Streaming of repeated noise in primary and secondary fields of auditory cortex

Daniela Saderi\textsuperscript{1,2}, Brad N. Buran\textsuperscript{2}, Stephen V. David\textsuperscript{2}

\textsuperscript{1}Neuroscience Graduate Program, \textsuperscript{2}Oregon Hearing Research Center, Oregon Health and Science University

Abstract

Statistical regularities in natural sounds facilitate the perceptual segregation of auditory sources, or streams. Repetition is one cue that drives stream segregation in humans, but the neural basis of this perceptual phenomenon remains unknown. We demonstrated a similar perceptual ability in animals by training ferrets to detect a stream of repeating noise samples (foreground) embedded in a stream of random samples (background). During passive listening, we recorded neural activity in primary (A1) and secondary (PEG) fields of auditory cortex. We used two context-dependent encoding models to test for evidence of streaming of the repeating stimulus. The first was based on average evoked activity per noise sample and the second on the spectro-temporal receptive field (STRF). Both approaches tested whether changes in the neural response to repeating versus random stimuli were better modeled by scaling the response to both streams equally (global gain) or by separately scaling the response to the foreground versus background stream (stream-specific gain). Consistent with previous observations of adaptation, we found an overall reduction in global gain when the stimulus began to repeat. However, when we measured stream-specific changes in gain, responses to the foreground were enhanced relative to the background. This enhancement was stronger in PEG than A1. In A1, enhancement was strongest in units with low sparseness (i.e., broad sensory tuning) and with tuning selective for the repeated sample. Enhancement of responses to the foreground relative to the background provides evidence for stream segregation that emerges in A1 and is refined in PEG.

Keywords: auditory cortex, streaming, sound

Introduction

Sounds generated by different sources impinge on the ear as a complex mixture, with acoustic energy generated by each source overlapping in both time and frequency. The auditory system groups these dynamic spectro-temporal sound features into percepts of distinct sources, in a process known as auditory streaming (Bregman 1990; Griffiths & Warren 2004). Streaming requires statistical analysis of sound sources: streams that come from the same sound source share statistical regularities, and the brain uses these properties as cues for stream integration or segregation (Bregman 1990; Carlyon 2004; Darwin 1997; McDermott 2009; Winkler et al. 2009).

Basic acoustic features, such as separation in frequency and time, are key perceptual cues for segregating simple, alternating sequences of pure tones (Bregman 1978; Bregman et al. 2000; Oberfeld 2014; van Noorden 1975). However, more complex natural sounds often overlap in frequency and time. Segregating spectrally overlapping sounds requires use of perceptual cues such as pitch (Akeroyd et al. 2005; Mesgarani & Chang 2012), timbre (Cusack & Roberts 2000; Roberts et al. 2002; Singh & Bregman 1997), spatial location (Carlyon 2004; Cusack & Roberts 2000; Mesgarani & Chang 2012; Michelley et al. 2007; Roberts et al. 2002; Singh & Bregman 1997), common onset (Elhilali et al. 2009; Shamma et al. 2011), and temporal regularity (Agus et al. 2010; Andreou et al. 2011; Bendixen et al. 2010; Szalárdy et al. 2014). McDermott et al. (2011) specifically tested for the benefit of temporal regularity with a set of naturalistic noise samples that lacked other cues for streaming. Non-repeating samples could not be distinguished from background noise, but humans could identify these samples when they were repeated. The neural basis of this perceptual pop-out remain unknown.

In contrast to the robust perceptual enhancement reported for a repeating foreground stream, studies of neurophysiological activity in auditory cortex demonstrate a suppressive effect of repetition (Pérez-González & Malmierca 2014). Single neurons undergo stimulus-specific adaptation (SSA), where responses to repeated tones adapt, but responses to an oddball stimulus, such as a tone at a different frequency, are less adapted or even facilitated, reflecting perceptual pop-out of the oddball sound (Nelken 2014; Ulansky et al. 2003). In human electroencephalography (EEG), a possibly related phenomenon is observed in a late event-related component, called the mismatch negativity (MMN). Although the dynamics are slower than SSA, MMN is also elicited by rare deviant sounds randomly interspersed among frequent standard sounds (Näätänen 2001). There is no evidence that links SSA or MMN with repetition-based grouping, but it is possible that these processes share some of the same neural circuits. How the brain might use adaptation to a repeating sound to enhance its perception is not known.

In this study, we investigated neural correlates of streaming induced by repetition of complex sounds in primary (A1) and secondary (PEG) fields of the auditory cortex. We first established the ferret as an animal model for streaming of repeating noise sounds by designing a behavioral paradigm that assessed animals’ ability to detect repetitions embedded in mixtures. We then recorded neural activity in A1 and PEG of passive, unanesthetized ferrets. We tested the prediction that auditory cortical neurons facilitate stream segregation by selectively enhancing their
embedded noise stimuli were composed of two overlapping continuous streams of brief (250- or 300-ms) broadband noise samples. The noise samples had second-order statistics (i.e., spectral and temporal envelope correlations) matched to natural sounds (1). Consistent with the goal of this study, the only streaming cue was repetition. These stimuli lacked other conventional streaming cues such as harmonicity and common onset time.

During the initial part of the stimulus, referred to as the random segment, samples for both streams were drawn randomly from a pool of 20 distinct noise samples (1-2.5-sec duration, Figure 1B). When all the samples are drawn randomly, they are perceived as a single stream. The random segment was followed immediately by the repeating segment, in which a target noise sample started to repeat in one sequence but not in the other. In humans, this repetition leads to perceptual separation of the two sequences into discrete streams (1). We refer to the sequence that contains the repeating target sample as the foreground stream, and the concurrent sequence with no repetition as the background stream (Figure 1B).

Two ferrets (O and H) were trained to detect the repetition of the target using a go/no-go detection paradigm. Head-fixed animals were required to withhold from licking a waterspout during the random segment and to lick after the onset of the repeating segment (Figure 1A-B). In each behavioral block (~50-100 trials), two noise samples were chosen as targets, each with 50% chance of occurring in a trial. Changing the identity of the targets between blocks avoided overtraining on a specific target. To measure behavioral performance in a task with continuous distractors and variable target times, we used a discrimination index (DI).

This metric uses hit rate (Ferret O: mean \pm SEM = 0.875 \pm 0.007, n = 636; Ferret H: mean \pm SEM = 0.869 \pm 0.007, n = 504), false alarm rate (Ferret O: mean \pm SEM = 0.522 \pm 0.007; Ferret H: mean \pm SEM = 0.516 \pm 0.010), and reaction time (250ms noise samples, Ferret O: mean \pm SEM = 0.649 \pm 0.017 seconds; Ferret H: mean \pm SEM = 0.826 \pm 0.020 seconds; 300ms noise samples, Ferret O: mean \pm SEM = 0.715 \pm 0.014 seconds; Ferret H: mean \pm SEM = 1.006 \pm 0.042) to compute the area under the receiver operating characteristic (ROC) curve for target detection (2, 3).

Both ferrets were able to learn the task and perform significantly better than chance computed by shuffling reaction times across trials before computing DI (Ferret O: mean \pm SEM = 0.603 \pm 0.005, n = 636; Ferret H: mean \pm SEM = 0.634 \pm 0.007, n = 504; Ferret O: mean shuffled DI = 0.541 \pm 0.004; Ferret H: mean shuffled DI = 0.528 \pm 0.005, p < 0.0001; pairwise t-test; Figure 1C). On average, ferrets responded after 2.9 target presentations (Ferret O: mean \pm SEM = 2.490 \pm 0.041; Ferret H: mean \pm SEM = 3.315 \pm 0.071). Thus, the animals were able to detect the repeating stream and they responded after the target began repeating.

