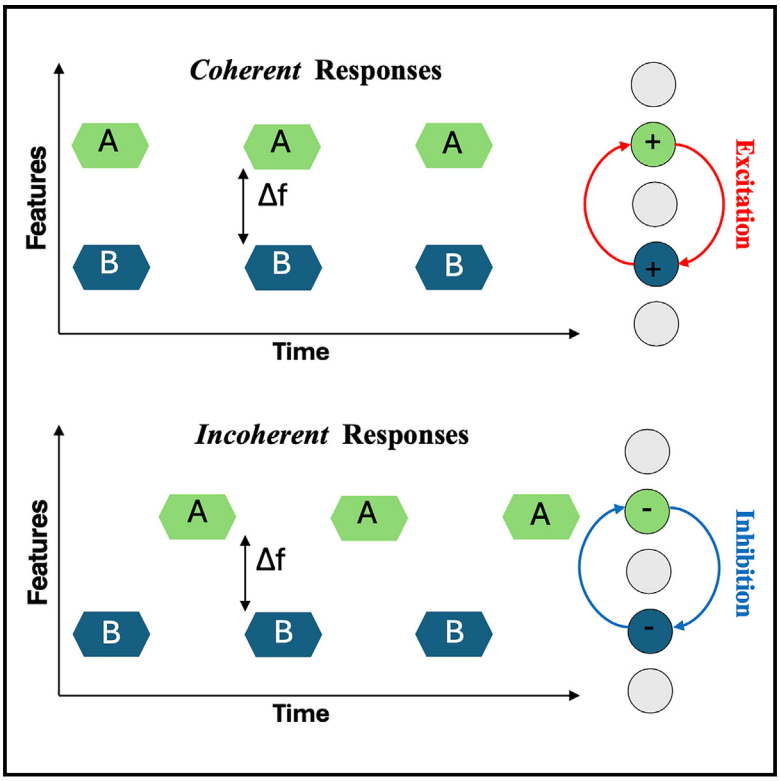
iScience

Article

Temporal-coherence induces binding of responses to sound sequences in ferret auditory cortex

Graphical abstract



Authors

Kai Lu, Kelsey Dutta, Ali Mohammed, Mounya Elhilali, Shihab Shamma

Correspondence

sas@umd.edu

In brief

Behavioral neuroscience; Sensory neuroscience

Highlights

- Binding of a source's features facilitates its perception
- Temporal coherence of feature-driven neuronal responses rapidly induces binding among them
- Temporally incoherent responses exhibit mutual competitive suppression
- Binding effects of enhancement and suppression are stronger with more coherent responses

Lu et al., 2025, iScience *28*, 111991
March 21, 2025
https://doi.org/10.1016/j.isci.2025.111991

CellPress

iScience

CellPress
OPEN ACCESS

Article

Temporal-coherence induces binding of responses to sound sequences in ferret auditory cortex

Kai Lu,[1] Kelsey Dutta,[2] Ali Mohammed,[2] Mounya Elhilali,[3] and Shihab Shamma[2,4,5,*]
[1]Department of Biology, Emory University, Atlanta, GA, USA
[2]Electrical and Computer Engineering Department & Institute for Systems Research, University of Maryland, College Park, MD, USA
[3]Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD, USA
[4]Déparment d'Études cognitives, L'École normale supérieure-PSL, Paris, France
[5]Lead contact
*Correspondence: sas@umd.edu
https://doi.org/10.1016/j.isci.2025.111991

SUMMARY

Binding the attributes of a sensory source is necessary to perceive it as a unified entity within its surrounding scene as in the cocktail party problem in auditory perception. It is postulated that coherent temporal modulation of a source's features binds them and enhances their perception. This study seeks evidence for rapid binding among coherently responsive single-neurons in ferret auditory cortex. In one experiment, ferrets attended to a sequence of noise bursts while we contrasted responses to simultaneous synchronized versus alternating tone sequences. We found that the contrast between synchronized (enhanced) and desynchronized (suppressed) responses rapidly increased, thus promoting their segregation. In another experiment, a sequence of an irregularly repeated multi-tone complex was embedded in a background of randomly dispersed tones. Single-unit and functional ultrasound imaging of responses to the temporally coherent tones of the complex became rapidly enhanced against the background responses, demonstrating the role of temporal-coherence in binding and segregation.

INTRODUCTION

Humans and other animals often perceive and manage auditory signals emanating from many simultaneously active sources in cluttered environments. The acoustic signals arrive to the ears as mixtures to be segregated, tracked, and recognized. This feat is achieved via multiple complex cognitive functions and neural processes including attention, memory, and rapid plasticity.[1,2] But a key role is played by a simpler process referred to as the *principle of temporal coherence*.[1,3] It postulates that a single source primarily evokes persistently synchronized (or coherent) responses representing its various attributes (e.g., pitch, location, timbre). Furthermore, activations due to other independent sources are typically mutually incoherent (e.g., two simultaneous speakers uttering different words). Consequently, binding the coherent attributes of a target source would segregate (extract) it from other incoherent sources. Numerous psychoacoustic tests, EEG/MEG and single-unit auditory cortical responses[3–6] have over the last decade confirmed many aspects and surprising predictions of this hypothesis in scene analysis. A review of these studies with detailed commentary can be found in the study by S. Shihab and M. Elhilali.[1]

The experiments described here build upon past investigations of the neural basis of perception of simple tone sequences as ''streams'', or stream formation.[2] For example, one specific study focused on auditory cortical responses as animals attended to synchronized vs. alternating tone sequences,[7] as depicted in Figure 1A. The rationale is that the evoked responses mimic those due to more complex source mixtures, with neurons tuned to these tones becoming coherently (top panel) or incoherently driven (lower panel). Such activations are postulated to induce rapid plasticity of the mutual connectivity among the responsive neurons, with coherent neurons forming mutually excitatory connections (depicted in red) that enhance both of their responses and gradually bind them together. The opposite is hypothesized to occur when neurons are driven incoherently (bottom panel), where mutually inhibitory connectivity forms suppressing both neurons' responses. The results in the study by Lu K. et al.[7] confirmed these postulates by demonstrating that during attentive listening, responses rapidly evolved (enhanced and suppressed) with dynamics of the order of 100 s of milliseconds.

Experiment I adds a fundamental twist with an auditory scene that is created with *multiple* sequences of different tokens (e.g., tones and noise) that could be perceptually organized in different ways depending on how attention is selectively deployed. Figure 1B depicts such a scenario where two alternating tone sequences (A and B) are presented with a 3rd sequence of high frequency noise bursts (**N**) which can be synchronous with either of the tone sequences (top and bottom panels of Figure 1B). The key rationale for the design of this experiment is that when the

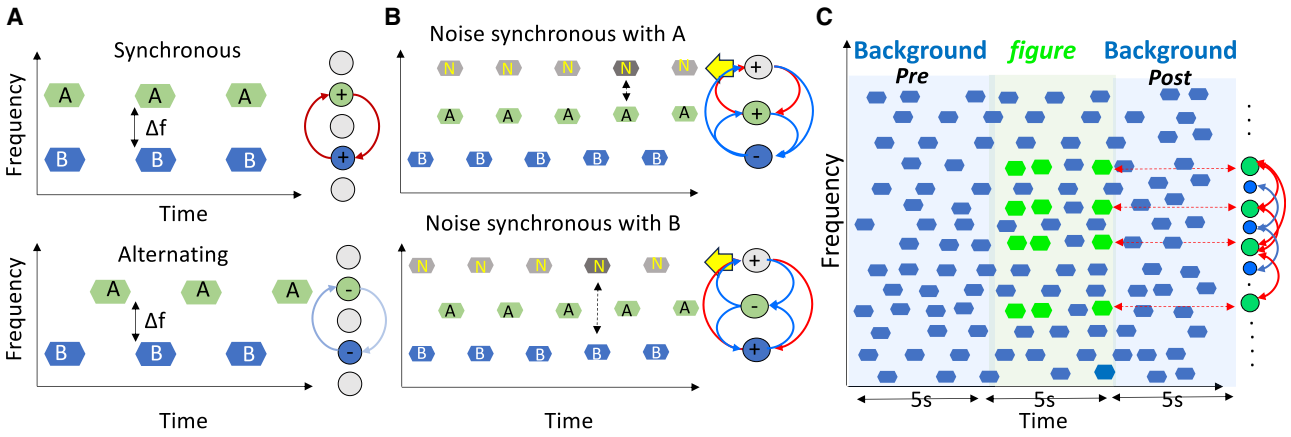


Figure 1. Temporal coherence and the binding hypothesis

(A) Binding in the classical two-tone streaming paradigm. Synchronous tones (A and B in top panel) form a single perceptual stream and induce coherently driven neurons to form mutually excitatory connections (red arrows) and enhanced responses. When asynchronous (bottom panel), the tones segregate into two streams and are hypothesized to induce mutually inhibitory connections (blue arrows) that mutually suppress the responses.

(B) Binding through selective attention. Attending selectively to a sequence of noise bursts (yellow arrows) makes it serve as an anchor that binds it to other synchronous sequences, while asynchronous sequences become suppressed. The role of sequences A and B can be interchanged to monitor the effects on the responses of the cells tuned to the two tones.

(C) The stochastic figure-background stimulus consists of a random tone-cloud (blue tones) with an intermediate epoch (*figure*) during which a sequence of several synchronized tones (4, 6, 8, or 10) are introduced interspersed irregularly with the random tones at a rate of ~4/s. It is postulated that the coherent *figure* tones become bound (red arrows), their responses mutually enhanced, and then pop-out perceptually.

ferret selectively attends to the noise burst sequence, we can examine the binding to the two other sequences without need to attend to either of tone sequences, i.e., N-A (*top panel*) or N-B (*bottom panel*). That simplifies significantly the behavioral task of the ferret in that it only requires it to attend to the noise. Of course, we can only assume based on previous related behavioral tasks on streaming with ferrets[8] that the alternating tone sequence will stream apart, while the synchronous one remains with the noise (see the study by Bregman A.S.,[2] p.29). How might this occur physiologically is depicted schematically by the pattern of dynamic connectivity postulated to form following the same process discussed in Figure 1A. To test this hypothesis, experiment I assessed the evolution of *binding* as evidenced by the changing responses to the tone-sequences when the animal attends in one condition versus another (*top* vs. *bottom panels*). It is specifically predicted that (1) in the top-panel condition, **A**-tone responses would *increase* relative to the B-tone responses; the opposite would occur in the bottom panel. Another prediction is that (2) evolution of the response changes (or presumably the connectivity) is directed or controlled by the attentional focus in favor of the tone sequence synchronous with the noise. Consequently, in the absence of attention (passive listening), this binding process and its concomitant suppression or enhancement are predicted to weaken or disappear.

In experiment II, the number of coherent tones and the complexity of the overall auditory scene is increased significantly (Figure 1C), while remaining relatively accessible and interpretable within the same hypothetical scenarios of Figures 1A and 1B. Here, a random cloud of incoherent tones (also referred to as the "Background") is heard typically for 5 s, followed by an additional embedded sequence of coherent tone complexes (referred to as the *figure*) that commences abruptly, and then repeats *irregularly at* about 4 per second for a total of 5 s. The tone-cloud continues for a final 5 s (post-*figure*) interval (Figure 1C). A more realistic (less schematic) rendition of a trial is depicted in Figure S1. It is hypothesized that the coherent tones of the *figure* will induce excitatory connectivity among the coherently driven neurons, binding them together and enhancing their responses, and hence making the *figure* perceptually pop-out becoming more salient after a short period of buildup. This type of stimulus (often referred to as the *stochastic figure ground [SFG]*) has already been fruitfully investigated with human subjects in numerous psychoacoustic and MEG/EEG studies.[9–11] But no single-unit recordings of the underlying responses in an animal model have been reported.

The results reported here from the two experiments of Figures 1B and 1C are broadly consistent with the idea of a binding process in which coherent neural responses become rapidly enhanced relative to the suppression of incoherent responses. In experiment I, the effect of selective attention reshapes the organization of the responses to the two-tone sequences (Figure 1B). In experiment II, responses in a passively listening animal exhibit the consequences of binding on the coherent tones of the *figure* (Figure 1C), in agreement with findings from EEG human recordings passively listening to similar kinds of stimuli.[11]

Finally, in addition to single-unit recordings, we employed functional ultrasound (fUS) imaging of the cortical responses to view large-scale spatiotemporal dynamics of brain activity during coherent activation and thus gain a multi-scale perspective of the binding process. As we shall discuss, the broad view afforded by this technique, especially in cortical depth reveals that the responses are not limited to a few driven cells but rather that it is of a global extent across the auditory cortex.

RESULTS

A total of 5 ferrets provided data in the various experiments. Two ferrets (U and R) participated in experiment I, and three (ferrets B, K, and U) in experiment II. Single-unit recordings were made in the primary auditory cortex (A1) unless stated otherwise. *fUS* recordings spanned a cortical region encompassing the medial and posterior ectosylvian gyri in ferrets U and Z (primary, anterior, and secondary [PEG] auditory cortical fields).

