the electrode channels bracketed the inversion of early AEP components and when the largest MUA was situated in the middle channels. Evoked responses to 50 presentations of each pure tone stimulus were averaged with an analysis time of 500 ms (including a 100-ms pre-stimulus baseline interval). Following determination of the BF, RMR stimuli were presented. 100-150 responses to individual tones comprising RMR sequences were acquired and averaged on-line, with a 225 ms analysis window whose onset was locked to the onset of the Distracter tones (even when there was no Distracter tone present, as in the Targets Alone condition). Single-trial data from which averaged responses were derived were stored for off-line statistical and PSTH analyses.

At the end of the recording period, monkeys were deeply anesthetized with sodium pentobarbital and transcardially perfused with 10% buffered formalin. Tissue was sectioned in the coronal plane (80 μm thickness) and stained for Nissl substance to reconstruct the electrode tracks and to identify A1 according to previously published physiological and histological criteria (Merzenich and Brugge, 1973; Morel et al. 1993; Kaas and Hackett, 1998). Based upon these criteria, all electrode penetrations considered in this report were localized to A1. However, the
possibility that some sites situated near the lower-frequency anterolateral border of A1 were located in the rostral field R cannot be excluded.

**General Data Analysis**

MUA and PSTH data analyzed in the present study were based on responses recorded within lower lamina 3, as identified by the presence of large amplitude initial current sinks that are balanced by concurrent superficial sources in upper lamina 3 (Steinschneider et al. 1992). The BF was defined as the pure tone frequency eliciting the maximal MUA or PSTH amplitude within a time window of 10-75 ms post-stimulus onset. This time window brackets the duration of the responses to the individual 50-ms tones comprising the RMR sequences.

It was not possible to conduct a statistically reliable analysis of responses in supragranular and infragranular layers, as frequencies of the stimulus components comprising RMR stimuli were tailored to match the BF of neurons located in lower lamina 3, which was not always identical to that of neurons located at more supragranular and infragranular depths. This potential discrepancy may be due in part to the insertion angle of the multi-contact electrode relative to a ‘cortical column’ after traversing many mm of brain tissue from the dorsal surface of the brain. However, previous findings based on MUA recorded in lower lamina 3 are generally concordant with results obtained from single neurons located in various cortical layers of A1 (e.g., Fishman and Steinschneider 2009; Scholl et al. 2010), suggesting that response patterns in lower lamina 3 are largely representative of those recorded in other cortical laminae.

Importantly, lamina 3 neurons in A1 send their output projections to non-primary auditory areas (Galaburda and Pandya 1983; Jones et al. 1995). Thus, responses in lower lamina 3 may reflect information transmitted to higher-level cortical areas for further sound processing.
Analysis of Neural Responses to RMR Sequences

Neural responses to the RMR sequences were analyzed as follows. For each electrode penetration and under each stimulus condition, the difference between the peak amplitude of spike activity evoked by the Targets and by the Distracters in single-trials was computed. Peak amplitude of the onset response was chosen for analysis instead of other measures (e.g., total spike count) because onset responses are the most prominent features of responses to short duration sounds and signal the occurrence of a new acoustic events, thus making them relevant for auditory scene analysis (e.g., Petkov et al. 2007). Peaks were defined as the maximal amplitude of spike activity occurring within a 50 ms time window beginning at the onset of the Distracter and Target tones. A non-directional paired $t$-test based on these peak response amplitude differences was used to test the aforementioned hypothesis (alpha level = 0.05). A significant positive difference between peak responses to Targets and to Distracters (when accompanied by simultaneous Flankers) would indicate a “Target-selective” neural population, whereas a significant negative difference would indicate a “Complex-tone-selective” neural population that responds more robustly to the Distracter-Flanker complexes than to the pure-tone Targets. Standard error of the mean difference between Distracter and Target responses was estimated via bootstrap re-sampling methods using a custom program implemented in MATLAB (20000 re-samples; Efron and Tibshirani, 1993). Similar paired (directional) $t$-tests were performed on mean data (averaged over single-trials) across sites.

We also examined whether sites displaying “Target-selective” and “Complex-tone-selective” neural response patterns differed with respect to other response properties, including BF and tuning bandwidth. To this end, we fitted FRFs with an Roexpr model (Glasberg and Moore 1990; Fishman and Steinschneider 2006) to derive estimates of the BF and 3-dB excitatory tuning bandwidth. Results of non-directional paired $t$-tests, used to characterize sites as “Target-
selective” or as “Complex-tone-selective” (as described above), were then analyzed as a function of the BFs and bandwidths derived from the model fits.

Computational Models Used to Relate Neural and Psychophysical Responses
To examine whether current models relating neural responses to auditory streaming percepts could account for the perceptual organization of RMR sequences, we implemented a within-channel ‘threshold’ model (Micheyl et al. 2005) and a ‘temporal-coherence’ model (Elhilali et al. 2009) and applied them to the neural responses evoked by the RMR sequences. Model predictions were then compared with psychophysical findings from RMR experiments in which the subjects’ task was to detect a Target rhythm embedded in a background of Distracters, presented in the absence or presence of synchronous or asynchronous Flankers (Turgeon et al. 2002, 2005).

Threshold Model
This model is similar to that used by Micheyl et al. (2005), which was found to successfully account for several features of the perceptual organization of alternating-tone sequences into auditory streams (see also Pressnitzer et al. 2008; Bee et al. 2010). The underlying principle is straightforward, and follows directly from basic signal detection theory, as applied to neural responses: spike counts are compared to a threshold—or, in the parlance of signal detection theory, a decision ‘criterion’ (Green and Swets 1966). If the number of spikes in the PSTH exceeds the threshold, a ‘detect’ decision is made. This process is assumed to take place within frequency-selective neural populations- hereafter referred to as ‘neural channels’. The time course of detect and non-detect decisions in the Target channel determines the percept predicted by the model. If a detect decision was made within the time frame (124-225 ms)
corresponding to the presentation of a Target, this was counted as a ‘hit’; if a detect decision was made within a time frame (1-123 ms) corresponding to the presentation of either silence (for the Targets Alone condition) or a Distracter (in the Targets & Distracters and Targets & Distracters + Flankers conditions), this was counted as a ‘false-alarm’.