Because two unique targets were used throughout each experimental block, we considered the possibility that animals learned the identity of the target samples during the first few trials and then responded to unique spectro-temporal features of the target samples rather than the repetition. If this were the case, then we predicted that animals would be more likely to respond (i.e., false alarm) to a target sample in the random segment relative to a non-target sample. Since the average reaction time was 2.9 samples and response rates increased over time within a trial, we could not attribute a response (i.e., lick) to a particular sample in the random segment. However, if animals did respond to the target sample, average reaction time measured from the onset of noise samples in the random phase would be expected to be shorter for target samples than for nontarget samples. We compared responses following target and non-target samples and found no apparent difference in mean reaction time between them (Figure 1D). To verify this similarity, we fit the reaction time data with a mixed linear model using sample duration, sample

**Results**

**Ferrets perceive repeated patterns embedded in noise**

To investigate the physiological underpinnings of repetition-based streaming in an animal model, we first developed a behavioral paradigm to assess ferrets’ ability to detect repetitions embedded in noise. Repeated
Relative enhancement of responses to foreground and background streams

Saderi et al., January 27, 2020 – bioRxiv

representation should be observed at the level of the auditory cortex. If this were true, we would expect the neural response to a target sample to change between random and repeating segments.

To test this prediction, we computed the average peristimulus time histogram (PSTH) response across all occurrences of the target noise samples in the random segment (excluding any occurrences during the first 250 ms of the trial), and compared it to the average PSTH response to a balanced number of targets in the repeating segment (Figure 2A). Since the background sample was randomly selected for each presentation of the target, responses to the background sample were averaged out, and the PSTH primarily reflected responses to the target. To quantify changes in the response, we computed the observed gain ($G_o$) term that scaled the PSTH for the random segment to best match the PSTH for the repeating segment. To allow for a direct comparison between gains generated by encoding models (see below), the gain was log-transformed. Thus, negative values indicate suppressed responses during repetition and positive values indicate enhanced responses. For most units in A1 and PEG, log gains were less than zero (Figure 2B), indicating that the average target response in the repeating segment was suppressed with respect to the random segment ($A_1$: mean ± SEM $G_o$ = -0.634 ± 0.038, $n$ = 300, one-sample t-test, null hypothesis population mean = 0, t-statistic = -16.72, $p < 0.0001$; PEG: mean ± SEM $G_o$ = -0.698 ± 0.049, $n$ = 259, one-sample t-test, null hypothesis population mean = 0, t-statistic = -14.11, $p < 0.0001$). Considering previous observations of neural adaptation to repeated stimuli in auditory cortex (6, 7), a decreased response to the target in the repeating segment is not unexpected.

Figure 2: Activity in both A1 and PEG was suppressed during the repeating segment

A. Schematic of two trials during electrophysiological recordings. The background stream consisted entirely of randomly-chosen noise samples (gray). The foreground stream contained randomly-chosen samples during the random segment, which sometimes included the target sample (T). The final pair of noise samples in the random segment included the target (light orange) as it had not yet begun to repeat. Average PSTH responses to each pair of samples that contained the target (thick rectangles) were computed separately for the random segment (black) and repeating segment (dark red). Target samples were subsampled from the repeating segment to match counts between random and repeating segments (see Material and Methods). To minimize effects of stimulus onset adaptation, the very first pair of samples at the beginning of the trial were always excluded from analysis, even if the pair included the target sample. B. Distribution of observed gain change ($G_o$) between random and repeating segments in A1 and PEG. The majority of target responses were suppressed ($G_o < 0$) during the repeating segment. Red dashed line indicates 0 (i.e., no difference between segments). Since results may depend on how well the unit responded to the target, all analyses of neural responses were performed separately for each unique unit-target pair ($A_1$: mean ± SEM $G_o$ = -0.634 ± 0.038, $n$ = 300, one-sample t-test, null hypothesis population mean = 0, t-statistic = -16.72, $p < 0.0001$; PEG: mean ± SEM $G_o$ = -0.698 ± 0.049, $n$ = 259, one-sample t-test, null hypothesis population mean = 0, t-statistic = -14.11, $p < 0.0001$).

Neuronal responses are suppressed during the repeating segment

We recorded multi- and single-unit neural activity in primary ($A_1$, $n$ = 149) and secondary (PEG, $n$ = 136) regions of the right auditory cortex of five ferrets passively listening to the task stimuli. One animal was trained on the repetition embedded noise task (Ferret O), and four animals were naive to the task. Although all physiological data presented here were recorded in non-behaving ferrets, we refer to the same trial structure terminology as in the previous section for consistency. During electrophysiological recordings, one of the target samples was chosen to match each unit’s tuning (eliciting a relatively strong response) while the other was chosen at random. The two targets had equal probability of occurring as the repeating sample on any given trial.

To investigate the neurophysiological underpinnings of streaming due to repetition, we first looked at the raw firing rates of the recorded units in response to the repeated noise samples. Given the enhanced perception of repeating stimuli observed in behavioral experiments (1, 4, 5), we reasoned that evidence for the selective enhancement of foreground

Figure 3: Schematic representation of the stream-dependent PSTH-based encoding model

During the random segment of each trial, the time-varying response of a neuron was modeled as the sum average responses to the two samples composing the stimulus at each point in time. During the repeating segment, responses to the foreground (repeating) and background samples were scaled by gain terms, $G_r$ and $G_b$, respectively, before summing. Thus, the time-course of the response to each sample was the same, but the relative magnitude of the foreground versus background response varied between the random and repeating segments.

Relative enhancement of responses to the repeating foreground stream

Simply comparing the average neural response to the target in the repeating segment to the random segment does not provide insight into any stream-specific effect that might emerge as a consequence of the repetition. To test for evidence of streaming in the neural response, we needed to independently assess the responses to the two streams. Even if the total response was suppressed, activity in the foreground stream in

Within each half of the trial, the total response was suppressed to some degree.

In humans, repetition-based streaming was found to be a pre-attentive phenomenon (4). All electrophysiological data that follow were recorded while animals (trained and task-naïve) were passively listening to the stimuli.
Figure 4: Example A1 units showing foreground enhancement of responses to repeating noise
A. Raster plots (top) and peri-stimulus time histogram (PSTH, gray shading, bottom) responses of an A1 unit to the target sample (T) in the random segment and during the first three repetitions in the repeating segment. Each trace is a response to the target sample and a simultaneous random background sample. Spontaneous rate (SR) is shown (1st column) for reference. Predictions from the stream-dependent model (orange) broken down into the constituent responses to foreground (blue, dashed) and background (black, dotted) streams. This unit had significant foreground enhancement (1.318), resulting from suppression of background ($G_b = -1.332$) but no change for foreground ($G_f = -0.014$). B. Same as in A for a second unit in A1. This unit had significant foreground enhancement (1.566) due to suppression of background ($G_b = -1.014$) and enhancement of foreground ($G_f = 0.552$). Colored dots near each unit name identify their points in Figure 6.

response to the repetition could be enhanced or suppressed relative to the background stream.

To test this prediction, we developed an encoding model in which the neural response was computed as the sum of responses to samples in each stream (stream-dependent model, Figure 3). Using regression analysis, the relative contribution of each noise sample to the PSTH response was computed from the response to the noise samples. In the random segment the response was modeled as the sum of responses to each of the two concurrent samples and a baseline firing rate (Eqn. 1). In the repeating segment, the response to each sample was scaled according to whether it occurred in the foreground or background stream before summing (respectively, gain terms $G_f$ and $G_b$, Eqn. 3). We compared this model to a stream-independent model, in which responses to samples in both streams were scaled equally by a single gain term in the repeating segment ($G_{opt}$, Eqn. 2). The gain terms were exponentiated before scaling the response. Thus, positive gain indicates stronger modulation of the unit’s response (i.e., greater excitation and/or inhibition relative to the spontaneous firing rate), and negative gain indicates weaker modulation.