Experiment I: Binding and selective attention

Ferrets U and R listened to simultaneous streams of sound sequences (Figure 1B) consisting of a high-frequency band of noise bursts, and two lower frequency tone sequences (A and B). The animals were trained to attend *only* to the N and to detect a small intensity change (denoted as a darker burst in Figure 1B) to receive a water reward. The center frequency of the noise burst was selected within a 5-octave range of the tones. Unbeknownst to the animals, two tone conditions were tested where either of the A or B-sequences were synchronous with the noise bursts (*upper* and *lower panels* of Figure 1B). Animals generally performed the task identically in the two conditions as exemplified by the matched performance of ferret R during the two sets of trials (Figure S2). Single-unit responses were recorded in ferrets R (76 neurons) and U (40 neurons) in both *passive* and *active* conditions to compare the effects of engagement and attention. In each recording, we first measured the best frequencies (BF's) of all simultaneously isolated neurons (on 4 independent electrodes). Then the frequency of one of the two tone-sequences was selected near the BF of one neuron, while the frequency of the other tone was selected 3/4 octave away and nearer to one of the other isolated neurons. This way, at least some neurons preferentially responded to either tone A or B in each recording.

Two types of trials were used to stimulate each neuron (e.g., the green neuron in Figure 1B whose BF is aligned with tone A sequence): (1) *coherent (or synchronized—**SYN**) trials* in which the neuron is driven by tone A synchronously with the N; (2) *incoherent (or asynchronized—**ASYN**) trials* in which the BF-tone of a cell (tone A) is asynchronous with the noise, while the other tone (tone B) drives the blue neuron synchronously with the noise bursts. We recorded responses of *each* neuron to *both* stimulus conditions while the animals passively listened and when they were actively engaged in the detection task. The neural responses from each pair of (A and B) tones are *averaged* over all their repetitions in each condition, and over the duration of all trials. The two post-stimulus histograms (PSTHs), thus generated from each neuron are labeled according to the trial type as: SYN_{pass}, SYN_{act}, $ASYN_{pass}$, and $ASYN_{act}$. A trial typically lasted a minimum of 1.5 s (4–20 repetitions of the A and B tone pairs) to allow for the postulated neuronal connectivity to adapt during the trial and become evidenced by the changes (enhancements or suppression) in the single-unit responses. Note that since a trial-type was chosen randomly, the changes observed are assumed to build up from scratch at every trial.

Evidence of binding in the patterns of rapid plasticity

Neuronal responses from a given cell depended on many factors, and hence were quite variable. They included the exact frequency of the A and B tones relative to the BF, and the excitatory and inhibitory fields of the cell's spectrotemporal receptive fields (STRFs) and whether they are thus affected by the noise bursts even if they were far away. Another important factor is the behavioral state of the animal (passive or active) because it has been commonly found that when animals engage in a task and attend to its stimuli, the overall responsiveness of primary auditory cortical cells often decreases significantly.[12–14] Therefore, to demonstrate the binding hypothesis based on such responses, it is critical to consider the *relative* changes of a cell's responses between the SYN and ASYN conditions and not simply whether it is enhanced or suppressed in absolute terms as we illustrate next.

Figure 2A panels display PSTHs from two neurons in ferret R. Each panel depicts the PSTH responses in the SYN (left half = 320 ms) and ASYN (right half = 320 ms) conditions as illustrated by the tone and noise burst symbols above the panels. The two cells' PSTHs are quite different yet both are consistent with the binding hypothesis. *Unit-1* (*top panel*; Figure 2A) exhibits responses that follow closely the postulate described in Figure 1B—with attention during the task, SYN responses (left side) increase while the ASYN responses decrease relative to the passive state. We quantify these changes as Δ^{ASYN} and Δ^{SYN} in Equation 1:

$$\Delta^{ASYN} = ASYN_{pass} - ASYN_{act} = \text{passive-to-active change in ASYN condition}$$

$$\left(\Delta^{ASYN} > 0 \text{ if cell is } \textit{suppressed} \text{ in active}\right)$$

$$\Delta^{SYN} = SYN_{pass} - SYN_{act} = \text{passive-to-active change in SYN condition}$$

$$\left(\Delta^{SYN} < 0 \text{ if cell is } \textit{enhanced} \text{ in active}\right) \qquad \text{(Equation 1)}$$

where SYN_{pass}, SYN_{act}, $ASYN_{pass}$, and $ASYN_{act}$ denote the *mean of PSTH* responses over the 320 ms interval comprising the two tones depicted in each panel of Figure 2A. Note that according to Equation 1, $\Delta^{ASYN} > \Delta^{SYN}$ in this neuron because $ASYN_{pass} > ASYN_{act}$ ($\Delta^{\mathbf{ASYN}} > 0$) and $SYN_{pass} < SYN_{act}$ ($\Delta^{SYN} < 0$).

Because of the depressive effects of task engagement and other factors alluded to earlier, it is common in cells' responses for the inequality $\Delta^{ASYN} > \Delta^{SYN}$ to hold even when both conditions cause a suppression of the responses as in *unit-2* (*lower panel*; Figure 2A). This cell showed deeper suppression during ASYN compared to SYN, hence $\Delta^{ASYN} > \Delta^{SYN}$, which is still in line with the binding hypothesis.

Figure 2B summarizes the response changes over all neurons in both ferrets (U and R). The two panels contrast the response modulations between passive vs. active states, separately under SYN (*left panel*) and ASYN (*right panel*) conditions. A significant suppressive change occurs in the ASYN condition, as indicated by the negative *effect size* and *p* value (see Methods) as the animal attends to the noise stream (Figure 1B). In the SYN condition, many cells are no longer suppressed, but there is no significant positive bias either. This suggests that the binding effects

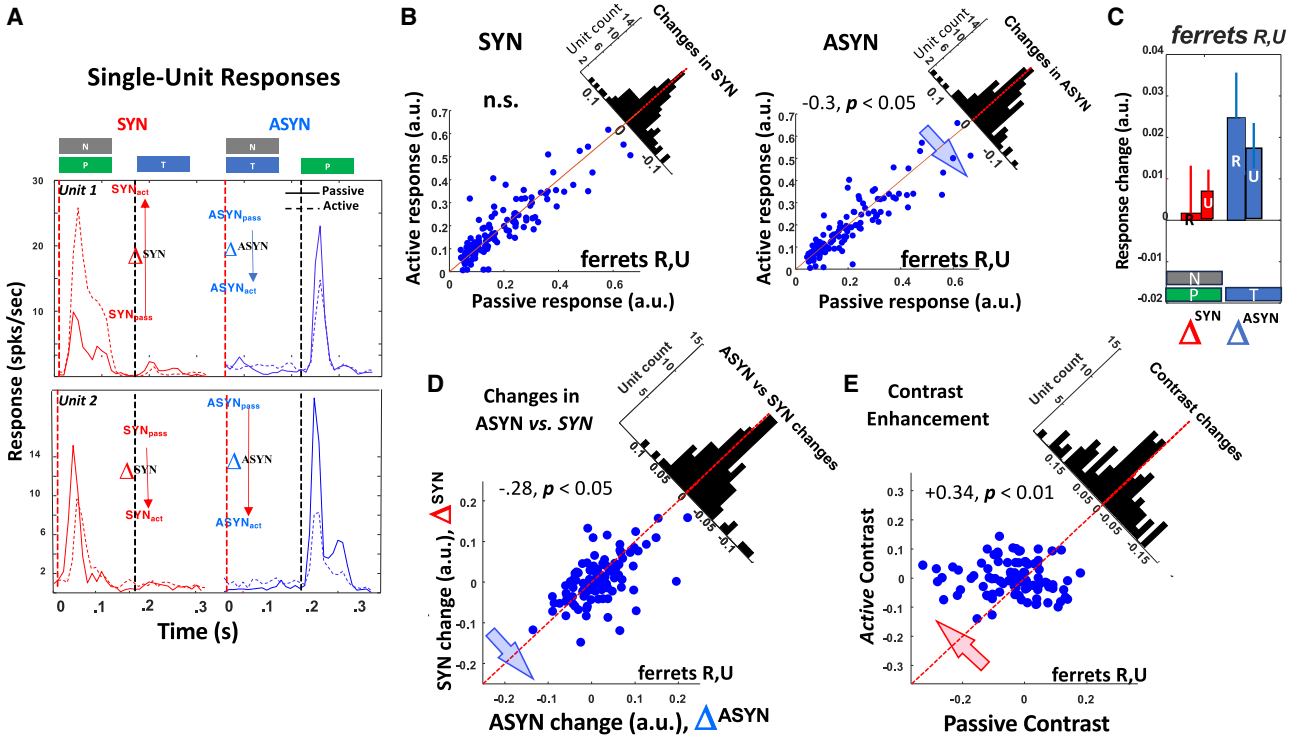


Figure 2. Response modulation reflects binding of different tone streams

(A) Enhancement and suppression in single cell examples during passive and active task engagement in ferret R. Each of the *top and bottom* panels depict the PSTHs of the SYN (*left-half*) and ASYN (*right-half*) trials of two units. The vertical dashed lines are the onsets of the SYN tones (in red) and the ASYN tones (in black). The stimuli are indicated by the colored symbols on top. The noise burst (N) sequence is synchronous with either of the two tone-sequences—the preferred tone (P) usually chosen near BF, or with the non-preferred tone T. *Unit-1* PSTH responses (*top panel*) change during the active state (dashed-curves) relative to the passive state (solid-curves). The *inset* arrows display the changes in the peak responses in the four different conditions (SYN_{act}, SYN_{pass}, $ASYN_{act}$, and $ASYN_{pass}$). This cell's responses are *enhanced* (Δ^{SYN} depicted by red arrow) when driven synchronously with N, and suppressed when driven asynchronously (blue arrow, Δ^{ASYN}). Both Δ^{SYN} and Δ^{ASYN} are defined as the *change* from *passive-to- active* (Equation 1 *in text).* Therefore, for *unit-1* $\Delta^{SYN} < 0$ and $\Delta^{ASYN} > 0$, or $\Delta^{ASYN} > \Delta^{SYN}$. In *unit-2* (*bottom panel*) both SYN and ASYN responses are suppressed in the active state, but the suppression is stronger in ASYN, and hence the inequality remains the same: $\Delta^{ASYN} > \Delta^{SYN}$. This inequality is taken to be consistent with the effects of binding.

(B) Scatterplots of response modulations in SYN and ASYN conditions in both ferrets R and U. (*Left panel*) The scatterplot of the total average responses (over 320 ms) of each unit in SYN trials during the active (y axis) versus the passive condition (x axis). (*Right panel*) The scatterplot of the same cells during the ASYN trials. The histograms summarize the scatter around the midline, and its effect-size is indicated if significant (see text and Methods). Only during the ASYN trials is there a significant bias of responses *below* the midline (suppression) (effect size = −0.3, $p = 0.023$).

(C) Suppression is strongest during the ASYN trials in both ferrets, as indicated by the bar-plots of the means and standard errors of spike rate differences, Δ^{ASYN} and Δ^{SYN} in both trial types.

(D) Evidence of binding in scatterplots of response modulations from all cells in ferrets R and U. Scatterplot of response changes during ASYN (Δ^{ASYN}) versus SYN (Δ^{SYN}) indicates a significant bias below the midline (effect size = −0.28, $p = 0.036$) implying an overall $\Delta^{ASYN} > \Delta^{SYN}$, taken to be as evidence of binding.

(E) Contrast enhancement due to binding. *Normalized* differences between SYN and ASYN (or the contrast; defined by Equation 2) increases during the active state, as indicated by the significant bias of points above the midline (effect size = +0.34, $p = 0.008$).

here are rooted more in the *suppression* of unattended ASYN responses or in the *relative* change between responses in the passive and active conditions. The net response changes in all cells from both ferrets (R and U) are integrated and depicted by the two bar plots in Figure 2C which show a *strong* net suppression during ASYN and a *weak* net suppression in SYN. All aforementioned findings were replicated individually in each of ferrets U and R (Figures S3A–S3B).

Figure 2D directly compares the response changes (Δ^{ASYN} vs. Δ^{SYN}) in each cell due to engagement. The negative bias confirms that the inequality $\Delta^{ASYN} > \Delta^{SYN}$ predominates, providing evidence of binding induced changes. However, despite the changes in SYN (Δ^{SYN}) being relatively small and unbiased (Figures 2B and 2C), they apparently still contribute to the overall plasticity patterns reflecting binding. This is demonstrated by shuffling the list of Δ^{SYN} associated with each of the cells, hence scrambling the pairing of Δ^{ASYN} with Δ^{SYN} in each cell and then recomputing the scatterplots as in Figure 2D. The Δ^{SYN} shuffling disrupts the condition $\Delta^{ASYN} > \Delta^{SYN}$ in many cells and eliminates the earlier negative bias (Figure S3C).

Evidence of binding through contrast enhancement

Another important consequence of binding's competitive and cooperative interactions (Figure 1B) is to *increase the contrast* between the responses of the attended coherent cells *relative* to the incoherent ones. To test if this change occurs in our

data, we defined for each cell the average contrast between the SYN and ASYN responses in passive and active conditions as:

$$\text{Contrast in passive} : C_{pass} = (SYN_{pass} - ASYN_{pass}) / (SYN_{pass} + ASYN_{pass})$$

$$\text{Contrast in active} : C_{act} = (SYN_{act} - ASYN_{act}) / (SYN_{act} + ASYN_{act}) \quad \text{(Equation 2)}$$

These normalized contrast measures are compared in the scatterplot of Figure 2E, which displays a significant *positive* bias with $C_{act} > C_{pass}$ (+0.34, $p < 0.01$) implying that task engagement increases the contrast of the coherent *relative to* the incoherent responses. Finally, to demonstrate the importance of Δ^{SYN} to contrast enhancement, we find that removing their contributions by setting $\Delta^{SYN} = 0$ ($SYN_{act} = SYN_{pass}$) abolishes the positive shift of Figure 2E, as demonstrated in Figure S3D.