The detection threshold was varied systematically over a range bounded, below, by the average spike-count corresponding to spontaneous activity and, above, by the largest observed spike count to compute a ‘receiver operating characteristic’ (ROC). The rationale for letting the criterion vary in the model is based on the assumption that, in psychophysical experiments, participants strive to achieve optimal performance - which requires that their detection ‘threshold’ (or, in terms of signal detection theory, their ‘internal decision criterion’) is adjusted so as to detect as many Targets as possible, while at the same time rejecting as many Distracters as possible. The probability of correct responses that is achieved by using this maximum-likelihood decision strategy in a two-interval forced-choice task corresponds to the area under the ROC (Green and Swets 1966). In order to assess statistical significance of differences between median areas under ROCs across stimulus conditions (un-paired, one-tailed t-test), standard error of the median was estimated using bootstrap (20000 re-samples).

Implementation details are as follows. First, for each stimulus condition and each cortical site, the number of spikes in the PSTH falling within 1-ms bins between 0 and 225 ms after the onset of the Distracter tone was tallied across all stimulus presentations (or trials)—as was done to produce the PSTHs shown in Figure 2. The resulting number of spikes was divided by the number of trials to estimate a time-dependent spike rate. We denote the spike rate as \( r[k] \), where \( k \) refers to the bin number. Second, the ‘false-alarm’ and ‘hit’ probabilities were estimated as:
\[ P_F (j) = 1 - \prod_{k=1}^{123} F(c[j], nr[k]), \quad (1) \]

and

\[ P_H (j) = 1 - \prod_{k=124}^{225} F(c[j], nr[k]), \quad (2) \]

\[ F(c, \lambda) \text{ denotes the cumulative Poisson mass function with parameter } \lambda \text{ evaluated at } c[j], \text{ the current value of the spike-count threshold } (j = 1, \ldots, m), \text{ and } n \text{ denotes the number of sites in the simulated pool. The Poisson distribution is the simplest possible theoretically motivated description of spike-count distributions, and it has been found to provide a good approximation in many previous studies. The products over } k \text{ in these two equations compute the probability that the spike count across the } n \text{ sites never exceeded the threshold, } c[j], \text{ during the timeframe indexed by } k. \text{ Since we used 1-time ms bins, the index, } k, \text{ is expressed in ms. One minus this product is simply the probability that the spike count exceeded the threshold at least once during the considered time frame. The number of A1 sites that are involved in human listeners’ decisions concerning the presence of a tone is currently unknown. Therefore, we explored the influence of the corresponding model parameter, } n, \text{ on model predictions and found that, while performance increased with increasing } n, \text{ the pattern of performance across the different stimuli remained qualitatively unchanged across a wide range of } n \text{ values (from } n = 3 \text{ to } n = 10000). \text{ The results shown here were obtained with } n \text{ set to 100, a number that was found to yield performance levels within the range observed in psychophysical studies. The ROC area was evaluated as,} \]

\[ A = - \sum_{j=1}^{m-1} P_H (j) [P_F (j + 1) - P_F (j)], \quad (3) \]
where $m$ denotes the index corresponding to the largest value of $c$, which was set to $4n$ times
the largest value of $r[k]$ observed for the current cortical site and stimulus condition. Although
indices for stimulus condition and A1 site have been omitted in the above equations for clarity of
notation, the above-described computations were carried out separately for each stimulus
condition and for each A1 site. If the threshold model is sufficient to account for RMR, then the
resultant areas under the model ROCs should parallel subjects’ ability to detect the Target
rhythm as reported in psychoacoustic RMR experiments.

Temporal Coherence Model

To examine whether a direct application of the temporal-coherence model to neural responses
in A1 could explain the enhanced detection of the Target rhythm in RMR, we followed an
approach similar to that described by Elhilali et al. (2009). In that study, the model tested
whether temporal coherence of responses across tonotopically organized neural channels in A1
could explain the perceptual organization of alternating and simultaneous tones. Testing the
temporal-coherence model requires that neuronal responses in the Target channel and the
upper and lower Flanker channels be compared and their correlation evaluated over time. As
we were unable to record simultaneously from three sites with BFs corresponding to the Target,
upper Flanker, and lower Flanker, we simulated responses in the Flanker channels elicited
under the Targets & Distracters + Flankers condition. A similar approach was used by Elhilali et
al. (2009) to simulate the tonotopic distribution of activity in A1 evoked by a stimulus comprised
of two tones of different frequency based on the responses recorded at a single site. To this
end, we presented two additional sets of stimulus conditions wherein the Flanker tone was set
at the BF of the site and the frequency of the Targets and Distracters was set at 0.25 octaves
above or below the BF. Thus, responses recorded under these conditions simulate what each
Flanker site ‘sees’ when the Distracters and Targets are presented simultaneously with the
Flankers under the $\Delta F=0.25$ condition. The temporal-coherence model was then tested by evaluating the correlations over time between neural responses evoked in the Target/Distracter channel and in the (simulated) Flanker channels.

To evaluate the temporal coherence of responses across the Target/Distracter and Flanker channels, we computed Pearson product-moment correlation coefficients between the temporal-response patterns evoked by Targets & Distracters + Flankers sequences with the Target (and Distracter) frequency set equal to the BF, and the temporal-response patterns evoked by Targets & Distracters + Flankers sequences with the upper and lower Flanker frequency set equal to the BF. We then computed the eigenvectors of the resulting three-by-three correlation matrix. The two eigenvector(s) associated with the largest significant eigenvalue(s) were taken to represent the channel groupings (or ‘streams’) predicted by the model, as in Elhilali et al. (2009). The eigenvectors were normalized so that their largest element was equal to 1. Normalized in this way, the eigenvectors can be interpreted as ‘weight vectors’, which indicate how much (as a proportion) each neural ‘channel’ (i.e., neural populations with a BF equal to the frequency of the upper Flanker, the lower Flanker, or the Target/Distracter) contributes to the corresponding stream. Thus, these normalized eigenvectors reveal the spectral content of the auditory stream percepts predicted by the temporal-coherence model. Thus, if the temporal-coherence model is sufficient to explain RMR, then the auditory stream percepts predicted by the model should correspond to the two perceptual streams heard during RMR, one comprised of the Targets and the other of the Distracter-Flanker complexes.
**Results**

Results are based on multiunit responses evoked by RMR sequences in a total of 48 electrode penetrations into A1 of 3 macaque monkeys (monkeys A, D, and E). Data from 2 additional electrode penetrations were obtained in a fourth animal (Monkey L) to examine effects of adding additional pairs of Flankers to the Distracters. A1 sites displaying multi-peaked tuning were uncommon (6 sites in the 3 animals) and are not considered in this report. All other sites displayed a clear BF and sharp frequency tuning characteristic of small neural populations in A1 (Fishman and Steinschneider 2009). Mean onset latency and mean 6-dB bandwidths of FRFs of MUA recorded in lower lamina 3 evoked by tones presented at 60 dB SPL were ~14 ms and ~0.75 octaves, respectively. These values are comparable to those reported for single neurons in A1 of awake monkeys (Recanzone et al. 2000). BF of recording sites ranged from 200 to 14,000 Hz. Data obtained under $\Delta F = 0.5$ and 1.0 octave conditions at sites with a BF less than 250 Hz or greater than 12 kHz were not included, as the level of Flanker tones under these conditions was significantly attenuated due to the frequency response of the speakers.