Figures 4 and 5 show the average response to the target samples and predictions by the stream-dependent model for example units in A1 and PEG, respectively. For some units in both areas, the gain during the repeating segment was unchanged for the foreground but negative for the background stream (Figure 4A: $G_f = -0.014$, $G_b = -1.332$; Figure 5A: $G_f = -0.009$, $G_b = -1.316$). In others, the change was negative for the background stream but positive for the foreground stream (Figure 4B: $G_f = 0.552$, $G_b = -1.014$; Figure 5B: $G_f = 0.612$, $G_b = -1.551$). In all these cases, however, foreground gain was greater than background gain during the repeating segment. Thus, even if the overall gain was negative, there could be a relative enhancement of response to the foreground stream over the response to the background.

Across the population, gain was negative for both foreground and background streams in the majority of unit-target pairs (Figure 6A; A1: mean $\pm$ SEM $G_f = -0.609 \pm 0.041$, mean $\pm$ SEM $G_b = -0.485 \pm 0.035$, n = 304 unit-target pairs; PEG: mean $\pm$ SEM $G_f = -0.935 \pm 0.038$, mean $\pm$ SEM $G_b = -0.518 \pm 0.040$, n = 276 unit-target pairs). Consistent with this result, the stream-independent model also showed a decrease in gain during the repeating segment (A1: mean $\pm$ SEM $G_{opt} = -0.498 \pm 0.038$, mean $\pm$ SEM; PEG: mean $\pm$ SEM, n = 304 unit-target pairs; $G_{opt} = -0.740 \pm 0.042$, n = 276 unit-target pairs). This overall suppression was also consistent with the decrease measured directly from the average target response (Figure 2B; $r = 0.626$ between $G_f$ and $G_b$, $p < 0.0001$).

To test for relative enhancement of responses to the repeated foreground, we measured foreground enhancement, the difference between $G_f$ and $G_{opt}$, for each unit-target pair (Figure 6B). Foreground enhancement was considered significant if the 95% credible interval (CI) for the fitted parameter did not bracket 0 (see Materials and Methods). A subset of unit-target pairs displayed significant foreground enhancement (41/304 in A1, 58/276 in PEG), meaning that during the repeating segment responses to the foreground stream were less suppressed than responses to the background stream. This could be due to a large suppression of responses to the background stream with no change in responses to the foreground stream (e.g., Figures 4A, 5A) or enhancement of responses to the foreground stream combined with a moderate suppression of responses to the background stream (e.g., Figures 4B, 5B). In contrast, fewer units showed foreground suppression in either area (26/304 in A1, 12/276 in PEG). Across the set of unit-target pairs, mean foreground enhancement was significantly greater than zero in A1 (mean $\pm$ SEM 0.124 $\pm$ 0.040, n = 304 unit-target pairs, $p = 0.004$, Wilcoxon signed-rank, sum of ranks = 18778; Figure 6A) and PEG (mean $\pm$ SEM 0.416 $\pm$ 0.040, n = 276 unit-target pairs, $p = 0.0001$, Wilcoxon signed-rank, sum of ranks = 10325).
Next, we wondered if the units showing significant foreground enhancement had distinct sensory encoding properties. For each unit, we quantified lifetime sparseness, a measure of selectivity for any one sample relative to the others (see Materials and Methods, Eqn. 4) (8). This metric is bounded between 0 and 1, where 0 indicates low sparseness (unit responds equally to all stimuli) and 1 indicates high sparseness (unit responds well to only one stimulus). Many units were sparse, responding strongly to only a few noise samples (example in Figure 7A). For each unit-target pair, we also computed target preference, the ratio of evoked response to the target sample versus the average response to all samples (see Materials and Methods, Eqn. 5). A target preference of 1 indicates that the modulation of the unit’s firing rate by the target (increased or decreased spike rate) is equivalent to the average modulation across all samples for all trials.

The relationship between each unit’s sparseness, target preference, and auditory area (A1 or PEG) and its foreground enhancement was quantified by a general linear mixed model with area, target preference, and sparseness as fixed effects and unit as a random effect (Eqn. 6, Figure 7B, C). All two- and three-way interactions between the fixed parameters were included, and results, including significance tests, are shown in Table 1. This model identified a significant relationship between target preference and foreground enhancement in A1 (e.g., for every increase in target preference of 1.0, foreground enhancement increased by 0.46). The effect of target preference was significantly modulated by sparseness. That is, foreground enhancement was stronger in units with high target preference, but this effect decreased with increasing sparseness (e.g., for...
We compared prediction accuracy of the stream-dependent STRF to two control models: a stream-independent STRF, in which stream identity was shuffled in time before fitting (Figure 8A, B, blue path), and a baseline STRF, in which both segment and stream identity were shuffled. By separately shuffling state variables related to repetition and stream identity, we controlled for the influence of each variable on response gain while keeping the number of free parameters in each model constant. We used a cross-validation procedure to prevent the possibility of overfitting to the model estimation data, testing the models on a common dataset that was not used for fitting.

The stream-dependent STRF predicted time-varying responses more accurately than the stream-independent STRF in both A1 (mean ± SEM 0.277 ± 0.013 and 0.267 ± 0.013, respectively, p < 0.0001, Wilcoxon signed-rank test, Figure 8D) and PEG (0.260 ± 0.015 and 0.250 ± 0.015, respectively, p < 0.0001). Thus, the STRF-based models confirm a significant influence of stream identity on relative gain.

**Figure 8:** STRF-based model corroborates PSTH-based model findings of stream-specific gain changes in A1 and PEG

A. Schematic of the spectro-temporal receptive field (STRF)-based encoding models. Spectrograms of the foreground and background streams (left) are scaled by context-dependent gain. Separate gain terms are applied to the foreground (G_f) and background (G_b) streams during the repeating segment. The sum of the scaled spectrograms then provides input to a traditional linear-nonlinear STRF, which in turn predicts the time-varying response. For the stream-independent model (bottom right), stream identity was shuffled, leading to similar gain for both streams. For the stream-dependent model, gains could differ. In this example, the higher gain for the foreground emphasizes the repeating noise sample (top right). B. Time-varying response for example unit in A for the actual unit response (grey), stream-independent model prediction (blue), and stream-dependent model prediction (orange). While incorporating stream-dependent gain improved prediction accuracy in both areas, predictions were generally more accurate in A1 than in PEG for both models. This difference is likely due to a greater prevalence of nonlinear encoding properties and non-auditory activity in PEG, compared to A1 (11). The same pattern of enhanced prediction accuracy was observed for stream-dependent PSTH-based models, when compared to fits with stream identity shuffled (Figure 8C).

The comparison of stream-independent and baseline STRFs measured the effect of repetition alone on evoked activity (independent of stream identity). On average, the stream-independent STRF had greater

---

**Table 1: Regression analysis of auditory tuning effects on foreground enhancement**

Results of the mixed linear model for foreground enhancement, with target preference, sparseness, and area as fixed effects. See text in Results and Materials & Methods for details of model fit. Results presented are a post-hoc test of contrasts. The effect and coefficient (Coef.) columns indicate the predictor/coefficients that were set to 1 for the post-hoc test of contrasts. Since each unit contributes two measures of foreground enhancement (one for each target), unit was included as a random effect in the model.

