Neural integration of acoustic statistics enables detecting acoustic targets in noise

A. Alishbayli[1,†], K. Przewrocki[1,†], P. van Heumen[1], B. Englitz[1,✉]

[1] Computational Neuroscience Lab, Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

[†] Contributed equally ✉ Corresponding author

Abstract

Sound detection amidst noise presents an important challenge in audition. Many naturally occurring sounds (rain, wind) can be described and predicted only statistically, so-called sound textures. Previous research has demonstrated the human ability to leverage this statistical predictability for sound recognition, but the neural mechanisms remain elusive.

We trained mice to detect vocalizations embedded in sound textures with different statistical predictability, while recording and optogenetically modulating the neural activity in the auditory cortex. Mice showed improved performance and neural encoding if they could sample the statistics longer per trial. Textures with more exploitable structure, specifically higher cross-frequency correlations improved performance, background encoding and vocalization activity. Activating parvalbumin-positive (PV) interneurons had an asymmetric effect, improving detection and neural encoding of vocalizations for low correlations, and impoverishing them for high cross-frequency correlations.

In summary, mice exploit stimulus statistics to improve sound detection in naturalistic background noise, reflected in behavioral performance and neural activity, relying on PV interneurons for temporal integration.

Short title: Exploiting Statistical Regularities in Noise

Teaser: The brain enhances sound detection by integrating statistical regularities in the background noise with the help of PV cells.

Highlights

- Mice integrate statistical information indicated by behavior and neural activity
- Encoding of background sounds stays stable in A1, while vocalizations are enhanced
- High cross-frequency correlations improve target detection and neural encoding
- Activating PV cells improves detection of sounds with low cross-frequency correlations

Keywords: sound textures, primary auditory cortex, hearing in noise, interneurons

Introduction

The perceptual environment of animals is structured at multiple hierarchical levels and characterized by statistical regularities. Many theories of neural function postulate that the nervous system has evolved to exploit such regularities efficiently and the capability to reduce redundant information may optimize energy expenditure and survival (*1–4*). The suppression of redundant information usually manifests itself through the process of neural adaptation to constant, repetitive, or predictable stimuli (*5*). However, the details of circuit-level mechanisms that underlie this ability remain unclear.

In the auditory domain, one area where the functional role of adaptation has been prominent is in the context of a well-known psychological phenomenon known as the cocktail party effect (*6*, *7*): here, the signal-to-noise ratio is typically poor, but we are still able to attend to important information by filtering out irrelevant noise. This feature of the auditory system is thought to rest on its ability to adapt to sounds at multiple stages of processing, from cochlear hair cells (*8*) to the auditory cortex (*9*); for review see (*10*). Along these processing stages auditory responses exhibit progressively more noise invariance (*11*, *12*). Recent studies into the neural basis of such noise suppression have highlighted the auditory cortex as an important processing stage, where in particular GABAergic interneurons may play separate but complementary roles due to the differences in their connectivity and physiology (*13–17*).

Traditionally, adaptation to background noise has been studied using artificial stimuli such as white noise and pure tones that give researchers tighter experimental control (e.g. (*14*)). Such basic stimuli fail to shed light on features that are present in complex naturalistic stimuli. A promising set of sounds in this respect are sound textures (*18*). A sound texture is composed of a limited set of sound elements that overlap in time and frequency and exhibit time-invariant statistics and can be represented compactly in terms of their time-averaged statistical parameters (*18*). These parameters can then be used to synthesize textures that are perceived as similar to the originals by human listeners, suggesting that the statistics are sufficient to account for the perceptually relevant information (*19*). This raises the possibility that the auditory system may also represent such statistical regularities and use them to suppress the response to them (*20*, *21*) on the level of the auditory cortex. This could provide the neural basis for improved detection and recognition of embedded target sounds (*22*). However, most previous research in this domain has investigated sound textures in anesthetized or passive conditions, usually using non-parametric stimuli and outside the most relevant task context, i.e. an animal actively detecting target sounds amidst a challenging acoustic background (*12*, *23–28*).

Here, we study the neural encoding of conspecific target sounds in the context of a background sound texture including animal behavior and cell-type specific optogenetic modulation. We trained mice to detect mouse vocalizations embedded in sound textures with parametrically controlled statistics and recorded the responses of neural populations in the primary auditory cortex, which has previously been shown to be required for challenging target-in-noise detection (*29*). We hypothesized that if the auditory system had more time to sample the statistics of the sound textures, the detection and neural encoding of the target signal should improve. We find that mice can indeed detect vocalizations better if they had more time within a trial to sample the sound texture, consistent with a benefit of prolonged exposure to the statistical properties in this trial. Alongside, the neural encoding of vocalizations and their decodability improved with longer texture exposure inside a trial. The encoding of the acoustic textures transitioned from being based on the details of a realization to the sound statistics, matching with human behavior (*30*).