**Responses to RMR Sequences: Individual Sites**

“Target-Selective” Sites

PSTHs of spike activity evoked by RMR sequences at a representative “Target-selective” site are shown in Figure 2. The FRF of the site is shown in Figure 2A (black: PSTH, gray: MUA). Frequency of Target (T) and Distracter (D) tones (set equal to the BF) and frequencies of Flanker (F) tones (set at 0.25 octaves above and below the BF) are indicated by the three dashed vertical lines superimposed upon the FRF. Responses evoked by Targets Alone (T Alone) sequences and by Target & Distracter (T & D) sequences are shown in Figure 2B.
Target & Distracter sequences evoke two onset responses, occurring at twice the rate as those evoked under the Targets Alone condition. Responses to Distracters and Targets have comparable peak amplitudes that are not significantly different from each other (paired, one-tailed t-test: \( t = -.99; p > .10; \text{df} = 149 \)). However, when Flankers (Fs) are presented synchronously with the Distracters (T & D + Fs), responses to Distracters are markedly diminished relative to responses to the Targets (Fig. 2C; downward arrowheads). Significant attenuation of responses to Distracters in the presence of Flankers is observed under all three \( \Delta F \) conditions tested (paired, one-tailed t-tests: all \( p < .00001; \text{df} = 149 \)). These latter results are consistent with psychoacoustic data indicating that RMR is largely independent of the \( \Delta F \) between Distracters and Flankers over a ~1 octave range (Turgeon et al. 2002). Findings illustrated in this example support the hypothesis that “Target-selective” sites may be generated by suppression of responses to Distracters by the synchronous Flankers. Quantitatively similar results are obtained for MUA data (not shown).

One prediction of our hypothesis is that, at “Target-selective” sites, suppression of responses to Distracters by Flankers will be maximal when Flankers are presented synchronously with the Distracters and reduced when Flankers and Distracters are asynchronous. Results are consistent with this prediction, as exemplified by PSTHs of spike activity recorded at three representative sites shown in Figure 3A-C, respectively. Onset responses evoked by Targets and Distracters under the Target & Distracter (T & D) condition have comparable peak amplitudes that are not significantly different from each other (paired, one-tailed t-tests; all \( p > .10; \text{df} = 149 \)). When Flankers (\( \Delta F = 0.25 \) octaves) are presented synchronously with the Distracters (T & D + Fs), responses to Distracters are significantly attenuated (downward arrowheads) relative to responses to Targets (paired, one-tailed t-tests; all \( p < .00001; \text{df} = 149 \)). However, attenuation of responses to Distracters disappears when Flankers are delayed by 40 ms relative to the Distracters (paired, one-tailed t-tests; all
sites \( p > .10; \text{df} = 149 \). Thus, the relative peak amplitudes of responses to Distracters and Targets under the delayed Flankers condition are comparable to those observed under the Targets & Distracters condition in the absence of Flankers.

“Complex-Tone-Selective” Sites

PSTHs of spike activity evoked by RMR sequences at three example “Complex-tone-selective” sites are shown in Figure 4A-C, respectively. FRFs of the sites are included on the left of the figure (same conventions as in Fig. 2 and 3). In contrast to the suppression of Distracter responses by the simultaneous Flankers at the “Target-selective” sites (Fig. 2 and 3), “Complex-tone-selective” sites display enhanced responses to Distracters when they are accompanied by the simultaneous Flankers. Responses to the Distracter-Flanker complexes are significantly larger than responses to Targets (paired, one-tailed \( t \)-tests; all sites \( p < .00001; \text{df} = 149 \)).

Distribution of “Target-Selective” and “Complex-Tone-Selective” Sites

The response patterns illustrated in Figures 2-4 support the hypothesis that A1 contains two functionally distinct populations of neurons, one which responds robustly to the pure-tone Targets and whose activity is suppressed by the Distracter-Flanker complexes, and the other which responds more robustly to the Distracter-Flanker complexes than to the pure-tone Targets. Together, these two neural populations may be sufficient to account for the perceptual phenomenon of RMR. The former would be suited for detecting the Target rhythm- the primary task evaluated in psychoacoustic studies of RMR, while the latter could account for subjects’ ability to attend selectively to the rhythm carried by the Distracter-Flanker complexes. To examine the prevalence of each of these two populations in A1, we characterized individual sites as “Target-selective” or “Complex-tone-selective” according to the results of non-
directional paired $t$-tests, with significant positive $t$-values (Target responses > Distracter-Flanker responses) and negative $t$-values (Target responses < Distracter-Flanker responses) indicating “Target-selective” and “Complex-tone-selective” responses, respectively. We then examined whether the different response patterns evoked by RMR sequences were associated with more basic response properties of small neural ensembles, including BF and excitatory tuning bandwidth.

Results of this analysis are represented in Figure 5, which show scatter-plots of $t$-values as a function of log relative excitatory tuning bandwidth (3-dB bandwidth divided by BF) and BF under the $\Delta F=0.25$ octave condition. Points above and below the horizontal dashed lines at critical $t$-values = +/-1.96 represent sites displaying statistically significant differences (alpha level = 0.05) between responses to Targets and Distracter-Flanker complexes (paired, non-directional $t$-test; df > 125 for all sites). Hence, points above and below the dashed lines represent “Target-selective” and “Complex-tone-selective” sites, respectively. The majority of our sample of A1 sites displayed a “Target-selective” response pattern (24 out of 47 sites), with only 7 sites displaying significantly larger responses to the Distracter-Flanker complexes (one site with exceptionally broad tuning is excluded from the plot of $t$-values vs. log relative excitatory tuning bandwidths, as its excitatory bandwidth could not be determined). Interestingly, $t$-values were significantly negatively correlated with log relative excitatory tuning bandwidths (Spearman $r = -0.62; p < .0001$) and positively correlated with BF (Spearman $r = 0.42; p < .005$). A significant, although smaller, negative correlation between $t$-values and log relative excitatory tuning bandwidths was also observed under the $\Delta F=0.5$ octave condition (Spearman $r = -0.35; p < .02$; Spearman correlations were not significant for the $\Delta F=1.0$ octave condition, perhaps due, in part, to the smaller sample size). The present findings suggest that while some sites in A1 may show significantly enhanced responses to Distracters in the presence of the simultaneous Flankers, suppression of responses to Distracters by Flankers (“Target-
selectivity”) represents the predominant response pattern in A1. Accordingly, to the extent that A1 is involved in the perceptual phenomenon of RMR, it would be engaged primarily in the task of detecting a narrow-band Target rhythm embedded in the background of Distracters, i.e., the task typically evaluated in psychoacoustic studies of RMR.