In summary, the results thus far reveal that selective attention to a sequence induces response changes consistent with the binding hypothesis ($\Delta^{ASYN} > \Delta^{SYN}$) but which take the form of larger suppression Δ^{ASYN} relative to Δ^{SYN} effects. Nevertheless, manipulating (shuffling or nulling) the Δ^{SYN} has significant effects on the change distributions and contrast indices reflecting their importance. Furthermore, as shown in Figure S4, the role of attention is critical because in the passive state there are only minimal biases in favor of SYN or ASYN responses.

Experiment II: The SFG stimulus

This experiment employs the SFG stimulus detailed in Figures 1C and S1. It is a versatile stimulus that allows for multiple simultaneous measurements including the neurons' STRFs, and the dynamics of their response changes as the *figure* exerts its strong influence by inducing binding among the coherently driven neurons. The stimulus consists of three epochs: (1) pre: the *initial* epoch of random tones (tone-cloud) that facilitates STRF estimation (referred to as *pre*-STRF); (2) mid: an *intermediate* period that contains in addition to the tone-cloud several irregularly repeated bursts of temporally coherent tones—the *figure*—at about 4 bursts per second. It is postulated that during this epoch binding evolves among the coherently driven neurons; finally (3) post: another epoch of tone-cloud that allows for a *post*-STRF measurement and assessment of the persistent effects of the sustained presentation of the synchronized *figure* tones during the intermediate epoch.

Note that the timing of the (random and coherent) tones throughout this stimulus was chosen carefully to ensure that when added together, the overall stimulus envelope did not exhibit large fluctuations related to the synchronous tones of the *figure*. Specifically, as elaborated upon in Methods, the tones of the tone-cloud epochs were overlapping and had roughly the same overall power throughout the *pre-* and *post-figure* epochs. Importantly, during the *figure*, the coherent tones were created by adjusting the timings of nearby random tones to keep the overall number at any given moment roughly constant. Furthermore, we also ensured that the synchronous onsets of the *figure* tones were balanced by removing any random tones that have nearby onsets. Three ferrets participated in the single-unit recordings of this experiment. Details of the numbers of isolated single-units, shared ferret participation across different experiments, and their ages are available in Methods.

STRF measurements during the tone-cloud epochs

The tone-cloud stimulus is akin to a noise that can be used to measure neural STRFs and hence the tuning and expected dynamics of the cells.[15–17] We first validated that this stimulus could recreate meaningful STRFs that resemble those due to our standard battery of measurements in A1, specifically using *temporally orthogonal ripple combinations* (TORC) stimuli.[16] With lag values of −10 to 120 ms, the cross-covariance between the response and the auditory spectrogram of the stimulus was computed and normalized by the auto-covariance of the stimulus. Ridge regression was performed with 37 log-spaced values for each trial and then averaged to compute the *pre*-STRF and *post*-STRF for an interval of 1 or 2 s just before and after the *figure* epoch. During the *figure* intermediate epoch, we computed three additional similarly windowed STRFs during the *early*, *mid*, and *late* portions of this epoch. While computed the same way as the *pre*-STRFs and *post*-STRFs, these measurements are different in that the stimulus auto-correlation is not white when the *figure* is included. The resulting artifacts due to the *figure* are quite evident, but the "STRF" estimates nevertheless are revealing and useful as explained next.

Figure 3A compares the STRFs measured with the tone-cloud versus TORC responses (ferret B). The two differ since many factors influence such estimate[15,16] including that TORCs contain a dense range of frequencies spanning 5 octaves which evolve over time in a continuous manner. The tone-cloud by contrast contains only a discrete set of 37 frequencies and are presented as discrete tone sequences spanning 3 octaves to 6 octaves (as indicated in specific animals and tests). This explains the striated appearance of the excitatory regions of the bottom-rightmost panel of Figure 3A. Despite these large differences between stimuli, the two measurements display approximately matched latencies, BFs, and arrangements of the major excitatory or inhibitory regions around the BF. The tone-cloud STRFs will be useful in monitoring how binding evolves over the different epochs and revealing the effects of the *figure* presentation during the intermediate epoch.

STRFs during *figure* presentations

Figure 3B illustrates a series of 5 STRFs computed from the responses to *unit 2* (Figure 3A): a *pre-STRF*, 3 *intermediate-epoch (early-, mid-, late-STRF),* and a *post*-STRF. Enlarged sections of some of these panels are shown for added clarity in Figure S5. In each trial, one of four different *figures* with varying sizes was randomly selected and presented, e.g., 4 tones (*top row*), 6 tones (*not shown*), 8 tones (*middle row*), and 10 tones (*bottom row*). The frequencies of the *figure* tones used in each panel are indicated by black-markers along the rightmost edge of each STRF panel, and the red-arrows along the rightmost edge of the *late-STRF* panels. Note that the *pre-* and *post-STRFs* resemble each other and exhibit a clear BF highlighted by the dashed oval circles.

STRFs in the intermediate (*figure*) epoch display strikingly different features of either *suppressed* (blue stripes in the

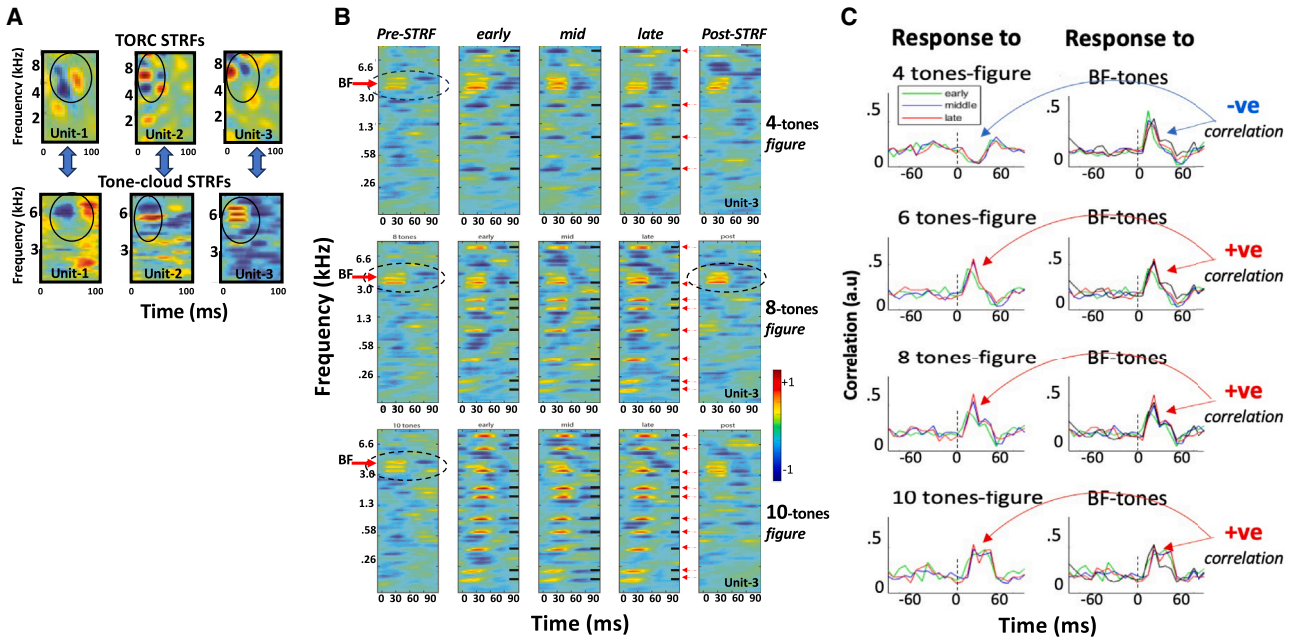


Figure 3. STRF measurements in relation to the *figure* in the SFG

(A) Comparing STRFs measured with TORCs and random tones. Examples of 3 units where TORC STRFs (top row) share prominent features with tone-cloud STRFs (bottom row) but differ in detail.

(B) STRFs of a single cell measured in different epochs of the stimulus. *Pre*- and *post*-STRF are computed from the pure tone-cloud responses, with the dashed circles highlighting the BF of the cell. In the intermediate epochs (*early*, *mid*, and *late*), the STRFs exhibit the effects of the *figure* presentation through the negative (blue) and positive (red) striations (see text and Methods for details).

(C) Measuring the correlation between the *figure*-triggered and BF-triggered responses. Each panel on the left depicts the PSTH response of the cell when triggered by the onset of the *figure* (defined as time = 0 ms). On the right is the cell's PSTH response when it is driven by a BF tone in the stimulus stochastic background. When the two panels are similar (i.e., their inner-product is positive), they are defined as positively (+*ve*) correlated as in the cases of the 6-, 8-, and 10-tone *figures*). Negatively correlated (-*ve*) PSTH responses are seen for the 4-tone *figure* (*top row of plots*) because the two panels are opposite. It is therefore postulated that the *figure* will excite the cell repeatedly and coherently with many other cells for the 6-, 8-, and 10-tone cases, but not with the 4-tone *figure.*

4-tone panels) or *enhanced* (red/yellow-stripes in the 8 and 10-tone *figure* panels). These narrowly tuned features emerge because the *figure* tones are exactly mutually synchronized (strongly correlated), and hence the stimulus is not spectrally white. Consequently, if one of the tones of the *figure* activates the cell near its BF at 6 kHz (e.g., the 2nd tone from the top in the 8-tone *figure* or the 3rd tone in the 10-tone *figure*), then the responses will be highly correlated with all the other *figure* tones resulting in an STRF with enhanced red-striped regions (*middle and bottom row panels* of Figures 3B and S5). A valuable insight gained from the emergence of these extra peaks is the evidence that the *figure* in fact activated the cell with one of its tones. When none of the *figure* tones activates the cell, or if they suppress the cell's responses as when aligned with its inhibitory sidebands, the *figure* tones become negatively correlated with the cell's responses, and thus induce suppressed (blue-striped) regions in the STRF (*top panels* of Figures 3B and S5A). Again, one can infer here that unlike with the 8- and 10-tone *figures*, the cell was incoherently activated (asynchronous) relative to cells that are positively driven by the *figure*.

In the interest of avoiding additional nomenclature, and since we use precisely the same algorithms to compute the *pre*-STRF and *post*-STRFs as we do during the *figure* presentations, we shall continue to refer to these measurements as STRFs (*early*-, *mid*-, and *late*-), although we are aware of the limitations of these estimates as strictly STRFs. Furthermore, because of the noisy character of all the STRFs, the estimates are smoothed by removing all but the most significant voxels (see *Data Analyses* in Methods) and then summing the remaining significant voxels to give an estimate of the "strength" of the STRF, or indirectly the strength of the cell's responses. Since the *pre*-STRFs and *post*-STRFs are measured with the same tone-cloud, comparing them allows us to directly estimate the STRF changes due to the intervening *figure*. By contrast, *early*-STRFs, *mid*-STRFs, and *late*-STRFs reveal the *dynamics* of STRF changes during *figure* presentation, and how the *figure* activations of a cell relate to its overall effects on the STRFs.

The effects of the *figure* on cortical cells are exemplified by the responses and STRF changes in the unit of Figure 3B. For instance, the detailed features of the *early*-STRFs, *mid*-STRFs, and *late*-STRFs are dependent on the arbitrary experimental alignment of the *figure* relative to the BF. Thus, on the one hand, the cell is *suppressed* by the 4-tone *figure* (*top panels*; Figure 3B) inducing inhibitory responses aligned to the *figure* tones. Consequently, this cell's STRFs exhibit gradually diminished BF peaks (*early* > *mid* > *late* STRFs) and an overall no-change between the *pre*- and *post*-STRFs (to be quantified later). On the other hand, the same cell is positively driven by the 8- and

10-tone *figures* (*middle and bottom rows* of Figure 3B *panels*), causing coherent responses that seem to enhance its STRFs. This condition is more likely to occur with bigger *figures* since they contain more tones that might align with the BF's.

To measure and summarize STRF changes from a large population of cortical cells, it is essential to consider the alignment of the *figure* tones with the BFs for *each cell* and *test* separately. Figure 3C illustrates how this relationship is quantified for one cell. We begin by first computing the *figure*-triggered PSTH (*left column of panels*), which appears suppressed for the 4-tone *figure* but excitatory for the 6, 8, and 10-tone *figures*. We then measure the BF-triggered PSTH of the cell during the *pre-* or *post-figure epochs* (*right column of panels*). The match (inner-product) between the two PSTHs is indicative of the nature of the *figure* activations: if the sign of the match is negative (*-ve*) as in the 4-tone *figure* case (*top row*), it indicates the presence of blue-stripes in the *early-*, *mid-*, and *late*-STRFs of the corresponding panels in Figure 3B; if the sign is positive (*+ve*), it reflects the presence of red-stripes during the *figure*.