<table>
<thead>
<tr>
<th>Area</th>
<th>Effect</th>
<th>Coef.</th>
<th>Std. err.</th>
<th>z</th>
<th>p [a]</th>
<th>CI lower</th>
<th>CI upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Intercept</td>
<td>0.15</td>
<td>0.04</td>
<td>3.57</td>
<td>0.0004</td>
<td>0.07</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>Sparseness</td>
<td>-0.42</td>
<td>0.32</td>
<td>-1.33</td>
<td>0.1835</td>
<td>-1.04</td>
<td>0.20</td>
</tr>
<tr>
<td>PEG</td>
<td>Intercept</td>
<td>0.42</td>
<td>0.32</td>
<td>1.33</td>
<td>0.1835</td>
<td>-1.04</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Target</td>
<td>-0.42</td>
<td>0.32</td>
<td>-1.33</td>
<td>0.1835</td>
<td>-1.04</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>Preference</td>
<td>0.42</td>
<td>0.32</td>
<td>1.33</td>
<td>0.1835</td>
<td>-1.04</td>
<td>0.20</td>
</tr>
</tbody>
</table>

---

**Foreground enhancement increases accuracy of spectro-temporal receptive field models**

To validate the gain changes observed in the PSTH-based model and to quantify their effect on sound-evoked activity, we modeled the same data with a spectro-temporal receptive field (STRF). In the classic linear-nonlinear (LN) STRF (see Material and Methods, Eqsns. 7-9), the time-varying neural response is modeled as a linear weighted sum of the stimulus spectrogram, followed by a static nonlinearity to account for spike threshold (9, 10). We developed a context-dependent model, in which spectrograms for each stream were scaled separately by a gain term before providing input to a traditional LN STRF (Figure 8A, B, orange arrows, Eqsns. 10-11). This scaling followed the same logic as the stream-dependent PSTH-based model described above. That is, a separate spectrogram for each stream was scaled by free parameters that depended on segment (random or repeating) and, during the repeating segment, stream identity (foreground or background). If neural responses were enhanced selectively for one stream, then the gain applied to that spectrogram was greater than gain for the other spectrogram. We refer to this model as the stream-dependent STRF.
Sound features that belong to the same source tend to begin and end at the same time. This phenomenon has been formalized for streaming in the temporal coherence model (15, 16). Teki et al. (2016) demonstrated that human listeners are sensitive to repetition of sounds presented in the context of a random mixture of chords. Similar to our findings, the authors observed that repeating sounds tend to fuse together into a foreground that emerges from a randomly changing background (17, 18). Here, we propose that enhancement of neural responses to the foreground contributes to streaming of repeating sounds.

However, it is important to note that an expectation of enhanced responses to “foreground” stimuli may reflect a biased expectation. There is no a priori requirement for sounds that perceptually pop out to evoke enhanced (or less suppressed) neural responses. For example, Bar-Yosef and collaborators investigated cat A1 activity during presentation of bird chirps in background noise, simulating a natural auditory scene (19, 20). To their surprise, neural responses were dominated by the background noise. The authors interpreted this finding in an evolutionary context, in which it is advantageous for prey to pay attention to subtle changes in the background to avoid predators using foreground sounds to mask the sound of their approach. Thus, it may be important to consider the broader behavioral context of the animal model when interpreting our findings. More experiments testing streaming in natural conditions will elucidate how the brain streams repeated sound features with behavioral relevance.

**Streaming analysis**

To capture differences in responses to simultaneous repeated and non-repeated noise samples, we relied on model predictions. In this paradigm, the neural response is the sum of responses to two simultaneous stimuli, and the component responses cannot be separated in the raw neural firing rate. Therefore, we constructed encoding models to computationally tease apart stream-dependent activity. This analysis showed that, even though most neural responses were suppressed by repetition (likely due to adaptation, Grill-Spector et al. 2006; Pérez-González & Malmierca 2014; Ulainovsky et al. 2004), responses to the foreground stream where less suppressed than the background or even enhanced. The same methodology could be applied to other datasets where there is a need to separate neural responses to simultaneously occurring inputs.

The challenge of separating responses to simultaneous sounds has been addressed for neural population activity using a similar approach (23). Ding and Simon recorded human brain activity via magnetoencephalography (MEG) while subjects attended to one of two background.

**Discussion**

In natural environments, temporally co-varying sound features tend to be grouped by the brain into a single object (12). Consistent with this observation, subjects are able to identify previously unheard noise samples when they are repeated simultaneously in a mixture with non-repeating samples (1). The goal of this study was to investigate the neural underpinnings of auditory streaming cued by repetition. We developed an animal model for repetition detection and found evidence for enhanced representation of a repeating foreground stream in single-unit activity in auditory cortex. This representation emerges hierarchically, as streaming effects are stronger in secondary (PEG) than in primary (A1) auditory cortical fields.

**Mechanisms of repetition-induced stream segregation**

Studies of streaming at the single-unit level have primarily used alternating sequences of pure tones (13, 14). Relevant to the current study, Michiel and collaborators presented sequences of “ABA” tone triplets to awake macaques and examined activity evoked in A1 (13). Tone A was chosen to be on the best frequency of the recorded unit, while tone B differed by 1-9 semitones from tone A. The authors found that, even if responses to both tones decreased relative to their presentation in isolation, responses to the non-preferred B tones decreased more. We observed a similar effect, that relative enhancement of the foreground stream was more pronounced in units well-tuned to the repeated noise sample. Thus, this study provides evidence that the same principles generalize to streaming of simultaneous, naturalistic sounds.

**Figure 9: Improved performance by the stream-dependent model predicts foreground enhancement**

Scatter plot compares foreground enhancement against the difference in prediction correlation coefficient between stream-dependent and segment only models in A1 (left) and PEG (right). Each circle represents a unit-target pair. Units showing significant enhancement for the stream-dependent model are indicated in purple. Headings indicate mean prediction correlation (Pearson’s $r$) and corresponding $p$-value.

Prediction accuracy than the baseline STRF in both areas (baseline STRF mean ± SEM prediction correlation, A1: $0.261 ± 0.013$, $p = 0.0007$, PEG: $0.247 ± 0.015$, $p = 0.014$, Wilcoxon signed-rank test, Figure 8D). Moreover, overall gain was suppressed during the repeating segment (mean ± SEM A1: $-0.098 ± 0.009$, PEG: $-0.047 ± 0.010$), as observed in the PSTH-based models above (Figure 6). Thus, the STRF-based models provide additional evidence for a streaming mechanism in which repetition leads to overall suppression of the neural responses, but with less prominent suppression of the foreground stream relative to the background.

To measure the relative enhancement between the two streams, we compared the stream-specific gain terms from the model fits, equivalent to $G_f$ and $G_b$ in the stream-dependent PSTH model. We observed a significant foreground enhancement in both A1 (mean ± SEM $0.180 ± 0.024$, $p < 0.0001$) and PEG (mean $0.298 ± 0.024$, $p < 0.0001$, Wilcoxon signed-rank test, Figure 8E). These changes in gain followed the same pattern as in the PSTH-based model above (correlation between foreground enhancement for PSTH- and STRF-based models in A1: $r = 0.39$, $p < 0.0001$ and PEG: $r = 0.22$, $p = 0.0005$).