Responses to RMR Sequences: Mean Population Data

Given the prevalence of “Target-selective” sites demonstrated above, we sought to determine whether overall population activity averaged across sites in A1 also displayed a “Target-selective” response pattern. Mean PSTHs (black lines) evoked by RMR sequences under synchronous and asynchronous Distracter+Flanker conditions are shown in Figure 6A and 6B, respectively. For display purposes, PSTHs have been temporally aligned relative to the peak of Target-evoked responses prior to averaging across sites to compensate for latency differences across sites with different BFs (e.g., Lakatos et al. 2005). Peak amplitudes of mean responses to Targets and Distracters under the T & D condition are comparable, whereas responses to Distracters are suppressed when Flankers are added synchronously to them. Suppression of responses to Distracters is observed under all $\Delta F$ conditions and disappears when Flankers are delayed by both 20 and 40 ms relative to the onset of the Distracters. Accordingly, results of paired one-tailed $t$-tests shown in Figure 6C indicate non-significant differences between peak amplitudes of Distracter- and Target-evoked responses under the T & D condition (TD) and significantly larger responses to Targets when Flankers are presented synchronously with the Distracters (TDF) under all $\Delta F$ conditions ($p < .0001$). However, when Distracters are delayed relative to Flankers, differences between Distracter- and Target-evoked responses are not statistically significant ($p > .05$). Mean population data in A1 thus show a “Target-selective” response pattern similar to that observed in the majority of individual sites, and are consistent
with the hypothesis that perceptual segregation of Targets from Distracter tones in RMR is facilitated by the suppression of Distracter-evoked responses by the synchronous Flankers. Similar results are obtained for complementary MUA measures of neuronal spiking activity (Fig. 7).

**Modeling Results**

There currently exist two main types of models describing the relationship between neural responses and auditory-stream percepts: models based on ‘spatial’ (e.g., tonotopic) separation between neural populations- the neural ‘population-separation’ model (here evaluated using a neural response threshold), and models based on temporal correlations between neural responses- the ‘temporal-coherence’ model. In the following, we quantitatively test whether these models, when applied to the neural responses obtained in the present study, can account for the perceptual organization of RMR sequences.

**Threshold Model**

Predictions of the threshold model are illustrated in Figure 8. Figure 8A shows the median ROCs derived from PSTH data obtained under each stimulus condition tested in the study. Figure 8B shows the median area under the ROCs in Figure 8A, which provides a quantitative measure of the neural discrimination of the Target rhythm under each stimulus condition. Based on the hypothesis that performance in the two-interval forced-choice psychophysical rhythm-discrimination experiments (Turgeon et al. 2002, 2005) depends on the detection of the Target tones by neurons tuned to the Target frequency, the area under the ROC should vary across stimulus conditions in a manner that parallels human listeners’ performance. Accordingly, the ROC area should be (a) close to 1.0 (perfect performance) in the T Alone condition, (b) close to
0.5 (chance performance) in the T & D condition, and (c) close to 1.0 in the T & D + Fs conditions. Moreover, for the model to account successfully for the effects of Flanker-Distracter asynchrony on RMR, the ROC area should be (d) smaller when Flankers are delayed than when they are presented synchronously with the Distracters. Finally, ROC area should be (e) larger in the 20-ms delay than in the 40-ms delay condition and close to chance (0.5) in the 40-ms delay condition (see Turgeon et al. 2002, 2005). The results shown in Figure 8B are consistent with these predictions. Median ROC areas in the T Alone and T & D + Fs conditions are significantly larger than the median ROC area in the T & D condition (asterisks; un-paired, one-tailed $t$-test; $p$-values are indicated in the figure). On the other hand, median ROC area in the 40-ms delay condition is not significantly different from that in the T & D condition. An intermediate value is obtained in the 20-ms delay condition. These results suggest that a simple threshold model is sufficient to account for the enhanced perceptual salience of the target rhythm in RMR.

**Temporal Coherence Model**

Whereas a simple threshold model appears to be sufficient to account for RMR, it may not be necessary. Hence, we evaluated whether the temporal-coherence model could also successfully account for RMR. Specifically, the auditory stream percepts predicted by the temporal-coherence model should correspond to the two perceptual streams that are heard during RMR, one comprised of the Targets and the other of the Distracter-Flanker complexes. Figure 9A (left) shows the mean MUA evoked under the T & D + Fs condition ($\Delta F=0.25$ octaves) for the three stimulus configurations relative to the BF that were used to evaluate the temporal-coherence model: BF equal to the Target frequency ('Target & Distracter Channel'), the lower-Flanker frequency ('Lower Flanker Channel'), and the upper-Flanker frequency ('Upper Flanker Channel').
Channel'). The resulting two-dimensional profile simulates the spatio-temporal response pattern
in A1 that would be expected if neurons with BFs corresponding to the Target tones and
neurons with BFs corresponding to the Flanker tones were recorded from simultaneously. The
resulting spatio-temporal profile is as predicted: it shows a strong response to the Flankers and
a weak response to the Targets in the two channels tuned to the Flankers, and larger responses
to the Targets in the channel tuned to the Distracters and Targets.

The across-channel correlation matrix is shown next to the MUA waveforms in Figure
9A. As expected from the similar temporal activation pattern in the two Flanker channels, the
correlation coefficient for these two channels is close to 1.0. Lower correlations are obtained
between responses in the upper- and lower-Flanker channels and those in the Target/Distracter
channel. These lower correlations reflect the fact that the response to the Target is large in the
Target/Distracter channel while it is small in the Flanker channels.