STRF response changes and binding in the neuron population

We consider next STRF changes due to *figure* presentations in a large population of diverse cortical cells. When a *figure* aligns with the BF in a portion of cortical cells, it stimulates coherently these cells and induces *+ve* correlations (Figure 3C). The proportion of such cells becomes larger with bigger *figures* because there are more tones that may align with the cells' BFs. Since coherent responses are hypothesized to bind the responsive cells (Figure 1B), we expect to detect more enhanced responses with bigger *figures*, which in turn leads hypothetically to the perceptual "pop-out" of the *figure*.[18]

We first examined response changes as a function of *figure* size in a large population of cells. Figure 4A displays the distributions of the STRF changes between the *pre-* and *post*-STRF in recordings from *all* 3 ferrets (1,668 tests in 277 cells), clustered according to the size of the *figures*. For each test, the STRF change is defined as:

$$\mathbf{\Delta_{post\text{-}pre}} = \Sigma_{f,t}\,(\text{post-STRF} - \text{pre-STRF})$$

As predicted earlier, the distribution of the 10-tone *figure* changes is significantly more positively shifted *($p < 0.001$)* relative to the 4- or 6-tone *figures* whose *means* are often insignificantly changed. Examples from the 3 animals individually are shown in Figure S6. To confirm that the alignment of the *figures* with the BF can make a significant difference to the STRF modulations, we combined in Figure 4B the tests of 8- and 10-tone *figures* in ferret B (125 tests) and clustered them according to their positive or negative correlation signs. As postulated, cells with **+**ve correlations exhibited significant STRF enhancements (blue distribution) compared to the centered distribution of the **-**ve correlated cells. Therefore, when cells respond coherently to a *figure,* their responses and STRFs become more enhanced, hypothetically because of the cooperative binding among them (Figure 1C).

A more succinct view of STRF *enhancements* in all cells and tests is shown in Figure 4C where bar heights indicate the increasing proportion of enhanced cells with the *figure* size. It is also evident that 8-tone *figures* are on average as effective as the 10-tone *figures.* Finally, the full range of STRF changes with all *figure* sizes are shown in Figure 4D where all tests were conducted with the identical parameters in ferret B (1,166 tests). Again, the 10-tone *figure* induced significantly more enhanced STRF changes (rightward shifts in the distributions; $p < 0.001$), less so for the 6- and 8-tone *figures* ($p < 0.01$), and no apparent net STRF changes for the small 4-tone *figures*. Interestingly, there are no substantial *net suppressive* STRF changes (leftward shifts in the distributions) likely because of the balanced mix of **+***ve* and **-***ve* correlated cells even for the 4-tone *figures*, causing the overall STRF changes to cancel out. Another possible reason is that STRF changes ($\Delta_{post\text{-}pre}$) represent the persistent effects *after* the end of the *figure*, and it is conceivable that suppressive effects do not persist as well as enhancements. These possibilities are explored next through the dynamics of STRF changes *while figures* are presented, i.e., changes relative to the *early-*, *mid-*, and *late*-STRFs.

Dynamics of STRF modulations during *figure* presentations

STRF changes between the *pre-figure* and *post-figure* measurements ($\Delta_{post\text{-}pre}$) presumably buildup gradually during *figure* presentations. They can be directly captured by tracking the strength of the STRFs in each of the *early-*, *mid-*, and *late*-epochs. We begin by quantifying the STRF strengths as:

$$S_{pre} = \Sigma_{f,t}\,(\text{pre-STRF})$$

$$S_{early} = \Sigma_{f,t}\,(\text{early-STRF})$$

$$S_{mid} = \Sigma_{f,t}\,(\text{mid-STRF})$$

$$S_{late} = \Sigma_{f,t}\,(\text{late-STRF})$$

$$S_{post} = \Sigma_{f,t}\,(\text{post-STRF}),$$

where S represent the integrated sum over all the *significant* voxels in the indicated STRFs. All are computed over 1-s intervals within the different epochs (see *Data Analyses* in Methods). The bar plots of Figure 5A illustrate the dynamics of the average STRF buildups estimated from 172 cells (371 tests) in ferret B and sorted according to the size of each *figure*. We note two characteristics of the responses. (1) During the early epoch, S_{early} becomes significantly enhanced (relative to the initial S_{pre}) only for the larger 10-tone *figure.* By the *mid*-epoch, within 2–3 s after the onset of the *figure,* the 6-, 8-, and 10-tone *figures* S_{mid} become significantly enhanced. (2) During the *late*-epoch, at about 4 s from the onset of the *figure*, the S_{late} begins to diminish. Nevertheless, after the end of the *figure*, STRF enhancements apparently persist into the *post-figure* period, resulting in net positive changes between *pre-* and *post*-STRFs ($\Delta_{post\text{-}pre}$).

STRF changes due to the *figure* activations may in fact be bigger and faster than indicated in Figure 5A because the tests included a mix of both *+ve* and *-ve* correlated tests. In Figure 5B,

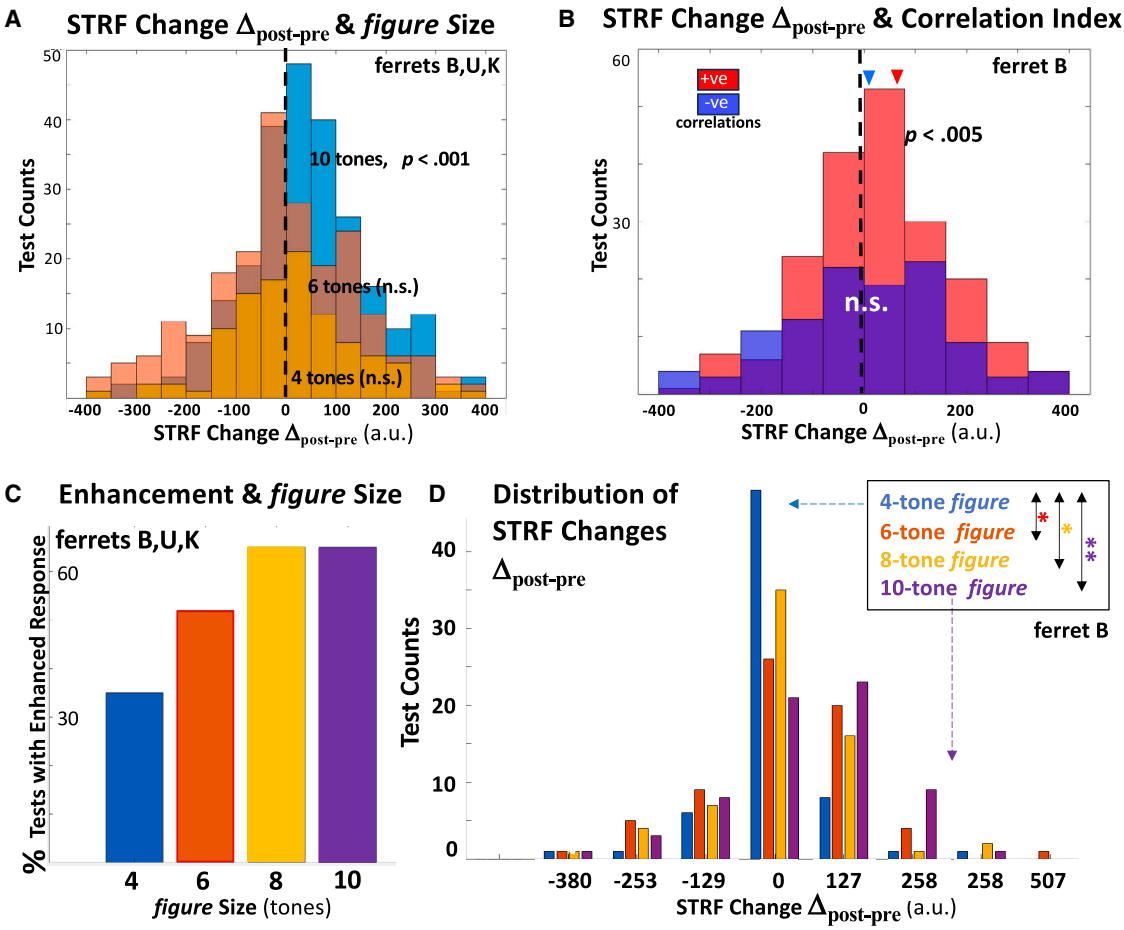


Figure 4. The *figure* size and the distributions of STRF changes $\Delta_{post-pre}$

(A) Distribution of all STRF changes $\Delta_{post-pre}$. All cells are included in this distribution regardless of their correlations with the *figures*. STRFs are more likely to be enhanced with bigger (10-tones) *figures* (p = 3.14e-06), and least for the smallest *figures* (4 and 6-tones), both of which were not significantly shifted.

(B) Distributions of STRF changes ($\Delta_{post-pre}$) for tests with only 8- and 10-tone *figures*, clustered according to their correlations. STRFs from *+ve* correlated tests (red) are rightward shifted (enhanced: p = 0.003), while of *-ve* correlated tests are randomly distributed (not significantly biased).

(C) % of tests with enhanced responses as a function of *figure* size.

(D) Distribution of all STRF changes ($\Delta_{post-pre}$) as a function of *figure* size. STRF changes for the 4-tone *figure* are nearly absent. STRFs of the larger *figures* are gradually more significantly rightward shifted (for 10-tone *figure*, ** is p = 0.001; for 8- and 6-tone *figures*, * are p = 0.03, 0.049, respectively). All significance tests are one-sample t tests.

we segregated the two contributions and depicted only the *-ve* correlated tests to illustrate how the STRFs do not experience any significant enhancements in these *figure* conditions. We conjecture that such **-***ve* correlated cells are inhibited by the *figures* (e.g., 4-tone *figure* in Figure 3B), or are simply unaffected, and hence do not bind due to their incoherent activation compared to the *+ve* cells, (hypothesis of Figure 1C).

A different segregation of tests is shown in Figure S7 where we designated the tests according to their final states as enhanced ($\Delta_{post-pre} > 0$) or suppressed ($\Delta_{post-pre} < 0$). In the enhanced group (*left panel*), all 6-, 8-, and 10-tone *figures* rapidly modulate up the cells' STRFs during the *early*-epoch, enhancements then diminish during the *late*-epoch but persist after the end of the *figure* to give a significant $\Delta_{post-pre}$. In the suppressed group (*right panel*), the STRFs rapidly *diminish* during all *figure* presentations, resulting in a relatively suppressed $\Delta_{post-pre}$.

Functional ultrasound imaging of SFG responses

The spatiotemporal distribution of the average neuronal activity in the auditory cortex during the SFG stimulus was tracked using fUS imaging in two ferrets (U and Z). This technology offers stable images of cerebral blood volume changes over large brain areas, which are assumed to reflect indirectly 1-2 s delayed neuronal responses evoked by the tone-cloud and *figure* over the stimulus epochs of Figure 1C. Details of the technical and physiological procedures are available in Methods and previous publications.[19–22] Figure 6 illustrates images from ferret U who also provided single-unit responses in experiments I and II.

Figure 6A illustrates an anatomical view of two adjacent cross-sectional planes approximately through the primary and secondary auditory cortex. These images are assembled from the average neuronal response during the initial pre-*figure* interval (Figure 1C). Two out of the 6 planes provided strong responses

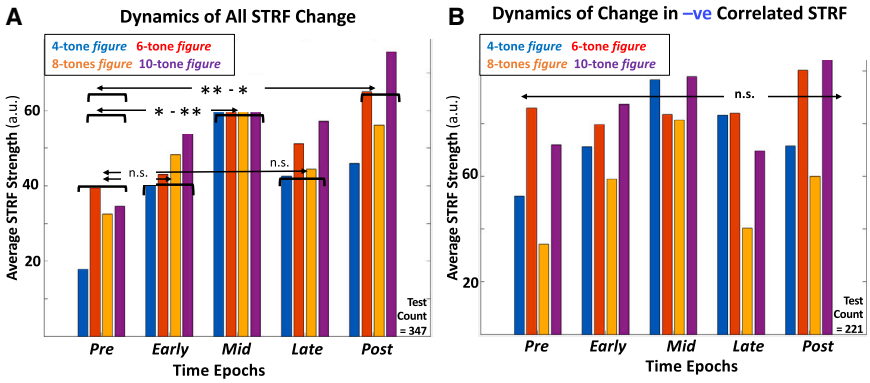


Figure 5. Dynamics of STRF modulations in ferret B responses

(A) Larger *figures* modulate STRFs rapidly (347 tests). For example, STRFs rapidly enhance reaching a peak by the mid-epoch about 2 s from the onset of the *figure* (*p = 0.005*, *n.s.*, *03, 0.01). Late*-STRFs decrease and become non-significantly enhanced relative to the *pre*-STRFs. In the *post-STRFs* most *figures* induce significant enhancements relative to the *Pre* (*p = 0.04*, *0.04*, *n.s*, *0.001).*

(B) Modulations of *-ve* correlated STRF (221 tests). No significant STRF modulations occur in any epoch of the *figure* presentations.

(planes 1 and 2; see Methods). Figure 6B displays the average responses from these two cortical slices imaged mostly during the *figure* epochs (6–12 s after stimulus onset). Both 4 and 8-tone *figures* were tested, with the 8-tone evoking significantly stronger responses in both cortical planes. The evolution of the responses throughout the stimulus, including the *figure* epochs, and the subsequent *post-figure* interval is depicted in Figure 6C. All responses are referenced to the average of the first 4 s interval (the initial two panels from the left). The *figure* responses emerge relatively rapidly in about 1s after onset, peak within 3 s, and decay in the post-*figure* epoch, a rising-falling pattern that resembles the dynamics of the single-unit responses of Figure 5. The increased responses to the 8-tone (but not 4-tone) *figure* persist during the *post-figure* epoch, remaining above the initial 4 s of the *pre-figure* which is consistent with the single-unit findings that STRF enhancements can be estimated from the *pre-* and *post-figure* epochs.