Since only some units in A1 or PEG showed foreground enhancement, we also expected that the performance of the stream-dependent STRF should vary and that the benefit of a stream-dependent model should be greatest for units showing the strongest foreground enhancement. We compared the difference in prediction accuracy between stream-dependent and stream-independent STRFs with foreground enhancement measured in the stream-dependent STRF (Figure 9). In both A1 and PEG, these values were positively correlated (A1: $r = 0.261$, $p < 0.0001$, PEG: $r = 0.335$, $p < 0.0001$). Thus, in units showing enhanced gain for the repeating stream, models incorporating stream dependent gain show the greatest improvement in prediction accuracy.
simultaneous speech samples. They fit a separate STRF (or more precisely a “TRF”, since MEG did not resolve spectral tuning) for each of the speech streams. Neural activity preferentially synchronized to the speech envelope of the attended speaker. Furthermore, the source location of the attended vs. non-attended signals suggested a hierarchy of auditory processing, in which representation of the attended object emerges in posterior auditory cortex (23). These results are consistent with the foreground enhancement observed in the current study, suggesting that top-down attention and bottom-up pop-out effects could be mediated by common mechanisms.

Another approach used to investigate the neural signature of streaming is stimulus decoding (23–25). A decoding model is effectively an inverse of the STRF, using neural population activity to reconstruct the sound stimulus. Using human MEG and electrocorticography recordings, attended stimuli can be reconstructed more accurately than simultaneous non-attended stimuli, consistent with enhanced coding of the attended stream (23, 25). In the current study, stimuli were not fixed across experiments. Thus, the data do not support straightforward population decoding. However, with appropriate changes to experimental design, we predict that stimulus reconstruction should be more accurate for the foreground stream than the background stream.

Relation of repetition enhancement to stimulus-specific adaptation

The ability of the brain to detect regularities is not only crucial for identifying an auditory object embedded in a noisy scene, but also for making predictions about the environment, thereby making the system sensitive to deviance (26, 27). Substantial effort has been devoted to understanding mechanisms of deviance detection, with a focus on the mismatch negativity (MMN) in humans (28) and on stimulus-specific adaptation (SSA) in animals (6, 7, 29). Evidence for SSA has been found in the non-lemniscal inferior colliculus and thalamus (30–32), but the first primary region in which SSA has been shown to be prominent is A1 (33, 34).

Selective enhancement of the neural response to a repeating sound might seem like an intuitive prediction, based on behavioral studies (1, 5). However, this enhancement may be surprising when viewed in the context of SSA (7, 35). If SSA affects responses to simultaneous stimuli, the same way as responses to sequential stimuli, one would expect a relative suppression of responses to the foreground stream. However, our results show the opposite effect, i.e., relative suppression of the simultaneous background. We propose that while SSA can account for the overall decreased response to both streams (6, 22), a separate mechanism must further suppress responses to sounds that occur simultaneously in the background. Furthermore, the fact that foreground enhancement is more prominent in secondary auditory cortical fields (PEG) than primary areas (A1), suggests a hierarchical process by which the enhancement emerges.

Feedback from local inhibitory networks and short-term synaptic plasticity are both believed to play a role in contextual modulation of sensory coding. Inhibitory interneurons contribute to SSA (36), contrast gain control (37), and top-down effects of behavior (38). Short-term plasticity may play a role in robust encoding of noisy natural sounds (39) and temporal integration of sound information (40). The most prominent effect of repetition in the current study may be broad inhibition of neural responses to both streams. However, disinhibition or adaptation of inhibitory feedback along specific pathways could mediate the relative enhancement of the foreground stream.

Animal models for streaming

Most behavioral studies of auditory streaming have been performed in humans (41–45). This is not surprising, as perceptual measurement of streaming in nonhuman species is challenging (46, 47). Within a small number of animal studies, however, the ferret has been identified as a model for streaming of alternating tone sequences and tone clouds (48, 49) and has been used to study its neurophysiological basis (15).

Here, we developed the ferret as a model for streaming complex, repeating sounds. Ferrets were able to report the occurrence of a repetition amidst random overlapping noise samples. Since the identity of the repeated sample was changed across behavioral blocks, we excluded the possibility that they used specific spectro-temporal features of the target sample to perform the task. While the ability to report the occurrence of repetitions in one stream is not direct proof that ferrets perceived two separate streams in the same way as humans, it confirms that they did perceive the repetitions.

Materials and Methods

All procedures were approved by the Oregon Health and Science University Institutional Animal Care and Use Committee and conform to United States Department of Agriculture standards.

Surgical procedure

Animal care and procedures were similar to those described previously for neurophysiological recordings from awake ferrets (Slee & David 2015). Five spayed, de-scented young adult ferrets (two females, three males) were obtained from an animal supplier (Marshall Farms, New York). Normal auditory thresholds were confirmed by measuring auditory brainstem responses to clicks. A sterile surgery was then performed under isoflurane anesthesia to mount a custom stainless-steel head-post for subsequent head fixation and expose a 10-mm2 portion of the skull over the auditory cortex where the craniotomy would be subsequently opened. A light-cured composite (Charisma, Heraeus Kulzer) anchored the post on the midline in the posterior region of the skull. The stability of the implant was also supported by 8–10 stainless self-tapping set screws mounted in the skull (Synthes). The whole implant was then built up to its final shape with layers of Charisma and acrylic pink cement (AM Systems).

During the first week post-surgery, the animal was treated prophylactically with broad-spectrum antibiotics (10 mg/kg Baytril). For the first two weeks the wound was cleaned with antisepsics (Betadine and Chlorhexidine) and bandaged daily. After the wound margin was healed, cleaning and bandaging occurred every 2–3 days through the life of the animal to minimize infection of the wound margin.

Experimental Design

Stimuli and Acoustics

Repeated embedded noise stimuli used in the present study were generated using the algorithm from McDermott et al. (2011). Brief, 250- or 300 ms duration samples of broadband Gaussian noise were filtered to have spectro-temporal correlations matched to a large library of natural sound textures and vocalizations but without common grouping cues, such as harmonic regularities and common onsets. The spectral range of the noise (125-16,000 Hz or 250-20,000 Hz) was chosen to span the tuning of the current recording site. Thus, the duration and spectral range of the stimuli differed from previous work in humans, but other statistical properties were identical (McDermott et al. 2011). An experimental trial consisted of continuous sequences of 10-12 noise samples (0 ms intersample interval) drawn randomly from a pool of twenty distinct samples (Figures 1B, 2A). The order of samples varied between trials. Either one stream of samples was presented (single stream trial) or two streams were overlaid and presented simultaneously (dual stream trial). At a random time (after 3–11 samples, median 6 samples), the sample in one stream (target sample) began to repeat. In dual stream trials, this repetition...
occurred only in one of the two streams, while samples in the other stream continued to be drawn randomly. In human studies, the repeating sample has been shown to pop out perceptually as a salient stream (McDermott et al. 2011). Thus, the stream containing the repeated sample is referred to here as the foreground, and the non-repeating stream as the background (Figure 1B). The period of the trial containing only random samples is referred to as the random segment, and the segment starting with the first repetition of the target sample is referred to as the repeating segment (Figures 1B, 2A). With the exception of the spectro-temporal receptive field analysis, the first sample of the random segment was excluded from all analyses to minimize the effect of onset-related adaptation, which was consistently complete by 250 ms following trial onset.

All behavioral and physiological experiments were conducted inside a custom double-walled sound-isolating chamber with inside dimensions of 8’ × 8’ × 6’ (L × W × H). A custom second wall was added to a single-walled factory chamber (Professional Model, Gretch-Ken Inc.) with a wooden frame and an inner wall composed of ¼” MDF board. The air gap between the outer and inner walls was 1.5”. The inside wall was lined with 3” sound absorbing foam (Pinta Acoustics). The chamber attenuated sounds above 2 kHz by more than 60 dB. Sounds from 0.2-2 kHz were attenuated 30-60 dB, falling off approximately linearly with log-frequency.