Figure 9B shows the first two normalized eigenvectors of the across-channel correlation
matrix. Each element in one of these eigenvectors indicates the proportion to which a given BF
channel ‘contributes’ to the corresponding stream (as explained in Materials and Methods);
thus, these two normalized eigenvectors represent the spectral content of the two streams
predicted by the temporal-coherence model. Figure 9B reveals that, according to the temporal-
coherence model, stimuli under the T & D + Fs condition should evoke a percept of two
streams, one corresponding mainly to the Flanker tones (unfilled symbols, dashed line), and the
other corresponding mainly to the Target and Distracter tones (filled symbols, solid line).
Importantly, while the temporal-coherence model is capable of ‘grouping together’ highly-
correlated activity occurring in the Flanker channels into a single stream and segregating this
activity from the less-correlated activity occurring in the Target/Distracter channel, it fails to
segregate responses evoked by the Targets from those evoked by the Distracters. Hence, when
applied to the responses of frequency-selective neural populations in A1, the temporal-
coherence model cannot replicate the perceptual organization heard during RMR.

On the other hand, the temporal-coherence model can successfully account for RMR if
one considers responses of ‘combination-selective’ neuronal populations that are tuned to
higher-level acoustic features, e.g., the spectral composition of the Distracter-Flanker
complexes. Combination-selective neurons, which show enhanced responses to particular
simultaneous tone combinations, have been found in both primary and non-primary auditory
cortex of monkeys (Schwarz and Tomlinson 1990; Kanwal and Rauschecker 2007; Sadagopan
and Wang 2009). Such neurons could be tuned so that they respond optimally when Flankers
are presented simultaneously with the Distracters and sub-optimally when the pure tone Targets
are presented. Indeed, this response pattern is exemplified by the “Complex-tone selective”
sites identified in the present study, which show larger responses to the Distracter-Flanker
complexes than to the pure tone Targets. As illustrated in Figure 9C, two neural populations that
are selectively tuned to the pure tone Targets and to the Distracter-Flanker complexes,
respectively, will respond in a temporally incoherent or ‘anti-correlated’ fashion when they are
stimulated by T & D + Fs sequences. Consequently, activity in the channel representing the
Targets will form a separate stream from that in the channel representing the Distracter-Flanker
complexes.

Regardless of which model (threshold or temporal coherence) is used to examine RMR,
some suppressive mechanism, similar to that exemplified in the present study, must be involved
in shaping neuronal selectivity to the Targets, such that neurons respond robustly to the Targets
and respond weakly to the identical Distracters when they are accompanied by the
simultaneous Flankers. Thus, in principle, both the threshold model and the temporal-coherence
model, if applied to neural responses selective to higher-level acoustic features, can account for
the perceptual organization of RMR sequences, provided that suppression and spectral
integration are included to shape neuronal selectivity to the Targets and to the Distracter-Flanker complexes.

**Effect of Number of Flankers**

As RMR can be demonstrated using more than a single pair of Flankers, it is important to determine whether the simultaneous suppression observed in "Target-selective" sites when a single pair of Flankers is presented also occurs when multiple pairs of Flankers are presented. Indeed, based on informal listening tests, the RMR effect tends to be perceptually more robust when the Flankers contain multiple frequency components than when they each consist of a single pure-tone. Hence, we examined the effects of including additional Flankers at 0.5 and .75 octaves above and below the Distracter frequency at two additional sites in a fourth animal (Figures 10A and 10B, respectively). Both sites display a “Target-selective” response pattern, with suppression of Distracter-evoked responses when Flankers are presented simultaneously with the Distracters, and diminished or absent suppression when Flankers are delayed by 20 and 40 ms. Suppression tends to be more pronounced when more than a single pair of Flankers is added to the Distracters. Moreover, suppression of Distracter-evoked responses by the simultaneous Flankers does not depend upon the presence of the Targets, as it is also observed when Target tones are omitted from the RMR sequences ('No T' condition in bottom row of Fig. 10A and 10B). Thus, the suppression observed when a single pair of Flankers is presented appears to represent a general neural phenomenon that is not specific to the particular RMR stimuli used in the present study.
Discussion

The present study builds on previous investigations of the neural bases of auditory scene analysis by examining neural correlates of RMR, a perceptual phenomenon that combines two key principles of auditory perceptual organization: sequential segregation and simultaneous grouping (Bregman 1990). Previous studies lend support to the neural ‘population-separation’ model of stream segregation, whereby different auditory streams are represented by separate populations of neurons that are selectively tuned to sounds comprising each stream (Fishman et al. 2001; Bee and Klump 2004; Micheyl et al. 2005; Itatani and Klump 2009). Given the frequency selectivity of neurons throughout the auditory pathway, spectral contrast is a key determinant of neural-population separation. Thus, population-separation models are consistent with the psychophysical observation that frequency separation is an important factor contributing to stream segregation (Bregman 1990; Bregman et al. 2000).

However, under certain conditions, sounds that share frequency components, and that therefore excite the same frequency-selective channels, may nonetheless be perceived as belonging to separate streams (Vliegen and Oxenham 1999; Grimault et al. 2002; Roberts et al. 2002). In RMR, frequency components comprising the Targets represent a subset of the components comprising the Distracter-Flanker complexes. Thus, a population-separation model based solely on neuronal frequency selectivity cannot explain the perceptual segregation of Targets from the spectrally identical Distracters. Another neural mechanism is needed. The present findings indicate that this mechanism is provided by simultaneous suppression of Distracter responses by the concurrently presented Flankers. Suppression effectively generates “Target-selective” neurons, and accounts for two key perceptual features of RMR: the enhanced ability to ‘hear out’ the Targets embedded in the sequence of identical Distracters and the reduced ability to ‘hear out’ the Distracters when they are accompanied by the simultaneous Flankers. Importantly, simultaneous suppression is needed to generate neuronal selectivity to
the Targets, even if the Targets and Distracters are complex tones, because neurons responding strongly to the Target complex must be prevented from responding when the Distracter complex (which is spectrally identical to the Target complex) is accompanied by a Flanker complex.

The present findings are thus qualitatively consistent with the neural population-separation model of stream segregation. While in previous studies the differential responses to the two tones, A and B, comprising ‘ABAB’ sequences could be explained in terms of frequency selectivity and forward (sequential) suppression, here, the differential responses to the Targets and the Distracter-Flanker complexes are attributed to simultaneous suppression across frequency channels. Several previous studies have reported suppression of responses in A1 by simultaneously presented tones, even when tones are separated in frequency by an octave or more and presented over a broad range of levels (Shamma and Symmes 1985; Sutter and Loftus 2003; Sadagopan and Wang 2011). Our findings are also consistent with results of psychoacoustic studies suggesting a role for wide-band non-linear spectral interactions in the perceptual organization of sound sequences (O’Connor and Sutter 2000; Holmes and Roberts 2006).