Similar pattern of response enhancements were measured in a second animal (ferret Z) whose images are shown in Figure 7. This animal was significantly older (6 years) than ferret U (2 years). There was a clear enhancement of responses with bigger *figures* as in the younger animal (Figure 6). The notable difference between the two cases is the significantly slower dynamics of buildup and decrease in the older animal, taking about 2–3 s longer to reach its peak (Figure 7). We speculate that age maybe the cause of the slowdown, but as we discuss in the following texts, we have no sufficient evidence to confirm this conjecture for now.

DISCUSSION

This study addressed the neural correlates of *binding*, the process that helps an animal segregate the sources in its surrounding auditory scene, by gluing together the myriad features that belong to one source (e.g., a voice's pitch, timbre, and location), while simultaneously distinguishing them from those of other speakers in a mixture. It has been postulated over the last decade that binding (or equivalently the segregation of auditory scenes) relies on the temporal coherence of the sources' signals, or the idea that temporally coherent sensory stimuli induce rapid plasticity among the synchronously activated neurons causing them to enhance their excitatory connectivity, while simultaneously inhibiting the responses of incoherently responding neurons (Figure 1; refer S. Shihab and M. Elhilali[1] and Shamma S.A et al.[3]). Such synaptic modulations are presumed to occur rapidly, on the order of 100–200 ms time constants[7,23] when an animal or a human switch their attentional focus from one source to another.

Psychoacoustic and physiological studies of binding

As mentioned earlier, both experimental stimuli in Figures 1B and 1C have been explored in human psychoacoustic studies of stream segregation.[1,10,18] Our physiological responses are consistent with these behavioral findings in all aspects investigated. For example, the proportional increase in STRF enhancements and rapid dynamics with the size of the *figures* (Figure 5) parallel well the higher detection accuracy and faster reaction times observed in humans[18] despite the absence of a behavioral task in our ferrets. The SFG stimulus mirrors many characteristics of signals in complex realistic settings, e.g., speech mixtures and orchestral musical streams. It is versatile in that it can readily be adjusted to different levels of difficulty by modifying the *figure* sizes, the jitter among its tones, or the density of the tone-cloud. Thus, human subjects' perception of this stimulus has been shown to reflect their performance on the quick speech-in-noise tasks as well as individual differences in working memory capacity and self-reported musicianship.[18]

While temporal coherence phenomena have already been demonstrated in physiological recordings in EEG/MEG/ECoG human studies,[11,24–28] the animal experiments here afforded us a more detailed assessment of the relationship between binding and the changes in neuronal responses and their dynamics under different attentional conditions.[7] For instance, experiment I (Figure 1B) deployed *multiple* simultaneous streams (sequences) of different tokens, e.g., tones and noise bursts that could be perceptually re-organized on a trial-by-trial basis in different ways depending on the attentional state of the animal. In experiment II (Figure 1C), coherent tone-sequences (4- to 10-tone *figures*) were embedded in a complex random auditory background that modulated binding efficacy among the *figure* tones facilitating assessment of its limits and dynamics even in the absence of selective attention.

Evidence of binding

A direct physiological test of the binding hypothesis (Figure 1) can conceivably be done by direct observations and

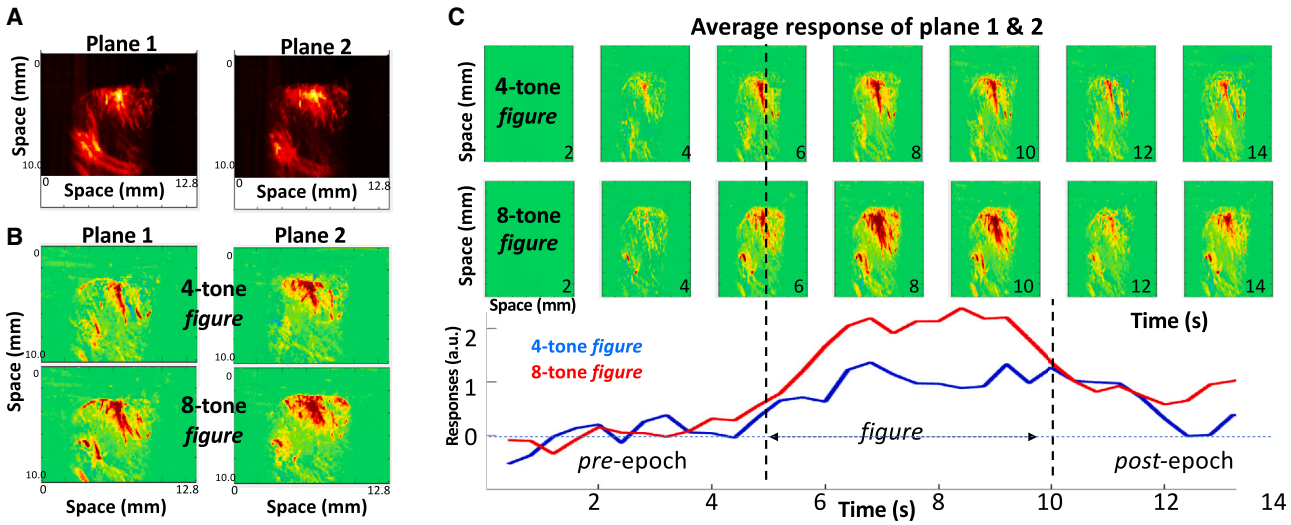


Figure 6. Functional ultrasound (fUS) imaging in ferret U

(A) Anatomical vasculature of the auditory cortex. It is revealed by the activity induced during the early random tone-cloud. Two planes of the auditory cortex are imaged located at the center of the primary auditory cortex, spanning 350–500 mm in thickness.

(B) Activations averaged over the entire *figure* presentations. The smaller 4-tone *figure* evokes significantly weaker activations than the 8-tone *figure* in both cortical planes.

(C) Dynamics of the response modulations. Sequence of 2-s average activations from all trials. Enhancements of activity are pronounced when the *figure* is large (8-tone *figure*), peaking at about 3 s from onset of the figure, decaying in the latter part of the *figure* and post-*figure* epoch.

measurements of connectivity between cortical cells in *in vitro* slices while undergoing synchronous versus asynchronous pulsatile stimulation, an experiment that has not been reported. Clearly, such experiments are difficult using an *in vivo* preparation such as our behaving ferrets. Instead, we have measured the postulated effects of synaptic modulations among neurons driven coherently or incoherently, e.g., by measuring the dynamics and strength of their responsiveness and STRFs. Thus, it is demonstrated in experiment I that attending to the noise sequence significantly suppresses responses of neurons that are incoherent with it relative to the coherent responses which remain unaffected or become weakly enhanced (Figure 2B). The role of attention in this process is critical since without favoring a neuronal response by attending to it, the two populations of coherent and incoherent responses would remain comparable and competitive and hence may not exhibit significant changes. In experiment II, selective attention is absent. Instead, the balance of coherent responses (induced by the single stream of synchronized tones of the *figure*) far outweighs the incoherent responses driven by the random tone-cloud, and thus the most effective binding is that due to the *figure* coherent responses, which explains the response enhancements (Figures 4 and 5).

Suppression versus enhancement in binding

When the animals behaved in Experiment I, suppression of incoherent responses was significant in all tests, whereas enhancement relative to the passive state was often statistically weak. This may well reflect findings from human MEG recordings in which the suppressed background of a scene plays a critical role in parsing the scene.[29] Nevertheless, despite the variability of the plasticity effects, the *relative* response changes were consistent with evidence for the binding hypothesis in three ways: (1) response changes in ASYN (Δ^{ASYN}) versus SYN (Δ^{SYN}) conditions significantly satisfied the inequality $\Delta^{ASYN} > \Delta^{SYN}$ (Figure 2E); (2) during behavior, the binding hypothesis predicts that *contrast* between coherent and incoherent responses should increase, as found by the inequality $C_{act} > C_{pass}$ (Figure 2F); (3) ASYN and SYN response changes (Δ^{ASYN}, Δ^{SYN}) in each cell were coordinated such that randomizing this association or setting $\Delta^{SYN} = 0$ across all neurons disrupted both inequalities ($\Delta^{ASYN} > \Delta^{SYN}$ and $C_{act} > C_{pass}$ in Figures S4C and S4D). Therefore, while the enhancements (Δ^{SYN}) in experiment I were small (Figure 2C), they were significant enough to affect the binding as measured by its consequences.

Finally, there was a substantial number of neurons that exhibited no response changes. One reason is that for binding to occur in our paradigm (Figure 1B) neurons had to be proximate enough to interact (e.g., nearby BF's) but far enough to avoid strong direct stimulation by both tones. The same applied to the noise stream which had to be far enough to avoid strongly stimulating the isolated cells but not too far to be irrelevant to the cells since binding effects are expected to weaken with spectral distance.[27] These compromises were often difficult to balance given the independent microelectrodes. While the positioning of the tones was carefully selected relative to the BF of one of the neurons, there were undoubtedly other confounds emanating from the interaction between the stimulus tones/noise and the excitatory/inhibitory fields of the cell which may cancel out changes in the neural response.

In experiment II, the balance of enhancements and suppression was the opposite, especially for the largest *figure* sizes

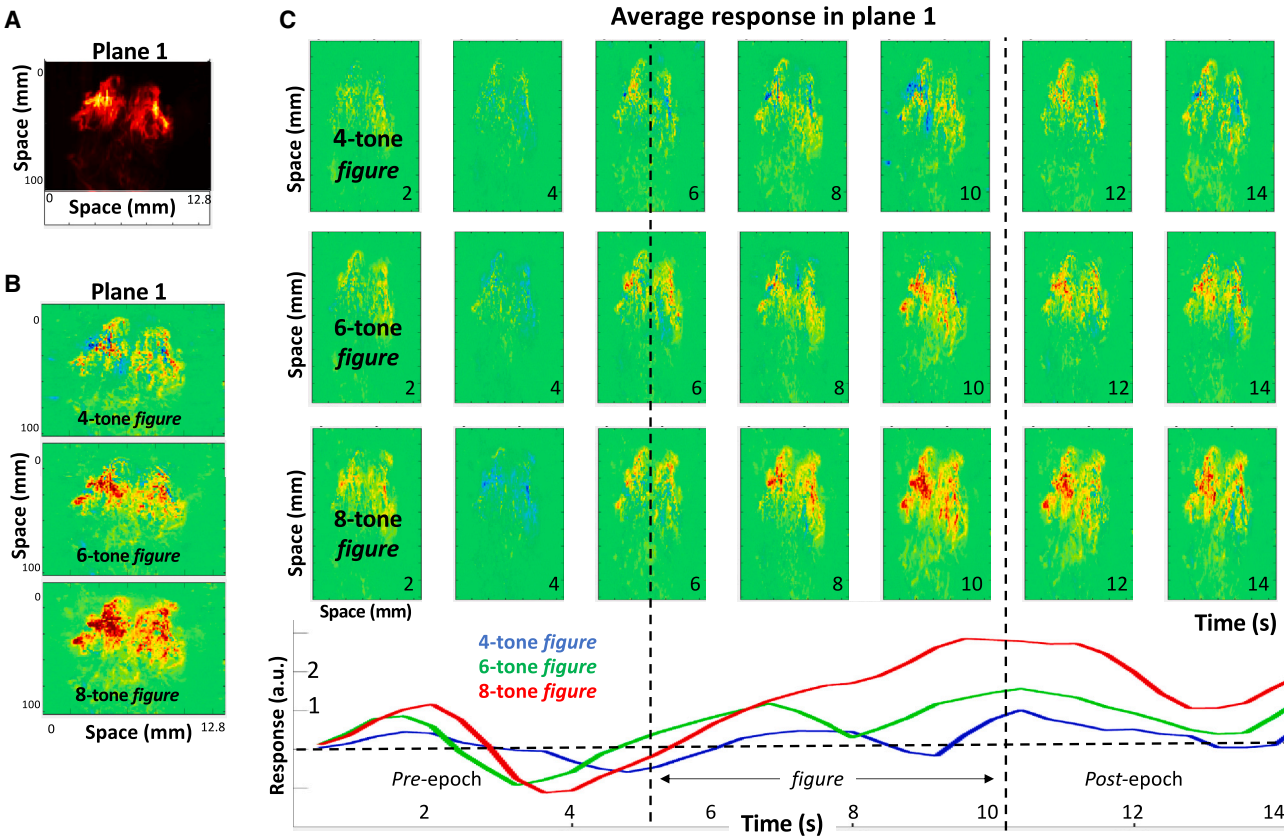


Figure 7. fUS imaged responses to SFG stimulus in an older ferret Z

(A) Anatomical vasculature of the auditory cortex in plane 1, measured by the activity induced during the early random tone-cloud. The image is the center of the primary auditory cortex, spanning 350–500 μm in thickness.

(B) Activations averaged over the entire presentations of three *figure* presentations. Activations gradually increase with the size of the *figures*.

(C) Dynamics of the response modulations. Sequence of 2 s average activations throughout the 3 types of trials. They are strongest for the 8-tone *figures*, compared to the other two *figures*. Responses however peak at about 5 s from onset of the figure, significantly delayed compared to the younger animal (Figure 6C). They decay in the latter part of the *figure* and post-*figure* epoch as in previous tests.