Stimulus presentation and behavioral control were provided by custom MATLAB software (Mathworks Inc.). Digitally generated sounds were D/A converted (100 kHz, National Instruments PCI-6229), and presented through a sound transducer (Manger W05) driven with a power amplifier (Crown D-75A). The speaker was placed one meter from the animal’s head, 30° contralateral to the cortical hemisphere under study. Sound level was calibrated using a ½” microphone (Bruel & Kjaer 4191). Stimuli were presented with 10ms cos2 onset and offset ramps.

Behavior

Two ferrets (one female, Ferret O, and one male, Ferret H) were trained to report the occurrence of repeated target noise samples in the repeated embedded noise stimuli using a go/no-go paradigm (David et al. 2012). Starting two weeks after the implant surgery, each ferret was gradually habituated to head fixation by a custom stereotaxic apparatus in a plexiglass tube. Habituation sessions initially lasted for 5 minutes and increased by increments of 5-10 minutes until the ferret lay comfortably for at least one hour. At this time the ferret was placed on a restricted water schedule and began behavioral training. During training and physiological recordings sessions that involved behavior, the ferret was kept in water restriction for five days/week, and almost all the daily water intake (40-80 ml) was delivered through behavior. Their diet was supplemented with 20 ml/day of high protein Ensure (Abbott). Water restriction was to be discontinued if weight dropped below 20% of the initial weight, but this did not happen with either ferret. Water rewards were delivered through a spout positioned close to the ferret’s nose. Delivery was controlled electronically with a solenoid valve. Each time the ferret licked the waterspout, it caused a beam formed by an infrared LED and photo-diode placed across the spout to be discontinued (Figure 1A). This system allowed us to precisely record the timing of each lick relative to stimulus presentation.

After trial onset, animals were required to refrain from licking until the onset of the repeating segment, i.e., after the occurrence of a repeated sample. Licks during the random segment were recorded as false alarms and punished with a 4-6 sec time-out. Licks that occurred in the repeating segment were recorded as hits and always rewarded with one to two drops of water (Figure 1B). Each behavioral block used two target samples, whose identity varied from block to block to prevent ferrets from learning to detect specific spectro-temporal features of the target. False alarm trials in which licks occurred before the target were repeated to prevent animals from simply responding with a short latency on every trial. The repeated trials were excluded from behavioral analysis (see below) to prevent artificial increase in performance from a strategy in which response time was gradually increased following each false alarm (David et al. 2012).

To shape the animal’s behavior, training started with a high signal-to-noise ratio (SNR) between random and repeating segments. SNR was then gradually decreased over subsequent training sessions to 0 dB (i.e., random and repeating segments were presented at the same intensity). Parameters such as spectral modulation depth of the two streams and length of the random segment/false alarm window were also adjusted over the training period.

Electrophysiology

Single- and multi-unit neural recordings were performed in one trained animal and four task-naive animals. A small (~1-2 mm diameter) craniotomy was opened over the right auditory cortex, in a location chosen based on stereotaxic coordinates and superficial landmarks on the skull marked during surgery. Initial recordings targeted primary regions of the ferret auditory cortex (A1), and recording location was confirmed by characteristic short-latency responses to tone stimuli and by tonotopic organization of frequency selectivity (Bizley et al. 2005). Recordings in secondary auditory cortex (PEG) were then performed in the field ventrolateral to A1. The border between A1 and PEG was identified functionally by a reversal in the tonotopic gradient.

Neurophysiological data were collected from animals in a passive state, i.e., while animals were head-fixed and unanesthetized but not performing any explicit behavior. Recording sessions typically lasted 2-4 hours. On each recording day, 1 to 4 high-impedance tungsten microelectrodes (FHC or A-M Systems, impedance 1-5 MQ) were slowly advanced into cortex with independent motorized microdrives (Alpha-Omega). The electrodes were positioned (Kopf Instruments) such that the angle was roughly perpendicular to the surface of the brain. Stimulus presentation and electrode advancement were controlled from outside the sound booth, and animals were monitored through a video camera. Neural signals were recorded using open-source data acquisition software (MANTA, Englitz et al. 2013). Raw traces were bandpass-filtered (0.3-10 kHz), amplified (10K, A-M Systems 1800 or 3600 AC amplifier), digitized (20 kHz, National Instruments PCI-6052E) and stored for subsequent offline analysis. Putative spikes were extracted from the continuous signal by collecting all events ≥ 4 standard deviations from zero. Different spike waves were separated from each other and from noise using principal component analysis and k-means clustering (David et al. 2009). Both single units (>95% isolation) and stable multunits (>70% isolation) were included in this study, resulting in a total of 149 A1 and 136 PEG units.

Between recording sessions, the exposed recording chamber surrounding the craniotomy was covered with polysiloxane impression material (GC America). After several electrophysiological penetrations (usually about 5-10), the craniotomy was expanded or a new craniotomy was opened to expose new regions of the auditory cortex. When possible, old craniotomies were covered with a layer of bone wax and allowed to heal.

Analysis

Behavior

Performance was assessed by a discrimination index (DI) computed from the area under the receiver operating characteristic (ROC) curve for detection of the target in the repeating segment (David et al. 2012; Yin et al. 2010). DI combines information about hit rate, false alarm rate, and reaction time. Higher values indicate better performance on the task.
\begin{align}
r_{\text{rand}}(S_f, S_b, i) &= \overline{r}(S_f, i) + \overline{r}(S_b, i) + r_0 \quad (\text{Eqn. 1}). \\
r_{\text{rep,ind}}(S_f, S_{bg}, i) &= [\overline{r}(S_f, i) + \overline{r}(S_b, i)] \exp(G_g) + r_0 \quad (\text{Eqn. 2}), \\
r_{\text{rep,dep}}(S_f, S_b, i) &= \overline{r}(S_f, i) \exp(G_f) + \overline{r}(S_b, i) \exp(G_g) + r_0 \quad (\text{Eqn. 3}).
\end{align}

Repeated trials following a false alarm were excluded from DI measurements to prevent artefactual inflation of DI estimates if animals used the strategy of gradually increasing their response time following each false alarm. Criterion was reached as the ferret performed at DI > 0.5, with 0 SNR and 0 modulation depth difference for four consecutive days. To determine statistical significance of each animal’s performance, we compared the actual DI to DI computed after shuffling response times across trials. We tested for a difference in mean actual and shuffled DI across blocks by a pairwise t-test.

Although the identity of the target samples varied from day to day, the ferrets could have developed a strategy in which, during the first few trials of a session, they learn to detect the spectro-temporal features associated with the target. Since some trials contained the target samples during the random segment, we predicted that ferrets using a spectro-temporal feature detection strategy would have a faster reaction time when the target was present during the random segment, whereas ferrets using a repetition detection strategy would have similar reaction times for both target and non-target samples. To test for this, we computed a reaction time for every noise sample in each trial. Reaction time was defined as the time from the onset of the noise sample to the first lick. We then aligned each trial relative to repetition onset, split the data by target vs. non-target, and computed the mean reaction time across trials (Figure 1D). Reaction time was fit using a mixed linear model based on sample identity, time until repetition, and the interaction between them. The majority of the experiments included only dual stream trials, but a subset contained interleaved dual and single stream trials (Ferret O = 224/635, Ferret H = 85/495). We found a small but significant difference in performance favoring dual stream trials in both animals (mean DI difference 0.038, \(p < 0.001\), pairwise t-test). Only the dual stream data were included in the behavior data reported in the Results.