Suppression of Distracter-evoked responses provides a simple cue for differentiating Targets and Distracters and for perceptually segregating the Target rhythm. Our modeling results demonstrate that this cue can be exploited by the simple and physiologically plausible mechanism of thresholding. With the threshold set high enough to selectively detect the larger Target responses, this model can replicate the qualitative pattern of perceptual performance across most stimulus conditions observed in psychophysical RMR experiments. Thus, owing to simultaneous suppression, a simple read-out mechanism is sufficient to account for the basic RMR phenomenon. A similar thresholding scheme was previously found to be sufficient to explain important aspects of auditory streaming based on neural responses to alternating pure
tones differing in frequency, both at the level of the cochlear nucleus (Pressnitzer et al. 2008) and A1 (Micheyl et al. 2005).

This simple threshold-based mechanism can also account for the finding that RMR is reduced by introducing an asynchrony between the Flankers and the Distracters (Turgeon et al. 2002, 2005). A1 responses to short-duration stimuli are predominantly phasic, with a sharp onset peak occurring within 20 ms following stimulus onset. Hence, if the Flankers are delayed, onset responses to the Flankers will no longer temporally overlap and suppress onset responses to the Distracters, thus yielding Distracter and Target responses that are comparable in amplitude. One aspect of the psychoacoustic data, however, is not fully accounted for by population responses in A1—namely the graded effect on RMR that is observed as stimulus onset asynchrony (SOA) between Flankers and Distracters is increased from 0 to 40 ms. Whereas listeners can reliably discriminate the Target rhythm at SOAs ranging from 0 to 25 ms, neural responses to Distracters tend to be fully restored from suppression at an SOA as short as 20 ms. This suggests that time constants of onset responses and simultaneous suppression in A1 are too short to account for the graded effect of SOA. While speculative, this effect may be explained by activity occurring in other auditory cortical areas which display longer response time constants. For instance, neural responses in the rostral auditory field (Area R) are more sluggish than A1 responses, exhibiting longer onset and peak latencies, diminished temporal precision, and slower decay rates (Recanzone et al. 2000; Bendor and Wang 2008; Scott et al. 2011). This temporal smearing (on the order of 10-20 ms relative to that observed in A1; Scott et al. 2011) would result in greater temporal overlap of onset responses to Distracter and Flanker tones when SOA is equal to 20 ms, and hence a longer time window during which suppression could occur. Thus, response time-constants in non-primary areas may be more consistent with those needed to account for the graded effect of SOA. Indeed, RMR stimuli are expected to activate multiple auditory cortical fields in addition to A1, all of which could
contribute to shaping the perceptual phenomenon of RMR. Physiological differences across species may also account for the discrepancy between the present physiological data and psychoacoustic data in humans.

Although the above-described population-separation model is sufficient to account for important aspects of the RMR phenomenon, results do not demonstrate that this model is necessary. Moreover, this type of model does not explicitly explain why tones which are widely separated in frequency— as in the 1-octave-separation condition— and (presumably) activate well-separated neural populations in A1, are nevertheless perceptually grouped (Elhilali et al. 2009). Indeed, in the psychoacoustic context, RMR has been explained in terms of perceptual ‘capture’ of the Distracters by the synchronous Flankers, leading to their ‘perceptual grouping’ (Turgeon et al. 2002, 2005). This problem is addressed by the ‘temporal-coherence’ model, which extends the neural ‘population-separation’ models by considering also the temporal correlations between responses of neurons tuned to different frequencies— or to other sound attributes (Shamma and Micheyl 2010). According to the temporal-coherence model, neural populations whose responses are temporally correlated are grouped together and represent a single stream, whereas neural populations whose responses are temporally uncorrelated are segregated and represent different streams.

However, we found that when a simple temporal-coherence model was applied to the responses of frequency-selective neurons evoked by T & D + Fs sequences, the predicted stream percepts did not correspond to those heard during RMR. Specifically, while the temporal-coherence model segregated neural activity in the Target/Distracter channel from that in the Flanker channels, it failed to segregate the Targets from the Distracters. On the other hand, the temporal-coherence model may successfully account for RMR if it operates on the output of neurons that are selectively tuned to more complex acoustic features, such as the Distracter-Flanker complexes (Fig. 9C).
Selectivity to these higher-level acoustic features may be achieved via spectral integration and suppression. Spectral integration is exemplified by ‘combination-selective’ neurons in A1, which show enhanced responses when several pure tone components are presented simultaneously (e.g., Sadagopan and Wang 2009), and by neurons in lateral belt regions of auditory cortex, which respond more vigorously to frequency-modulated sweeps and band-pass noise than to pure tones (Rauschecker and Tian 2004; Tian and Rauschecker 2004). Although limited in number, the “Complex-tone selective” neural populations identified in the present study provide additional support for the existence of combination-selective neurons in A1. As shown in Figure 5, “Complex-Tone-selective” response patterns were more common at sites with larger relative excitatory tuning bandwidths and with lower BFs, which tend to exhibit broader tuning bandwidths than sites with higher BFs (e.g., Fishman and Steinschneider 2009). These observations suggest a role for spectral integration in generating the enhanced responses to the Distracter-Flanker complexes at “Complex-Tone-selective” sites. While relatively rare in our sample of sites, neurons with multi-peaked frequency tuning functions, which are also indicative of spectral integration, may further contribute to combination-selectivity in A1 (Kadia and Wang 2003). Thus, from a more general perspective that considers neural populations tuned to higher-level acoustic features, both the population-separation and temporal-coherence models of streaming are compatible with the perceptual phenomenon of RMR. This conclusion is concordant with the view that the brain will exploit whatever cues are available (e.g., spectral or temporal differences) to perceptually organize sounds comprising auditory scenes (Darwin and Carlyon 1995). Accordingly, a hybrid model which incorporates both neuronal population separation and temporal coherence may be best suited for explaining all of the relevant perceptual phenomena related to stream segregation.
An interesting question is whether neurons that respond preferentially to pure tones (Targets) and neurons that respond preferentially to complex tones (Distracter-Flanker complexes) occupy different anatomical locations in A1. We were unable to identify any clear differences between these two populations with respect to their location within A1. While our methods may lack sufficient resolution to discern a systematic organization, these two populations may instead form a ‘patchy’ distribution in A1, an organization which is still consistent with the population-separation and temporal-coherence models of stream segregation.