(Figure 4). Suppression was still evident when we focused on the plasticity in the **-*ve*** cells (Figure S7), but the dominance of such suppressive influences seen in experiment I is absent. This is likely because of the use of sizable 6- to 10-tone *figures* that recruited and synchronized many neurons *(+ve* cells; Figure 3C) thus inducing response enhancements. All these effects were measured in passive animals (as was the case in some of the human experiments[9–11]), but we conjecture, based on all our previous experiences,[7,12,13,17,27,30] that the binding plasticity would be significantly stronger if the SFG paradigm had included a behavioral task.

Dynamics of the binding process

Both enhancements and suppression were evident in experiment II, where response dynamics occurred rapidly, within fractions of a second (Figures 5A and S7), suggesting that binding may result from modulations of *already* existing interneuronal connectivity (as opposed to formation of new connections which would presumably be slower). While the onset of response modulations was rapid, it was relatively slow to peak taking approximately 3 s after stimulus onset (Figures 3B, 5 and S5, and S7), and did not persist for long as it waned although remaining elevated for seconds afterward (Figures 3B and 5A). It is likely that a task-engaged animal may have exhibited different and even faster temporal evolutions. For instance, in humans attending to and segregating speech mixtures,[23] binding of different neuronal populations to change the listener's focus and select the desired target must be significantly faster than seen here. In ferrets, segregating speech streams also occurs rapidly within a fraction of a second (Figure 6B in the study by Joshi N. et al.[31]).

Another speculative observation concerning the dynamics of binding is that this process enables listening in real-world cluttered environments. As such, if it weakens or slows down with age, it becomes presumably more difficult to isolate a target voice from a mixture even if hearing remains objectively normal by most measures. The fUS measurements in the older ferret K (Figure 7) speculatively illustrates this point in that while the *figure* enhancements appeared robust, the dynamics of the process were noticeably slower. Aged human listeners also exhibit difficulties detecting low-rate stochastic FM modulations[32] as well as poorer speech comprehension in noise,

much like the SFG subjects in the study by Johns M.A et al.[18] This speculation clearly needs confirmation in more aged animals, but if it does, this stimulus may well serve as a model for the mechanisms that cause the deterioration in the binding process.[33–35]

Conclusion

Binding of temporally coherent stimulus features is the fundamental process that ties together all human and animal psychoacoustic and physiological threads that we addressed in this study. More critically, temporal-coherence provides an overall theoretical framework to predict and interpret the results. And, despite the absence of direct evidence demonstrating binding as modulations of synaptic efficacy between co-activated cells, the indirect physiological consequences remain compelling as illustrated here and in previous experiments with simpler stimuli.[7] Future studies however would benefit from experiments on two ends of the spectrum. On one extreme, one may measure the direct connectivity between cell pairs or ensembles "*in vitro*" (using intracellular recordings in a brain slice) as they are stimulated coherently or incoherently. On the other extreme are experiments to engage animals in segregating and selecting a target voice in speech mixtures while recording the effects of binding across the cortical spectral channels of the target and distractor speakers.[31] Simulations of these processes[36] have already proven their efficacy at segregating complex sound mixtures. Thus, despite its relative conceptual simplicity, binding through temporal coherence may well underlie a substantial part of our remarkable abilities at disentangling sources in cluttered environments.

Limitations of the study

This study offers suggestions for two key future explorations to elaborate and validate further its conclusions. The first involves measurements of the direct connectivity between cells or ensembles (e.g., "*in vitro*" using intracellular recordings in a brain slice) as they are stimulated coherently or incoherently. The goal would be to test if the coherent stimulation modulates up the connectivity between the cells or populations. The second set of experiments would be to replicate the *fUS* measurements in aged vs. younger animals (as in Figures 6 and 7) to establish the generality of the result found thus far in just a single pair of young vs. aged animals.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and fulfilled by the lead contact, Shihab Shamma (sas@umd.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Neural and behavior datasets for single-unit recordings in this study have been deposited in the figshare database and are publicly available. The DOI is listed in the key resources table.[37]
- Scripts for single-unit analysis for Figures 2, 4, and 5 have been deposited in the figshare database and are publicly available. The DOI is listed in the key resources table.
- Datasets for fUS recordings, code for STRF reconstruction and fUS data analysis, and any additional information required to reanalyze the data reported in this paper is available from the lead contact upon reasonable request.

ACKNOWLEDGMENTS

The authors would like to acknowledge funding from the National Institutes of Health (Program Grant from NIA (P01 AG055365), NIDCD (R01-DC016119 and DC005779), and grants from AFOSR (FA9550-19-1-0408) and ONR Muri (#2005869237), and a training grant NIH-T32 to K.D.). Partial funding was also made available by an Advanced ERC (Neume) grant.

AUTHOR CONTRIBUTIONS

K.L. participated in the design of experiment I, single-unit recordings, analysis of the data, and writing of the relevant results in MS; K.D. participated in the design of experiment II, all its single-unit recordings, and the initial stages of data analyses; A.M. conducted the *fUS* recording experiments; M.E. provided the conceptual guidance to the experiments, and helped with data analyses, and review of the manuscript at various stages; S.S. worked on the design of the experiments, data analyses, and writing of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interest.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Subjects
- METHOD DETAILS
  - Experiment I
  - Experiment II
  - Functional ultrasound imaging
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical analysis
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.111991.

Received: May 28, 2024
Revised: July 25, 2024
Accepted: February 6, 2025
Published: February 12, 2025

REFERENCES

1. Shihab, S., and Elhilali, M. (2021). Temporal coherence principle in auditory scene analysis. In the *The Senses* A Comprehensive Reference, 2nd edition (Elsevier).
2. Bregman, A.S. (1994). Auditory Scene Analysis: The Perceptual Organization of Sound (MIT press).
3. Shamma, S.A., Elhilali, M., and Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. Trends Neurosci. *34*, 114–123.
4. Bizley, J.K., and Cohen, Y.E. (2013). The what, where and how of auditory-object perception. Nat. Rev. Neurosci. *14*, 693–707.
5. Middlebrooks, J.C., and Bremen, P. (2013). Spatial stream segregation by auditory cortical neurons. J. Neurosci. *33*, 10986–11001.

6. Itatani, N., and Klump, G.M. (2017). Animal models for auditory streaming. Philos Trans R Soc Lond, B, Biol Sci *372*, 20160112.
7. Lu, K., Xu, Y., Yin, P., Oxenham, A.J., Fritz, J.B., and Shamma, S.A. (2017). Temporal Coherence Structure Rapidly Shapes Neuronal Interactions. Nat. Commun. *8*, 13900.
8. Ma, L., Micheyl, C., Yin, P., Oxenham, A.J., and Shamma, S.A. (2010). Behavioral measures of auditory streaming in ferrets (Mustela putorius). J. Comp. Psychol. *124*, 317–330.
9. Teki, S., Chait, M., Kumar, S., von Kriegstein, K., and Griffiths, T.D. (2011). Brain bases for auditory stimulus-driven figure–ground segregation. J. Neurosci. *31*, 164–171.
10. Teki, S., Chait, M., Kumar, S., Shamma, S., and Griffiths, T.D. (2013). Segregation of complex acoustic scenes based on temporal coherence. Elife *2*, e00699.
11. O'Sullivan, J.A., Shamma, S.A., and Lalor, E.C. (2015). Evidence for neural computations of temporal coherence in an auditory scene and their enhancement during active listening. J. Neurosci. *35*, 7256–7263.
12. Elgueda, D., Duque, D., Radtke-Schuller, S., Yin, P., David, S.V., Shamma, S.A., and Fritz, J.B. (2019). State-dependent encoding of sound and behavioral meaning in a tertiary region of the ferret auditory cortex. Nat. Neurosci. *22*, 447–459.
13. Yin, P., Strait, D.L., Radtke-Schuller, S., Fritz, J.B., and Shamma, S.A. (2020). Dynamics and Hierarchical Encoding of Non- compact Acoustic Categories in Auditory and Frontal Cortex. Curr. Biol. *30*, 1649–1663.e5.
14. Otazu, G.H., Tai, L.-H., Yang, Y., and Zador, A.M. (2009). Engaging in an auditory task suppresses responses in auditory cortex. Nat. Neurosci. *12*, 646–654.
15. (2013). Handbook of modern techniques in auditory cortex. In Chpt. 1: A Linear Systems View to the Concept of STRFs, M. Elhilali and D. Depireux, eds..
16. Klein, D.J., Depireux, D.A., Simon, J.Z., and Shamma, S.A. (2000). Robust Spectro-temporal Reverse-Correlation for the Auditory System: Optimal Stimulus Design. J. Comput. Neurosci. *9*, 85–111.
17. Jonathan, F., Shamma, S., Elhiliali, M., and Klein, D. (2003). Rapid task-dependent plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat. Neurosci. *6*, 1216–1223.
18. Johns, M.A., Calloway, R.C., Phillips, I., Karuzis, V.P., Dutta, K., Smith, E., Shamma, S.A., Goupell, M.J., and Kuchinsky, S.E. (2023). Performance on stochastic figure-ground perception varies with individual differences in speech-in-noise recognition and working memory capacity. J. Acoust. Soc. Am. *153*, 286.
19. Bimbard, C., Demene, C., Girard, C., Radtke-Schuller, S., Shamma, S., Tanter, M., and Boubenec, Y. (2018). Multi-scale mapping along the auditory hierarchy using high-resolution functional UltraSound in the awake ferret. Elife *7*, e35028.
20. Landemard, A., Bimbard, C., Shamma, S., Norman-Haignere, S., and Boubenec, Y. (2019). Functional Segregation of Ferret Auditory Cortex Probed with Natural and Model-Matched Sounds (Universitätsbibliothek der RWTH Aachen).
21. Montaldo, G., Urban, A., and Macé, E. (2022). Functional ultrasound neuroimaging. Annu. Rev. Neurosci. *45*, 491–513.
22. Demené, C., Bimbard, C., Gesnik, M., Radtke-Schuller, S., Shamma, S., Boubenec, Y., and Tanter, M. (2016). Functional Ultrasound Imaging of the thalamo-cortical auditory tract in awake ferrets using ultrafast Doppler imaging. In 2016 IEEE International Ultrasonics Symposium (IUS) (IEEE), pp. 1–4. https://doi.org/10.1109/ULTSYM.2016.7728659.
23. Mesgarani, N., and Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature *485*, 233–236.
24. Teki, S., Barascud, N., Picard, S., Payne, C., Griffiths, T.D., and Chait, M. (2016). Neural correlates of auditory figure-ground segregation based on temporal coherence. Cereb. Cortex. *26*, 3669–3680.
25. O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G.M., Sheth, S.A., Mehta, A.D., and Mesgarani, N. (2019). Hierarchical encoding of attended auditory objects in multi-talker speech perception. Neuron *104*, 1195–1209.e3.
26. Chambers, C., Akram, S., Adam, V., Pelofi, C., Sahani, M., Shamma, S., and Pressnitzer, D. (2017). Prior context in audition informs binding and shapes simple features. Nat. Commun. *8*, 15027.
27. Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J., and Shamma, S.A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. Neuron *61*, 317–329.
28. Rezaeizadeh, M., and Shamma, S. (2021). Binding the acoustic features of an auditory source through temporal coherence. Cereb. Cortex Commun. *2*, tgab060.
29. Brodbeck, C., Jiao, A., Hong, L.E., and Simon, J.Z. (2020). Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers. PLoS Biol. *18*, e3000883.
30. Jonathan, F., Shamma, S., and Elhiliali, M. (2005). Differential dynamic plasticity of A1 receptive fields during multiple spectral tasks. J. Neuroscience *25*, 7623–7635.
31. Joshi, N., Ng, W.Y., Thakkar, K., Duque, D., Yin, P., Fritz, J., Elhilali, M., and Shamma, S. (2024). Temporal coherence shapes cortical responses to speech mixtures in a ferret cocktail party. Commun. Biol. *7*, 1392.
32. Sheft, S., Shafiro, V., Lorenzi, C., McMullen, R., and Farrell, C. (2012). Effects of age and hearing loss on the relationship between discrimination of stochastic frequency modulation and speech perception. Ear. Hear. *33*, 709–720.
33. Holmes, E., and Griffiths, T.D. (2019). Normal' hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception. Sci. Rep. *9*, 16771.
34. Guo, X., Dheerendra, P., Benzaquén, E., Sedley, W., and Griffiths, T.D. (2022). EEG Responses to auditory figure-ground perception. Hear. Res. *422*, 108524.
35. Choi, I., Gander, P.E., Berger, J.I., Woo, J., Choy, M.H., Hong, J., Colby, S., McMurray, B., and Griffiths, T.D. (2023). Spectral Grouping of Electrically Encoded Sound Predicts Speech-in-Noise Performance in Cochlear Implantees. J. Assoc. Res. Otolaryngol. *24*, 607–617.
36. Krishnan, L., Elhilali, M., and Shamma, S. (2014). Segregating complex sound sources through temporal coherence. PLoS Comput. Biol. *10*, e1003985.
37. Lu, K., Dutta, K., Mohammed, A., Elhilali, M., and Shamma, S. (2025). Data for Temporal-Coherence Induces Binding of Responses to Sound Sequences in Ferret Auditory Cortex. Figshare. https://doi.org/10.6084/m9.figshare.28266323.
38. David, S.V., Fritz, J.B., and Shamma, S.A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. Proc. Natl. Acad. Sci. USA *109*, 2144–2149.
39. Blake, D.T., and Merzenich, M.M. (2002). Changes of A1 receptive fields with sound density. J. Neurophysiol. *88*, 3409–3420.
40. Eggermont, J.J. (2011). Context dependence of spectrotemporal receptive fields with implications for neural coding. Hear. Res. *271*, 123–132.
41. Tomita, M., and Eggermont, J.J. (2005). Cross-correlation and joint spectro-temporal receptive field properties in auditory cortex. J. Neurophysiol. *93*, 378–392.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
Ferret (Mustela putorius)	Marshall Farm	https://www.marshallferrets.com/
Software and Algorithms		
MATLAB 2010a-2024a	Mathworks	https://www.mathworks.com/
Custom MATLAB code	Neural Systems Laboratory	https://doi.org/10.6084/m9.figshare.28266323
Deposited Data		
Single-unit data	This study	https://doi.org/10.6084/m9.figshare.28266323

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Subjects

Five adult female ferrets (Mustela putorius, Marshall Farms, North Rose, NY) participated in the experiments described here, with one ferret (**U**) taking part in all experiments. Two (ferrets **R,U**) were trained for the neurophysiological **Experiment I**. Three ferrets (**B,U,K**) contributed to **Experiment II** data with the SFG stimuli in a passive state. Two animals (ferrets **U,Z**) were imaged in the *fUS* experiments without task performance.