Next, we varied the correlation between frequency channels - a major source of redundancy in natural sounds - in the sound textures to check if the behavior and neural activity in response to the target sound is sensitive to this statistical property. Mice showed better performance for high cross-frequency correlations (CFC) conditions, analogous to classical studies with comodulated maskers (CMR, (*31*, *32*). We find a double-dissociation of modulating PV interneuron activity: increasing PV activity improved both detection behavior and neural encoding for low CFCs contexts, while the opposite occurred in high CFC contexts. Activation of SST interneurons did not substantially change the behavior or neural encoding. The present results highlight the processing of complex natural stimuli in the auditory cortex and identify a specific role for PV interneurons in narrow-band processing of stimulus statistics, while suggesting that they are not constructively involved in wideband modulation processing, such as in CMR.

Results

We trained mice (transgenic PV::ChR2, n=7; SST::ChR2, n=3) to detect vocalizations embedded in sound textures in order to understand the dynamics of adaptation to such acoustic contexts and how long-term statistics affect the processing of textural sounds in the auditory cortex. On each trial, a sequence of conspecific, chevron-type vocalizations (*33*, *34*) embedded in a background sound texture was embedded after a period of exposure to background noise alone (**Figure 1A**). A set of carefully designed acoustic textures was used, which differed parametrically in their marginal distribution in frequencies and cross-frequency correlations (see Methods & Supplementary Figure 1). These statistical properties could be varied largely independently (*18*). Texture stimuli were drawn randomly per trial and served as challenging background sounds for the vocalization detection task. Vocalizations were presented at pseudorandomized spectral locations, to prevent the animal from a selective listening approach. The duration of pre-exposure was randomly varied from 0 to 6 seconds to make vocalization onset unpredictable in any given trial. Licking during the vocalization presentation was considered a hit trial and rewarded with water, via a water spout positioned close to their mouth. Pupil diameter and neural activity from the primary auditory cortex (A1) were recorded concurrently. Peristimulus time histogram (PSTH), stimulus reconstruction and nonlinear decoding were used to understand the sensitivity of the system to statistical features of the background textures (**Figure 1B**). PV and SST interneurons expressing channelrhodopsin-2 (ChR2) were identified using opto-tagging and stimulated in a random subset of the trials to investigate their role in the encoding of textural statistics (**Figure 1C**). The main property of the background textures that was investigated in this study was cross-frequency correlations which was varied independently from the frequency marginal statistics in different trials (**Figure 1D-G)**.

Mice learn to detect conspecific vocalizations embedded in textural noise

Head-fixed animals were trained for 4-6 weeks in the behavioral task, before being implanted with chronic electrodes. We adopted an individual training program for the mice to assist them in learning the task: multiple parameters of the paradigm were adapted to gradually make the task more difficult over time, e.g. by reducing the signal-to-noise ratio of vocalizations over background, extending the range of random onset times and shortening the response window. This training strategy prevented frustrating the animals early on and guided them to focus on the intended response strategy (detect vocalizations) over alternative response strategies (e.g.

guessing or average-timing-based responses). We increased the difficulty such that the hit rate increased to and remained around 20-30%, thus keeping the task challenging for the mouse to motivate learning (*35*). To verify that each animal was responding to the vocalization, we estimated the chance performance by recomputing the chance hit rate in surrogate data, created by permuting the trial order (N=100) while keeping the original lick times and then taking the mean of the resulting distribution of chance hit rates (μ_{rnd}). A session in which the animals performed significantly above chance, was defined as one where the actual hit rate was ≥ [$\mu_{rnd} + 2\times\sigma_{rnd}$] (red dots in Fig. 2C; gray indicates insignificant). The performance ratio between actual and random performance served as an index for the learning progress (**Figure 2C**). Using a linear discriminant analysis (LDA), we found that the significant and insignificant performance ratios on the single recording level were best separated by a threshold performance ratio of >1.5. Using another LDA on the same data but separating along session number, we found that the group of animals require ~22 sessions when a greater fraction of sessions are significantly above chance. However, we also observed considerable variability across animals in the number of recordings required to reach the desired level of performance (**Figure 2D**). Together these metrics indicated that mice are able to learn to detect vocalizations inside a textural background occurring at a random time.

Neural population response and pupil dynamics reflect the behavioral state

The task performance confirmed that the mice learned to actively listen for the vocalizations inside the textural background sound. Next, we checked whether the neural activity in the auditory cortex and the pupil diameter also reflected this engagement. Recordings were carried out both under active and passive listening conditions using identical acoustic stimuli. In the passive condition, the reward spout was moved away from the animal, hence, no lick response was registered and no reward was subsequently given.