While the present findings demonstrate parallels between neural response patterns in A1 and the perceptual organization of RMR sequences in humans, the relevance of A1 responses to RMR is tempered by the absence of behavioral data related to the animals’ perception. A similar caveat applies to previous studies of the neural bases of auditory streaming in experimental animals. While there is behavioral evidence that many non-human animals perceptually organize auditory sequences similarly to humans (Fay 2008), the connection between neural response patterns and perception has been indirect. Thus, the present results should be conservatively interpreted as representing a ‘proof of principle’ concerning how neural response patterns at the cortical level may provide the physiological “building blocks” underlying the perceptual organization of RMR sequences. Present findings reinforce the view that pre-attentive aspects of auditory scene analysis rely on relatively basic neural response properties at the cortical level, including frequency selectivity, forward and simultaneous suppression, spectral integration, adaptation, and temporal coherence (Fishman et al. 2001; Micheyl et al. 2005; Elhilali et al. 2009). Additional mechanisms, including synchronized gamma-band activity in auditory cortex, may also contribute to perceptual grouping (or “binding”) of simultaneous acoustic components (Bidet-Caulet et al. 2008), such as those comprising the Distracter-Flanker complexes. As auditory cortical areas outside of A1 are
also involved in stream segregation (Gutschalk et al. 2005), future studies should examine
neural responses to RMR sequences in non-primary auditory cortical fields, and investigate how
these responses are modulated by learning and attention (Fritz et al. 2007; Yin et al. 2007).

Acknowledgments

We are grateful to Jeannie Hutagalung and Kyoko Kamishima for providing invaluable
assistance with animal training, surgery, and data collection, and to Dr. Steven Walkley for
providing histological facilities. We also thank Dr. Shihab Shamma for helpful comments.

Grants

Supported by NIH grants DC00657 and DC07657

Author Contributions

Yonatan Fishman and Mitchell Steinschneider designed the experiments and collected the data.
Yonatan Fishman and Christophe Micheyl analyzed the data. Yonatan Fishman, Christophe
Micheyl, and Mitchell Steinschneider wrote the paper.

References

Bee MA, Klump GM. Primitive auditory stream segregation: a neurophysiological study in the
Bee MA, Micheyl C. The cocktail party problem: what is it? How can it be solved? And why
Bee MA, Micheyl C, Oxenham AJ, Klump GM. Neural adaptation to tone sequences in the
songbird forebrain: patterns, determinants, and relation to the build-up of auditory


**Figure Caption**

**Figure 1.** Schematic representation of the “rhythmic masking release” (RMR) stimuli presented in the study. Frequency and time are represented on the ordinate and abscissa, respectively. Stimulus sequences are comprised of tones presented in a continuous loop. Tones belonging to each perceptual stream typically heard by human listeners are represented by blue and red rectangles, respectively (rectangle length represents 50 ms). In the “Targets Alone” condition (panel A), Target tones (T) with frequencies equal to the best frequency of the recorded neurons are presented in a regular rhythmic pattern, which is heard as a single perceptual stream. In the “Targets & Distracters” condition (panel B), Distracter tones (D) are interspersed among the Target tones, thereby perceptually “masking” the rhythmic pattern in A within a single stream. In the “Targets & Distracters + Flankers” condition (panel C), Flanker tones (Fs) are added synchronously to the Distracters. This leads to the perceptual segregation of the Targets from the Distracters and an enhanced ability to discriminate the original Target rhythm - an effect referred to as RMR. RMR is associated with the perception of two auditory streams, one comprised of Targets (blue rectangles in panel C) and the other comprised of the complex tones formed by the summation of Flankers and Distracters (red rectangles in panel C). However, when Flankers are delayed relative to the Distracters (panel D), RMR disappears and listeners hear two perceptual streams, one comprised of Targets and Distracters and the other comprised of the Flankers (blue and red rectangles, respectively, in panel D).

**Figure 2.** PSTHs of spiking activity evoked by RMR sequences at an individual “Target-selective” site, illustrating effects of frequency separation ($\Delta F$). A. Frequency response function (FRF) of the site (black line: PSTH, gray line: MUA). Frequencies of Targets (T), Distracters (D),
and upper and lower Flankers (F) at a $\Delta F$ of 0.25 octaves are indicated by the superimposed vertical dashed lines. Duration of tones is represented by the horizontal bars below the PSTHs.

B. PSTHs evoked by “Target Alone” (T Alone) sequences (left) and by “Targets & Distracters” (T & D) sequences (right). C. PSTHs evoked by Targets and Distracters in the presence of Flankers, presented synchronously with the Distracters (T & D + Fs), under the three $\Delta F$ conditions tested in the study (.25, .5, and 1.0 octaves). Downward arrowheads indicate attenuated responses to Distracters in the presence of the Flankers. Response attenuation is observed under all three $\Delta F$ conditions. Calibration bar: 50 spikes/sec/stimulus.

**Figure 3.** PSTHs of spiking activity evoked by RMR sequences at three “Target-selective” sites, illustrating effects of onset asynchrony. PSTHs under the T Alone, T & D, T & D + Fs, and T & D + Fs Delayed 40 ms conditions for each of the three sites are shown in A, B, and C, respectively. FRFs of each site (black line: PSTH, gray line: MUA) are shown on the far left in A, B, and C. Same conventions as in Fig. 2. $\Delta F$=0.25 octaves. Downward arrowheads indicate attenuated responses to Distracters in the presence of the simultaneous Flankers, compared with responses to the Targets. Response attenuation is not observed when Flankers are delayed by 40 ms relative to the Distracters. Calibration bars: 100 spikes/sec/stimulus.

**Figure 4.** PSTHs of spiking activity evoked by RMR sequences at three “Complex-tone-selective” sites. PSTHs under the T & D and T & D + Fs conditions for each of the three sites are shown in A, B, and C, respectively. FRFs of each site are shown on the far left in A, B, and C. Same conventions as in Fig. 2. $\Delta F$=0.25 octaves. Asterisks indicate significantly smaller
responses to Targets, compared with responses to the Distracter-Flanker complexes.

Calibration bars: 50 spikes/sec/stimulus.

Figure 5. Paired t-values plotted as function of log relative excitatory tuning bandwidth (left) and BF (right) of sites under the ΔF=0.25 octave condition. Points above and below the horizontal dashed lines at critical t-values = +/-1.96 represent sites displaying statistically significant differences (alpha level = 0.05) between responses to Targets and Distracter-Flanker complexes (paired, non-directional t-test; df > 125 for all sites).