Stream-dependent changes in sound-evoked activity

To assess the effect of repetition on overall responsiveness, we first measured changes in the response to the target sample between random and repeating trial segments. We computed the peristimulus time histogram (PSTH) response to each occurrence of target sample in the stimulus separately for the random segment and repeating segment, using data from dual-stream trials only. Spontaneous rate was subtracted from the PSTH to ensure the fraction term reflected changes in the evoked response. We then computed the observed gain change \((G_g)\) that minimized the least squares difference between evoked responses in the two segments. Log of the measured gain is reported to allow for direct comparison with the results of subsequent modeling analysis (see below).

PSTH-based models

Auditory cortical neurons could support segregation of the repeated stream either by changing the overall gain of their response to the repeating stream (stream-independent) or by differentially enhancing responses to one or the other stream (stream-dependent). To test these alternative predictions, we fit the data using stream-independent and stream-dependent models. In both models, responses were predicted using a weighted sum of time-varying responses to each noise sample. During the random segment, the time-varying response was the linear sum of a response to the foreground stream, response to the background stream, and spontaneous spike rate:

Here, \(S_f\) and \(S_b\) are the identity of samples in the foreground and background streams, respectively, and \(r_0\) is the spontaneous rate. \(\overline{r}(S, i)\) is the contribution of sample \(S\) to the evoked spike count in the \(i\)-th time bin following sample onset.

For the stream-independent model, responses during the repeated segment were computed, where a global gain term \((G_g)\) scales responses to both streams. For the stream-dependent model, responses during the repeated segment were computed, where the gains for the foreground \((G_f)\) and background \((G_g)\) modulate the respective stream responses separately before they are summed. The use of an exponent simplifies interpretation of gain changes such that values of \(G > 0\) indicate enhancement and values of \(G < 0\) indicate suppression. \(\overline{r}(S, i)\) can be negative, which allows for suppressed responses relative to the spontaneous rate. In this case, if a unit has both enhanced and suppressed responses, \(G\) will scale both responses equally (e.g., if \(G > 0\), there will be a decrease in spike rate during negative responses and an increase in spike rate during enhanced responses). The difference \(G_f - G_g\) is the relative enhancement between streams, here referred to as foreground enhancement. If \(G_f > G_g\), then the neural response to the foreground stream is enhanced relative to the background stream.

The repeating segment had many more presentations of the target sample than the random segment. To minimize potential bias when fitting the data, we randomly discarded target samples from the repeating segment such that the number of target samples in the repeating segment matched the number of target samples in the random segment. As mentioned earlier, the first sample of the random segment \((i.e., \text{the very first sample in the trial})\) was excluded from analysis to minimize the effect of on-set-adaptation in our analysis.

In this model, \(r(S, i)\) can contain negative values because it was added to the spontaneous rate, \(r_0\) (Figure 3). These negative values indicate that there was suppression of the spontaneous rate. The mechanisms producing this suppression are not specified \((e.g., \text{inhibition, adaptation})\), but similar suppression is captured by negative coefficients in an STRF. Models were fit to maximize Poisson likelihood of free parameters using Bayesian regression. A normal prior with mean 0 and standard deviation 10 was set on both \(r_0\) and \(\overline{r}\). A normal prior with mean 0 and standard deviation 1 was set on all \(G\) parameters. The model was fit three times using a different set of random starting values for each coefficient. Two thousand samples for each fit were acquired with a No-U-Turn Sampler, an extension to Hamiltonian Monte Carlo that eliminates the need to set \(r_0\) from \(G_f\). Units for which the 95% credible interval \((\text{CI})\) for \(E\) (as derived from the posterior) was less than 0 were considered to have significant suppression and those with a 95% CI greater than 0 were considered to have significant enhancement. For data shown in the Results, the mean of the relevant posterior is plotted.

To validate the stream-dependent model as a means of measuring stream-specific gain, we simulated the activity of 100 neurons with stream-dependent gain. For each neuron, we simulated the responses to 20 noise samples by generating a simulated PSTH for each sample from the sum of one to nine Gaussians (each amplitude, mean, and standard deviation...
randomized). Eighty trials for each neuron were simulated by assembling two sequences of noise samples, one for the foreground stream and one for the background stream, as in the repeated embedded noise task. The sequences were then scaled by randomly-chosen $G_f$ (uniform distribution, -0.5 to 1) and $G_b$ (uniform distribution, -1 to 1). The resulting PSTH for the trial provided a time-varying mean to a Poisson spike generator to produce simulated single-trial spiking responses. The regression model accurately recovered the responses to the individual samples ($r = 0.87$, $p < 0.0001$) and the baseline rate of each unit ($r = 1.00$, $p < 0.0001$). Predicted values for $G_f$, $G_b$ and $E$ were significantly correlated with the simulated values ($r = 0.65$, 0.92 and 0.82, respectively; $p < 0.0001$ for all). The model tended to underestimate foreground enhancement, suggesting that the results presented in this manuscript are a conservative estimate of foreground enhancement.

To validate results of the stream-dependent PSTH model applied to the data, measurements of prediction accuracy were obtained by 10-fold cross validation, in which a separate Bayesian model was fit to 90% of the data then used to predict the remaining 10%. This procedure was repeated 10 times with non-overlapping test sets, so that the final result was a prediction of the entire time-varying response. Prediction accuracy was then measured as the correlation coefficient (Pearson’s $r$) between the predicted and actual response. Mean prediction coefficients for A1 were 0.165 (± 0.01 SEM) for the stream-dependent and 0.161 (± 0.01 SEM) for the stream-independent model. Mean correlation coefficients for PEG were 0.125 (± 0.01 SEM) for the stream-dependent model and 0.119 (± 0.010 SEM) for the stream-independent model. For both A1 and PEG, the stream-dependent model had significantly higher correlation coefficients than the stream-independent model (A1: paired-sample $t$-test, $t$-statistic 3.27, $p = 0.0014$; PEG: paired-sample $t$-test, $t$-statistic 4.88, $p < 0.0001$).

**Lifetime sparseness and target preference**

We quantified sparseness ($S$), a measure of unit selectivity for a given sample relative to the others in the collection (8),

$$ S = \left[ 1 - \frac{\sum_{i=1}^{n} (r_i^2)}{\sum_{i=1}^{n} (r_i^2)} \right] \left[ 1 - \frac{1}{n} \right] $$

(Eqn. 4),

where $r_i$ is the standard deviation of the PSTH (computed using the average of response to the noise sample in the random segment of single-stream trials) for the $i^{th}$ sample and $n$ is the total number of noise samples. We also quantified target preference ($TP$), a measure of how well the target sample modulates the unit’s response,

$$ TP = \frac{r_{tar}}{\left( \sum_{i=1}^{n} (r_i^2) \right)} $$

(Eqn. 5),

where $r_{tar}$ is the standard deviation of the target PSTH, and the other terms are defined as for sparseness. The use of standard deviation to measure response magnitude means that strong suppression or enhancement yield similar response strength.

To assess whether there was a significant effect of sparseness, target preference, and/or area on foreground enhancement, $E$, we used a general linear mixed model:

$$ E = \beta_0 + \beta_1 A + \beta_2 S + \beta_3 \text{AS} + \beta_4 T + \beta_5 AT + \beta_6 ST + \beta_7 \text{AST} + U_1 $$

(Eqn. 6),

where sparseness ($S$), target preference ($T$), and area ($A$) are fixed effects. Since each unit was tested with two target samples and, therefore, provided two measures of foreground enhancement, unit ($U_i$) was included as a random effect for the $i^{th}$ unit.