Trained animals were placed on a water-control protocol in which they obtained water as rewards during behavioral sessions or as liquid supplements if the animals did not drink sufficiently during behavior. All animals also received *ad libitum* water freely over weekends, and their health was always monitored, maintaining above 80% of their *ad libitum* weights. Ferrets were housed in pairs or trios in facilities accredited by the Association for Assessment and Accreditation of Laboratory Animal Care (AAALAC) and were maintained on a 12-h light-dark artificial light cycle.

All animal experimental procedures were conducted in accordance with the National Institutes of Health's Guide for the Care and Use of Laboratory Animals and were approved by the Institutional Animal Care and Use Committee (IACUC) of the University of Maryland.

METHOD DETAILS

Experiment I

Headpost implant surgery

After reaching behavioral criteria on the task, a stainless steel headpost was surgically implanted on the ferret skull under aseptic conditions while the animals were deeply anesthetized with 1%–2% isoflurane. The headpost was secured in the skull using titanium screws and embedded in Charisma (ferrets R and U). The area around the auditory cortex was covered in a single layer of cement, whereas surrounding areas were covered with 5–7 mm-thick cement to protect the exposed skull.

Neurophysiological recording

After recovery from surgery (2–3 weeks), the animals were placed in a double-wall soundproof booth (IAC) and were habituated to the head-fixed setup. They were re-trained in the head-constrained version of the task. Neurophysiological recordings began after the animals regained consistent criterion levels of performance (Hit Rate $\geq$ 75%, Discrimination Rate $\geq$ 40%) in three sequential behavioral sessions. All behavioral data in this paper were obtained after implantation, including regaining performance to pre-surgical levels.

To expose a part of the auditory cortex for recording, small 1–2 mm craniotomies were made in the skull. Recordings were conducted using 4 tungsten microelectrodes (2–5 MΩ, FHC) simultaneously advanced through the craniotomy and controlled by independently movable drives (Electrode Positioning System, Alpha-Omega) with 1 μm precision until well isolated spiking activity was observed. Raw neural activity traces were amplified, filtered, and digitally acquired by a data acquisition system (AlphaLab, Alpha-Omega). Single units were isolated offline using customized spike-sorting software based on PCA, K-means clustering, and subsequent template matching. Only units with greater than 70% isolation and a typical refractory period were conserved as single units.

Localization of recording sites on the tonotopic map

Recording locations in the primary auditory cortex were characterized by their locations along the dorsoventral and rostro-caudal axes. Artificial marks were created by drilling a depression in the headcap on either side of the craniotomy as reference landmarks for localizing recording positions. Furthermore, the neuron's BF was measured at each electrode penetration. The BFs obtained from all penetrations were then aligned to form a tonotopic map for all animals.

Computing the best frequency

The best frequency (BF) of each neuron was measured by analyzing their responses to tone pips with varying frequency and intensity (see Experimental Procedures and Stimuli). A two-dimensional frequency x intensity response matrix was then created by taking the mean evoked response to tones at each frequency and intensity level. The response matrix was baseline-corrected by subtracting the mean and dividing by the standard deviation of the baseline activity from 100 ms before tone onset. The longest iso-response contour line in the normalized matrix was defined as the neuron's tuning curve, and the frequency corresponding to the lowest intensity on the tuning curve was the neuron's BF. A penetration site's BF was computed by taking the median value of all isolated single units in that site.

PSTH calculations

Peri-stimulus time histogram responses (PSTHs) were obtained by binning the neural responses into 10-ms time bins and averaging these windowed spike data across trials and stimulus presentations per cell, then further averaging across cell. Unless otherwise specified, responses during engagement included only hit trials, i.e., trials where the animal successfully refrained from licking the spout until they detected the target change in noise level in **Experiment I**. PSTH responses used in further analysis are without baseline correction or normalization unless specified as "referenced to spontaneous activity," in which case the average of neural responses during a 250-ms pre-stimulus time window was subtracted from the PSTH for each cell before averaging over population. In most plots, we have opted to show averages of PSTH of a neuronal population (unless specified otherwise). We shall refer to the response units as "**a**rbitrary **u**nits" (**a.u**.) as they become far-removed from the spike/sec rates measured with single cells.

Behavioral task

Two young female ferrets (**R,U**) are trained on an appetitive, selective attention task to detect an auditory intensity cue in one portion of a sound sequence. The rationale behind the design of this experiment was that the animals did not need to attend to the tones, but rather only to the noise sequence, which was placed far enough that it did not interfere perceptually or physiologically with the tone responses. This design is also consistent with the entire focus of the study, that the target of attention will induce binding to all other coherently responsive neurons regardless of whether they are explicitly attended to or not. As explained in the Introduction, this rationale is also consistent with the way we postulate speech segregation occurs.[1] A listener wishing to segregate a specific voice in a mixture (e.g., the male voice), needs to attend only to one attribute of their voice (e.g., location or pitch), and the rest of the voice's details and attributes would then bind with the attended response because of their temporal coherence with it.

Animals are normally placed on water restriction in days prior to the experimental sessions and also during the period when they are engaged in task performance. They are required to withhold licking from a waterspout during presentation of a series of reference sounds until the target sound cue is detected. Each stimulus trial consists of a sequence of two alternating pure tones and one sequence of narrowband noise bursts (bandwidth 1/2 octave) that is synchronized at any given trial with one of the tone sequences. Each tone token is 80ms in duration with 5ms cosine ramps at onset and offset. The inter-tone interval is 80ms. All reference tones and noise are kept at a constant intensity until the target cue occurs, usually after 4–30 alternating burst epochs. The target cue is a single noise burst played 10–20 dB louder than the bursts of the reference portion, followed by 1.5 s of continued sequences that are the same as the those in the reference. If an animal licks the waterspout within 1 s after the onset of the target cue, the response is considered to be a hit and the animal is rewarded. If an animal does not lick, or licks after the 1 s response window, the trial is considered a miss. The false-alarm rate is calculated for each trial as the ratio between the number of references during which animals licked the waterspout and the total number of references before the target. During the training phase, the false alarm rate is calculated for each reference sound online. If the false-alarm rate exceeds 0.5 (the subject licked more than 50% reference sounds), the trial is terminated immediately to ensure that the animal is not using a timing cue and simply waiting a fixed period of seconds before licking. All trials with false alarms are excluded from analysis of neural activity. After each training session, false-alarm rates across all trials are averaged and denoted as the overall false-alarm rate of the training session. The "miss" is quantified as a No-response to the target cue within 1s. A training session ends when animals are no longer thirsty and do not respond for 3 consecutive trials. Task performance is quantified as the discrimination rate (DR), and is defined according to the formula below[38]:

$$DR = \text{Hit rate}^{*} \ (1 - \text{False Alarm rate})^{*} \ 100\%$$

Once they performed consistently above chance level, where consistency was defined DR $\geq$ 40%, we considered the animal ready for implantation. Animals' performance during electrophysiology recordings were also quantified as d', because it is a commonly used measurement in psychophysics. d' is defined as below:

$$d' = z(\text{False Alarm rate}) - z(\text{Hit rate})$$

While z-scores were calculated as the inverse of cumulative distribution function of the given false alarm rate or hit rate values. Animals generally performed this task identically in the two conditions (A-SYN and B-SYN) as exemplified by the matched performance of **ferret R** during the two sets of trials (Wilcoxon signed-rank test: $p = 0.079$; Figure S2).

Electrophysiology in behaving animals

Each recording began with 250-ms random tone pips of varying frequency (125–32000 Hz, 4 tones/octave) and intensity (0 to −50 dB range, 10 dB increment) to determine the best frequency (BF) and latency of each individual recording site. Two pure tone frequencies A and B were chosen based on the tuning properties of the neuronal units measured at a given recording depth. Effort was made to choose frequencies near one or more units' BFs that would produce distinct responses in all units. Animals performed the task

in 2 conditions: with tone A synchronous with the noise stream (*A-SYN*); and with tone B synchronous with the noise stream (*B-SYN*), or equivalently with tone A asynchronous with the noise (A-ASYN).

Initially, animals performed all trials of each A-SYN and B-SYN conditions in two consecutive blocks, but later experiments randomly interspersed the two conditions. Tone frequencies were chosen for each recording based on the response properties of the neurons isolated. Ideally, tones are chosen so that one but not both tones fall near the BF of each neuron. The narrowband noise is placed with center frequency at least one octave above the higher pure tone. In many experiments, noise-only trials are also played in the passive and active conditions. The parameters selected for the tone and noise separation and presentation rates in this paradigm are typical of what has been found in streaming perception experiments in ferrets.[8] Furthermore, since the animals succeeded in attending and reacting to the noise sequences, we can safely assume that they potentially could stream the tone sequences apart and treat them as distractors. However, we hypothesized that the coherent tone would be perceptually bound with the attended noise and therefore neurons responding to this tone would have enhanced responses. Analogously, neurons tuned to the asynchronous or incoherent tone were hypothesized to be suppressed. During recordings, both animals performed above the threshold. They achieved average hit rates of 93.30% ± 9.9% and 90.01% ± 11.87% with false alarm rates of 18% and 41%, respectively.

Analysis responses to tone-noise-sequence

To exclude any lick-related artifacts, we only analyzed neural responses to sounds before the target while also excluding all trials with false alarm responses. Single-unit neuronal responses were sorted and binned into 10ms windows and averaged over each 320 ms epoch of the stimulus (or two 80ms tones, each followed by 80ms gap) thus defining the PSTHs used in Figure 2A. Responses were analyzed from the third of reference epochs because we assumed that it may require a few repeats for stream formation to occur. Synchronous (SYN) and alternating (ASYN) conditions for each neuron were assigned by comparing average spike rates during passive presentation of the two alternating tone stimuli. Tone A response values in the passive state were calculated by summing tone A-noise (A-SYN, first half) with tone A alone (B-SYN, second half) responses. The same procedure was followed for tone B response values. Responses were then organized into two PSTHs labeled SYN and ASYN (Figure 2A). They were assigned as SYN or ASYN *relative* to each cell based on which of these response values was greater. For example, if a neuron responds more during (A-noise + A-alone) than (B-noise + B-alone), i.e., A is the *preferred*-tone (**P** in Figure 2A), then we consider A-SYN to be the SYN stimulus for that cell and B-SYN to be the asynchronous (A-ASYN) stimulus. Only neurons with a minimum 10% difference in spike rate were used in the analysis. This selection also ensured that neurons driven only by the noise stream and not tone A or B were removed. Many neurons showed changes in responses during the 4 experimental conditions (SYN_{pass}, $SYN_{act,}$ $ASYN_{pass}$ $ASYN_{act}$). Finally, responses as expected were globally suppressed during task engagement (see text related Figure 2A; *bottom panel*). However, they exhibited more suppression when driven in the $ASYN_{act}$ condition.

Computing the bias size around the midline

In all scatterplots of the data from this Experiment we computed a measure of the overall bias of the points around the midline. The measure captures the *effect-size* of the changes in the population. It is computed by first measuring for each point (cell or test) the signed-distance to the midline (positive/negative if above/below the midline, respectively) and then computing the mean of these signed-distances, *normalized* by the mean of the unsigned-distances. The *effect-size* therefore varies between **+1** to **-1** if all cells' values fall above or below the midline.

To determine if a scatterplot bias is significant, we computed a one-sample t-test of the distribution of all distances of scatterplot points around a midline (as described above) to determine if its mean shifted significantly above or below the midline, as reflected by the ***p-values*** indicated on each panel. If the mean shift is insignificant (***p >0.05***) then both *effect-size* and its ***p-values*** are eliminated from the plots, and the null-hypothesis (no change occurred) is not rejected.

Experiment II

Electrophysiology in Experiment II

Three ferrets were used in the single-unit recordings of this experiment: ferret **B** (172), **U** (53), **K** (52) isolated single-units. Analyzed data were combined from all animals in most plots, although in some cases one animal's data were used because of the need for consistency of stimulus parameters across all the tests. Supplemental data provide results from each animal separately. Since each cell underwent many tests with different *figure* sizes and frequency ranges, we often provided in the text and figures the more relevant total number of tests used in the analyses rather than the number of cells.