Figure 6. PSTHs of spiking activity evoked by RMR sequences averaged across all sites sampled in the study. A. Mean PSTHs under the T Alone, T & D, and T & D + Fs conditions for ΔFs=0.25, 0.5, and 1.0 octaves, as indicated. B. Mean PSTHs under the T & D + Fs condition when Flankers are delayed relative to the Distracters by 20 and 40 ms, as indicated. Gray shading above the PSTHs indicates SEM. N indicates number of sites contributing to each average. Downward arrowheads indicate attenuated responses to Distracters in the presence of the simultaneous Flankers, compared with responses to the Targets. Response attenuation is not observed when Flankers are delayed by 20 or 40 ms relative to the Distracters. Arrows indicate responses to the delayed flankers. Calibration bar: 100 spikes/sec/stimulus. C. Results of paired t-tests comparing peak amplitude of responses to Targets and Distracters in the absence (TD) and in the presence of Flankers (TDF), presented simultaneously with the Distracters or delayed by 20 or 40 ms, as indicated below the histograms. Height of bars represents mean difference between Target and Distracter responses; error bars represent
SEM. Asterisks indicate a significant difference between Target and Distracter responses (for all significant differences, $p < .0001$).

**Figure 7.** MUA evoked by RMR sequences averaged across all sites under each condition tested in the study. Same conventions as in Fig. 6. Calibration bar: 1 $\mu$V. Gray shading above and below the mean MUA waveforms represents +/- SEM. Results based on mean MUA data are qualitatively similar to those based on mean PSTH data shown in Fig. 6.

**Figure 8.** Threshold model predictions. A. Median ROCs predicted by the model under the different stimulus conditions, as identified in the legend (TD: T & D; TDF: T & D + Fs). The dashed line along the diagonal indicates chance discrimination of the Target rhythm. B. Corresponding median area under ROCs, which quantifies predicted performance (discrimination of the Target rhythm) under the different stimulus conditions. Results are color coded according to the legend in panel A. Error bars represent sample bootstrap estimates of the standard error of the median. The dashed line at 0.5 indicates chance discrimination performance. Asterisks indicate conditions under which median ROC area was significantly larger than that under the T & D condition (black bar; one-tailed t-test; double asterisk: $p < .0001$; single asterisk: $p < .03$). Note that, since the median is a non-linear function of the data, the median area under the ROC need not equal the area under the median ROC; thus, for example, the median ROC area for the T & D condition (black bar in B) is above 0.5, even though the corresponding median ROC (black curve in A) is below the diagonal.
Figure 9. Temporal coherence model predictions. A. Mean MUA evoked under the T & D + Fs condition (ΔF=0.25 octaves) for the three stimulus configurations used to test the temporal coherence model: BF equal to the Target frequency ('Target & Distracter Channel'), to the lower-Flanker frequency ('Lower Flanker Channel'), and to the upper-Flanker frequency ('Upper Flanker Channel'). The corresponding across-channel Pearson correlation matrix is shown next to the MUA waveforms. A duplicate set of MUA waveforms, rotated by 90-degrees, is placed above the correlation matrix to illustrate how correlation coefficients were derived. Filled cells represent self-correlations; Bolded coefficients represent high correlations between Lower and Upper Flanker channels. B. The first two normalized eigenvectors of the across-channel correlation matrix, each of which represents a separate perceptual stream predicted by the model. The model predicts that stimuli under the T & D + Fs condition should evoke a percept of two streams, one comprised largely of the Target and Distracter tones (filled symbols and solid line) and the other of Flanker tones (unfilled symbols and dashed line). Thus, the streams predicted by the model do not correspond to streams that are heard during RMR. C. Illustration of how the temporal coherence model can correctly predict perceptual streams heard during RMR if one considers responses of “Target-selective” neurons (upper MUA waveform), and “Combination-selective” neurons tuned to the Distracter + Flankers complex (lower MUA waveform). The “Target-selective” and “Combination-selective” responses are identical, respectively, to the ‘Target and Distracter Channel’ and ‘Upper Flanker Channel’ waveforms shown in A. Thus, the Upper Flanker Channel response is used to ‘simulate’ the response of a “Combination-selective” neural population tuned to the Distracter + Flankers complex. As described in the text, selectivity to the Targets in the ‘Target & Distracter Channel’ is achieved by suppression of responses to the Distracters by the simultaneously presented Flankers. Each waveform has been normalized to the peak response amplitude for display purposes. Note the temporal incoherence between response patterns for these two neural populations, which,
according to the temporal coherence model, implies two perceptual streams, one comprised largely of Targets and the other of Distracter-Flanker complexes.

**Figure 10.** Effect of number of Flankers at two “Target-selective” sites (panel A and B, respectively). PSTHs under the T Alone, T & D, and T & D + Fs conditions wherein 2, 4, and 6 Flankers were presented are shown, as indicated. Responses under the T & D + Fs condition when 6 Flankers are delayed by 20 and 40 ms are also shown, as are responses under the 6 Flankers condition in the absence of Targets (D + 6Fs, No T). The FRF of each site is shown on the bottom left of A and B. Frequencies of Targets (T), Distracters (D), and the 6 Flankers (Fs) are indicated by the superimposed vertical dashed lines. Downward arrowheads indicate attenuated responses to Distracters in the presence of the simultaneous Flankers, compared with responses to the Targets. Arrow indicates response to the Flankers when they are delayed by 40 ms. The response to Distracters presented alone, while redundant with the response to Targets alone, is provided at the bottom left for ease of comparison with responses under the ‘D + 6Fs, No T’ condition. Calibration bar in A and B: 50 and 100 spikes/sec/stimulus, respectively.
Figures

Rhythmic Masking Release (RMR)

A

"Targets" Alone

Frequency

T T T T T

1 Stream

B

"Targets" & "Distracters"

Frequency

D T D T D T D T

1 Stream

C

"Targets" & "Distracters" + "Flankers"

Frequency

DFs T DFs T DFs T DFs T

2 Streams RMR

D

"Targets" & "Distracters" + "Flankers" Delayed

Frequency

F D F T D F T D F T

2 Streams No RMR

Figure 1
Figure 2
Figure 3
Figure 4
Figure 5
Figure 6
Figure 7
Figure 8

A

B

Proportion Correct

T Alone  TD  TDF 0.25 Oct  TDF 0.5 Oct  TDF 1.0 Oct  TDF Delay 20 ms  TDF Delay 40 ms

**  **  **  **  * ns
Figure 9