Data from the trained animal represented 40% of all units in A1 and 46% in PEG. To test for potential effects of training on neural coding, we included training status (i.e., trained vs. naïve), along with all possible one, two, and three-way interactions to the general linear mixed model. Results suggested that the effects were weaker in the trained subset (i.e., there was no significant effect of foreground enhancement in A1). However, since all the “trained” data was from one animal and the statistical power was weaker overall, we felt there was insufficient data to draw conclusions regarding potential differences between units from naïve versus trained animals. Thus, we combined the data from naïve and trained animals for the results reported in this manuscript.

**Spectro-temporal receptive field models**

In addition to the PSTH-based models, which fit responses to individual noise samples, we confirmed that the same streaming effects were captured by a context-dependent spectro-temporal receptive field (STRF) model (51). The classic linear-nonlinear (LN) STRF models neural activity as the linear weighted sum of the preceding stimulus spectrogram, the output of which passes through a static nonlinearity to predict the time-varying spike rate response (52, 53). The STRF, $h(x, u)$, is as a linear weight matrix that is convolved with the logarithm of the stimulus spectrogram, $s(x, t)$:

$$ r_{LIN}(t) = \sum_{x=1}^{X} \sum_{u=1}^{U} h(x, u) s(x, t - u) $$

(Eqn. 7),

where $x = 1...X$ are the frequency channels, $t = 1...T$ is time, and $u$ is the time lag of the convolution kernel. Taking the log of the stimulus spectrogram accounts for nonlinear gain in the cochlea. Free parameters in the weight matrix, $h$, indicate the gain applied to frequency channel $x$ at time lag $u$ to produce the predicted response. Positive values indicate components of the stimulus correlated with increased firing, and negative values indicate components correlated with decreased firing.

The output of the linear STRF is passed through a static nonlinear sigmoid function to account for spike threshold and saturation (10),

$$ r(t) = F[r_{LIN}(t)] $$

(Eqn. 8),

where

$$ F(x) = r_0 + A \exp \left[ -\exp (\kappa (x - x_0)) \right] $$

(Eqn. 9).

Free parameters of the static nonlinearity are $x_0$, inflection point of the sigmoid; $r_0$, spontaneous spike rate; $A$, maximum spike rate; and $\kappa$, the slope of the sigmoid.

We developed a modified LN STRF to account for stream-dependent changes in gain. The input spectrogram for each stream was scaled by a gain term that depended on stream identity (foreground or background) and trial segment (random or repeating). We refer to this model as the stream-dependent STRF model. The stimulus was modeled as the sum of two log spectrograms, computed separately for the foreground and background streams, $s_1$ and $s_2$, respectively. In the random segment, the total stimulus, $s(x, t)$, was modeled as the linear sum of these two stimuli:

$$ s(x, t) = s_1(x, t) + s_2(x, t) $$

(Eqn. 10).

In the repeating segment, each stimulus was scaled by the gain, $G$, for the respective stream,

$$ s(x, t) = G_1 s_1(x, t) + G_2 s_2(x, t) $$

(Eqn. 11).

All model parameters were estimated by gradient descent (10, 51, 54). STRF parameters were initialized to have flat tuning (i.e., uniform initial values of $h$) and were iteratively updated using small steps in the direction that optimally reduced the mean squared error between the time-varying
spike rate of the unit and the model prediction. To maximize statistical power with the available data, the STRF was fit using both single- and dual-stream data. For single-stream trials, the second stimulus spectrogram was fixed at zero, \( s_2(x,t) = 0 \), and a separate gain term was fit for those trials to prevent bias in estimates of \( G_r \) and \( G_s \). Gain parameters and STRF parameters were fit simultaneously (51). Measurements of prediction accuracy were obtained by 20-fold cross validation, in which a separate model was fit to 95% of the data and then used to predict the remaining 5%. Fit and test data were taken from interleaved trials. This procedure was repeated 20 times with non-overlapping test sets, so that the final result was a prediction of the entire time-varying response. Prediction accuracy was then measured as the correlation coefficient (Pearson’s \( r \)) between the predicted and actual response. Standard error on prediction correlation was measured by jackknifing (55), and only units with prediction error significantly greater than zero were included in model comparisons (\( p < 0.05 \); jackknife \( t \)-test).

To quantify effects of segment- and stream-dependent gain, we also fit models using the same data and fitting procedure, but where stream identity (stream-independent STRF model) or both segment and stream (baseline STRF model) were shuffled in time. An improvement in prediction accuracy for a model with a non-shuffled over shuffled variable indicated a beneficial effect of the corresponding gain parameter on model performance, and thus of a stream-dependent change in sound encoding. Significant differences in model performance were assessed by a Wilcoxon rank sum test between prediction correlations for the set of units fit with each model.

Statistical analysis

As described above, the effect of repetition on neural responses to the simultaneous streams was quantified using a Bayesian regression model. Unlike conventional (i.e., frequentist) approaches, Bayesian analysis does not generate standard \( p \)-values. Instead, Bayesian analysis quantifies the probability that the true value for a parameter falls between two points. These distributions can be used to calculate the probability that there is a true difference between groups, which is typically the information that is intended to be reflected in \( p \)-values (56). In our analyses, we report the mean and 95% confidence interval (CI) for all the gain terms. The CI should be interpreted as the interval in which we are 95% certain the true value is contained. Therefore, if the 95% CI does not include the reference value (e.g., \( E = 0 \), which indicates no significant change in response from the random segment), we treat it as significant. Where possible, we have translated these determinations into more standard \( p \)-value.

To quantify the relationship between neural tuning and the effects of repetition on stream segregation, we used a linear mixed effects model fit with the Python statsmodels toolbox (57), using algorithms as described by Lindstrom and Bates (58). Area (A1 vs. PEG), target preference and sparseness were treated as fixed effects. All two- and three-way interactions between the effects were included. Random intercepts for each unit were incorporated into the model. The effect size, \( z \)-score, \( p \)-value and 95% confidence interval are reported for all fixed effects and their interactions in Table 1.

In the STRF model analysis, to assess statistical significance of the prediction correlation and thereby determine which units to include in the model comparisons, we used the jackknife \( t \)-test (55). Standard error was computed by jackknife resampling of the predicted and actual response. Significant units had a prediction correlation at least two standard errors greater than zero (i.e., \( p < 0.05 \)). Significant differences in performance by two models on a single neuron were also determined by a jackknife \( t \)-test. In this case, the prediction correlation for the two models had to be separated by at least two standard errors.

To test for differences between two neural populations (e.g., \( G \) in A1 vs. PEG), we used a Wilcoxon rank-sum test, Student’s \( t \)-test or independent two-sample \( t \)-test. The specific test used, along with the test statistic, degrees of freedom and \( p \)-value are reported alongside the relevant result.

Acknowledgments

The authors thank Dr. Josh H. McDermott for providing the code to generate the noise samples, and Zachary Schwartz and Henry Coney for assistance with behavioral training and neurophysiological recordings. This study was supported by grants from the NIH (S.V.D., R00 DC 010439, R01 DC 014950; B.N.B., R21 DC 016969; D.S., F31 DC014888) and Tartar Trust (D.S.).

References


Englitz B, David SV V, Sorenson MDD, Shamma SA. 2013. MANTA – An open-source, high density electrophysiology recording suite for MATLAB. Front. Neural Circuits. 7:69


Nääätänen R. 2001. The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). Psychophysiology. 38(1):1–21


Roberts B, Glasberg BR, Moore BCJ. 2002. Primitive stream segregation of tone sequences without differences in fundamental frequency or passband. J. Acoust. Soc. Am. 112(5 Pt 1);2074–85

Seabold S, Perktold J. 2010. Statsmodels: Econometric and Statistical Modeling with Python