Stochastic figure ground stimuli

Random tones. A cloud of random tones (or background) consisted of a group of fixed-length (50 ms) tone pips with 5 ms cosine ramp, which occurred at random onset times within a set of discrete frequency channels. Such a cloud is generated by first selecting an octave range and a fixed number of tones logarithmically spaced per octave. For animals **B** and **U** we used 6 octaves with 6 semitones per octave, centered at 1600 Hz. For ferret **K** the same parameters were used in addition to two spectrally compressed versions, with 3 octave ranges and 12 semitones per octave, starting at the lowest frequencies of 500 and 1500 Hz. Each frequency channel had a mean tone rate of 4 Hz, i.e., on average 4 pips/s in each channel. Onset times were randomly and uniformly generated for a 2s window within a minimum of 50ms spacing between consecutive onsets, and then are jittered +/− 25ms to avoid randomly generated coherence. To obtain a 5s background stimulus, we generated 6 s and kept the first 5, eliminating any pips that overlapped the borders.

Perceptual Figure. For the "*figure*" portion of the stimulus, an initial set of random onset times is generated at an average rate of 4Hz as described above. To control the precise timing of the first *figure* onset and last offset, we required that one onset occurs at t = 0 and one at t = 4.95s. Tone pips occurred in each of the *figure* frequency channels at these onset times. Then, the remaining channels were filled-in randomly as described above. Importantly, the background onsets did not overlap with the *figure* to keep the final stimulus envelope relatively flat.

Figure frequency channels were chosen by randomly generating a subset of 10 of the tone background channels. For the 8-tone *figure*, a subset of 8 of these 10 channels is randomly chosen, etc. for the other tone number conditions. For ease of interpretation, this procedure was performed one time, and then the same *figure* tones were used for the remainder of each experimental session. The same applied to the set of background stimuli which was pre-generated and used as frozen random tones in each session. However, The figures and background however were varied across experimental sessions depending on the cells' tuning and the range of frequencies tested, with the goal of maximizing the chances of having driven cells (BF aligned with the *figure* tones). Contrasting the effects of responses that are aligned versus misaligned to the *figure* tones was a key objective of the experiments. However, it was difficult for 2 reasons to do this systematically in the same cells by for instance shifting up & down the *figure* and completing two sets of measurements in each condition: (i) We wished to maximize the exposure time of an aligned cell to the *figure* to accumulate enough responses (spikes) to estimate the STRFs, and also more importantly to allow the plasticity effects (binding) to become measurable. (2) Another difficulty is that we typically recorded with at least 4 or more separate electrodes. That means, it was inevitable that some of the cells were aligned with the figure, while others were not. This made it, on the one hand, difficult to manipulate the stimuli without affecting all the other isolated cells. Therefore, although it would have been ideal to get the contrast between the two conditions, we felt that having many cells recorded with the same *figure* meant that we will obtain diverse examples of aligned and non-aligned cells which could be contrasted just as desired, but of course not in the same cell.

Finally, each trial typically consisted of 5s of background, followed by 5s of background with an embedded *figure*, then 5s background. We hypothesized that the temporal coherence of the *figure* tones would cause them to be bound into a stream, and that this effect-size would systematically vary with the size of the *figure,* or the number of tones which make up the *figure*. Typically, multiple *figure* sizes were generated and tested with at least 50 repetitions per *figure* size.

Analyses of responses to stochastic figure ground

The background of tones is an auditory stimulus akin to coarse noise, which is commonly used to assess neuronal receptive fields. We first validated that our version of this stimulus can re-capitulate response fields of A1 generated by our standard tuning battery, specifically TORC responses.[16] Using lag values of −10 to 120ms, the cross-covariance of the response and stimulus (auditory spectrogram) was divided by the auto-covariance of the stimulus. Ridge regression was performed with 30 log-spaced values of λ and the results averaged. This procedure was performed for each trial and then averaged, using stimuli and responses to 1s windows immediately before the figure onset (*pre*), immediately after the first onset (*early*), halfway through figure presentation (*mid*), the last second of figure presentation (*late*), and immediately following the last figure onset (*post*). The background responses can approximately reconstruct STRFs generated from TORC responses.[16] Previous studies have indicated that tuning maps generated with background stimuli can vary systematically with tone density, with more spectro-temporally dense clouds producing sharper tuning.[39–41] We chose our density parameters based on intuition of human perceptual limits, but it is possible that a denser background may produce better tuning and different effects. It is important to note also that TORCs contain a continuous range of frequencies spanning a 5-octave range, while the SFG contains only a discrete set of 37 frequencies spanning typically 6 octaves. Thus, excitatory or inhibitory regions that span multiple channels and appear smooth are in fact interpolated from a few frequency samples by the spectral blurring inherent in the auditory spectrogram.

Neurons demonstrated typical responses to SFG stimuli. Units that were tuned to frequencies far from any of the *figure* tones had relatively stable responses throughout the stimulus. Some units showed strongly suppressed responses to the *figure* tones while maintaining excitatory responses to their best frequency (BF). Still others were tuned at a frequency equal to a *figure* tone and hence showed strong excitation in response to *figure* tones. The reasons these STRF features emerge is because the *figure* tones are correlated during the *figure* presentation, and hence it is not possible to disambiguate responses driven by any one of the tones. Therefore, we examined the response field during the post-STRF period to identify any tuning modulation driven by the *figure* presentation (see text related to Figure 3). Previous studies have shown that attention-dependent plasticity can persist for minutes or up to hours,[17] so the 1-s window immediately following the *figure* should retain some of the modulatory effects.

STRFs were computed by standard procedures detailed in.[15,16] The dimensions of the STRF in this study typically consisted of 37 frequency (y axis) and 13 lag-bins (x axis), referred to as voxels. To test for STRF significance, we generated a null distribution of voxel values using the same procedure as described above for STRFs, but mismatching responses and stimuli. We then down-selected the 128 frequency bins in the auditory spectrogram by removing all channels which did not correspond to the pure tone frequencies in the stimulus. A Wilcoxon paired t-test was used to select voxels from the data with significant distributions ($\boldsymbol{p} < 0.05$). Summing over the 37x13 (frequency x time) bins in the response map yielded mean correlation values for each of the 5 epochs described above (see text related Figure 3). Subtracting the *pre-figure* STRFs from the *post-figure* STRFs yields a summary value for the size and direction of change in response - stimulus correlation (see text related to Figure 3). We also averaged over the response from 20 to 80ms lag to select the strongest response area corresponding to the BF of the isolated cell.

On a population level, we found that the degree of change between *post-figure* and *pre-figure* increased systematically with *figure* size. To quantify all these changes, we defined the strength of the STRF in each epoch (*Pre, Post, Early, Mid,* and *Late*) as follows:

$$S_{pre} = \Sigma_{f,t}\,(\text{pre-STRF})$$

$$S_{early} = \Sigma_{f,t}\,(\text{early-STRF})$$

$$S_{mid} = \Sigma_{f,t}\,(\text{mid-STRF})$$

$$S_{late} = \Sigma_{f,t}\,(\text{late-STRF})$$

$$S_{post} = \Sigma_{f,t}\,(\text{post-STRF}),$$

where the sum is over all *significant* voxels in the STRF, regardless of their signs. The strength therefore may be negative or positive depending on the inhibitory and excitatory fields of the STRF. Differences between the STRFs over a population of cells were judged to be significant if they exhibited distributions with shifted means and $\boldsymbol{p}<0.05$ in a one sample t-test. The resulting distributions are shown throughout the manuscript, especially Figures 4 and 5. For example, in Figure 4D, the shift in 4-tone *figure* distribution is **n.s.**, This *figure* is perceptually difficult for humans to detect. For 10-tone *figures*, which produce a strong perceptual pop-out,[10,18,24] the curve is right-shifted, with a peak in the positive region. In Figure 5, the STRF strengths are plotted directly to measure the dynamics of the changes.

Functional ultrasound imaging

Functional Ultrasound imaging (*fUS*) is a recording technology based on blood flow imaging.[20] It is a variation of Doppler-based ultrasonic imaging that measures the ultrasonic energy back-scattered from red blood cells (proportional to blood volume) in each voxel of the image. The technique is widely used in medical imaging - but major innovations have turned it into a powerful tool for neuroscience such as an ultrafast imaging scanner able to acquire thousands of images per second and filtering to remove global coherent signals.[21,22] These developments have significantly boosted signal-to-noise ratio (SNR) and increased spatial resolution to $\sim$100μm, allowing us to measure subtle changes in blood ow due to local neuronal activity.

We implanted two ferrets (**U**, age <3 years**; Z**, age >5 years) for the *fUS* recordings. Ferrets were first surgically implanted with a stainless steel headpost that was attached to the sagittal interparietal suture and secured in the skull with titanium screws and cement then areas surrounding the auditory cortices were covered with radiopaque bone cement (PALACOS), leaving 1–2 cm^2 cavities for easy access to the auditory cortex in both hemispheres. During surgery ketamine (35 mg kg^{-1} intramuscularly) and dexmedetomidine (0.03 mg kg^{-1} subcutaneously) are used to anesthetize ferrets and 1$-$2% isoflurane to maintain anesthesia also atropine sulfate (0.05 mg kg^{-1} subcutaneously) was used to control salivation and to increase heart and respiratory rates. Electrocardiogram, pulse, and blood oxygenation were monitored, and rectal temperature was maintained at $\sim$38 °C. During surgery the skull was surgically exposed and the areas of auditory cortex were determined and marked on both sides of the skull. Following surgery, antibiotics (cefazolin, 25 mg kg^{-1}subcutaneously) and analgesics (dexamethasone 2 mg kg^{-1} subcutaneously and flunixin meglumine 0.3 mg kg^{-1} subcutaneously) were administered. ferrets were allowed to recover for 2 weeks before being habituated to a head restraint in a customized horizontal cylindrical holder for a period of 2 weeks. Then, the next surgery for the fUS window was performed using a surgical micro drill to remove cement and skull bone, yielding relatively large craniotomies over the auditory cortex $\sim$15 × 10 mm window over the brain. After clean-up and antibiotic application, the hole was sealed with an ultrasound-transparent TPX cover, embedded in an implant of radiopaque bone cement (PALACOS). Sterile polysiloxane impression material (EXAMIX NDS; GC America, Inc.) was placed in the wells between recording sessions, which allowed the craniotomies to be kept well-protected from the environment and were cleaned and treated with topical antiseptic drugs (povidone-iodine). The skin surrounding the implant was cleaned three times per week with warm saline and treated with povidone-iodine and sulfadiazine cream ointment.

Ultrasound Imaging

All acoustic stimuli were presented at 65–70 dB SPL. Sounds were digitally generated at 40 kHz with custom-made MATLAB functions and A/D hardware (PCI-6052E; National Instruments) and presented with a free-field speaker positioned 30 cm in front of the animal's head. During imaging sessions, the ultrasonic probe was placed in contact with the implant cement and acoustic coupling was assured via degassed ultrasound gel. Experiments were conducted in a double-walled sound attenuation chamber. We used a custom miniaturized probe (15 MHz central frequency, 128 elements) inserted in a four degree of freedom motorized setup. The probe was driven using a stereotaxic manipulator with a custom-made holder to control its position.

Imaging

Functional Ultrasound (FUS) imaging has rapid acquisition of 300 frames at a 500 Hz frame rate (lasting 600ms, that is one to two ferret cardiac cycles) that are filtered to remove tissue motion from the signal using a spatiotemporal clutter filter, each frame being a compound frame acquired via 11 tilted plane wave emissions ($-10°$–$10°$ with 2° steps) fired at a pulse-repetition frequency (PRF) of 5500 Hz. And the Image reconstruction is performed using GPU-parallelized delay-and-sum beamforming. A CBV image is averaged

every second to capture the dynamics of the cerebral blood physiological response. High sampling rate is a key asset to cancel any respiratory or tissue pulsatile motion artifacts in the final averaged images.

Power doppler is then computed for each voxel (100 x 100 x ~400 μm, over the 300 time points which is proportional to blood volume. A certain band of Doppler frequencies can be chosen before computation of the power using a bandpass filter enabling the selection of a particular range of axial blood flow speeds such as a slow blood flow of capillaries and arterioles vs. fast blood flow of big vessels. We focused on small vessels with axial velocity lower than 3.1 mm s^{-1}. Power UfD signal was normalized toward the baseline to monitor changes in Cerebral Blood Volume (%CBV).

The imaging progressed by first positioning the ultrasound transducer probe at plane 1 as illustrated in Figure S8. We then ran the SFG stimulus as described above for a total of 50 repetitions. In each presentation, the stimulus consisted of a pre-*figure* (5s), *figure* period (5s), and a post-*figure* (5s), ending with silence for 5s. The probe was then moved 400μm to Plane 2 and the same procedure was repeated. The process was repeated up to 6 or 8 planes as necessary. The localization of the fUS images on the auditory cortex is based on the anatomical vasculature of the auditory cortex (e.g., Figures 6A and 7A) which provide familiar landmarks based on our previous publications.[19,20,22]

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

All statistical analyses were performed with MATLAB (MATLAB 2010a-2024a, Mathworks). We used one-sample t-tests to compare data with matched samples. When data did not meet the requirements for a t-test, we used the Wilcoxon signed-rank test (a non-parametric version of a one-sample t-test) instead.

ADDITIONAL RESOURCES

Further information and requests for resources and reagents should be directed to and fulfilled by the Lead Contact, Shihab Shamma (sas@umd.edu).