Both the population response and pupil diameter exhibited an increase at the sound onset and elevated response following the initial adaptation period (**Figure 3A-B**). Neural and pupil data were normalized to their spontaneous average per neuron and animal, respectively (see Methods for details, and **Supplementary Figure 2** for alignment-based normalization). Active listening resulted in heightened responses for both metrics in comparison to passive listening (**Figure 3A,** aligned to sound onset: average normalized $\Delta FR_{Active\text{-}Passive}$=0.193; n=3621 matched units; Wilcoxon signed rank test, $p=2.53\times10^{-95}$, Cohen's *d*=0.23; average normalized $\Delta Pupil\varnothing_{Active\text{-}}$

$_{Passive}$=0.002; 2-way ANOVA, for behavioral state, df=1, F=303.9, p=4.74×10^{-68}, Cohen's *d*=0.01; **Figure 3B,** aligned to the vocalization onset: $\Delta FR_{Active-Passive}$=0.135, Wilcoxon signed rank test, p=7.36×10^{-49}, Cohen's *d*=0.18; $\Delta Pupil\varnothing_{Active-Passive}$=0.003, F=140.79, p=1.80×10^{-32},Cohen's *d*=0.02). When aligned to the time of the first lick (**Figure 3C**, active trials only), the population activity increases before the lick, while the pupil diameter remains unchanged. After the lick, the pupil diameter increases directly and substantially, indicating a close relation to the lick and the subsequent arousal (*36*).

Next, we analyzed the data separately for different trial outcomes in active sessions. Similar to the differences observed between active and passive conditions, in trials where the animals were more engaged with the task (Hit and FA) the average sound evoked firing rates were higher in comparison with Miss trials, i.e. when the animal responds after vocalization offset (**Figure 3D,** top row, Kruskal-Wallis test, p=5.16×10^{-5}). This difference was even more substantial in the case of the response to the vocalizations, where in miss trials the response was nearly absent (**E,** top row, Wilcoxon signed rank test between hit and miss trials, p=3.85×10^{-20}, Cohen's d=0.10).

Changes in pupil diameter also exhibited dependence on the trial outcome, but the time course of pupil diameter differed from that of the population PSTHs in important ways. While changes in pupil diameter in response to sound onset displayed a similar magnitude distinction between trial outcomes (**Figure 3D,** bottom, 2-way ANOVA, for outcome types, df=2, F=1.46×10^{3}, p=0), vocalization onset-aligned pupil responses point to a closer relationship between pupil diameter and motor behavior or its relation with arousal (**Figure 3E,** bottom, F=717.8, p=1.46×10^{-311}). After the initial peak in pupil diameter following sound onset, pupil diameter in FA trials remained elevated, while on hit and miss trials the pupil size went back to pre-trial levels. This likely reflects arousal after licking early in FA trials (**Figure 3E,** bottom, yellow). The average reaction time of mice in hit trials was 0.56 s (relative to first vocalization onset), which coincides well with the timing of the rise in pupil diameter in the same trials (**Figure 3E,** bottom, blue), consistent with earlier studies (*37*, *38*). On miss trials, the animals fail to respond on time to the vocalization, which is also consistent with the absence of arousal and thus flat pupil response observed in these trials (**Figure 3E,** bottom, red). The relation with arousal is further corroborated by aligning the responses to the first lick: While magnitude differences of population response under different trial outcomes are preserved (**Figure 3F,** top), the increase in pupil diameter only starts *after* the lick (thus effectively ruling out motor preparation as a source) and is larger in hit

trials, where the reward is likely to increase arousal (**Figure 3F**, bottom, 2-way ANOVA, for outcome types, df=2, $F=3.61\times10^3$, p=0).

Taken together, these findings show that average neural population response and pupil dynamics can distinguish between behavioral states in our paradigm. Consistent with previous research on detection tasks in the context of statistically defined stimuli, we observe general adaptation in response rate but also an increase in neural response prior to decision-making (*39*, *40*), while the pupil dilation appears to primarily reflect arousal. The increase in population activity before the lick may reflect the integration of evidence about the presence of the vocalizations (*39*, *40*), a change in attention due to the detection of a target (similar to (*41*)), motor preparation (*42*) or non-aligned sensory responses. We consider the latter two unlikely, given the long build-up (compared to much shorter reaction times in mice, (*43*)) and similarity of the build-up between trial outcomes (Fig. 3F).

Detectability and neural encoding of vocalizations are improved by continued exposure to texture sounds

Next, we investigated whether the detection performance improved *within* a trial if the animals could listen to the background statistics for longer, as demonstrated previously in human subjects (*39*). For this purpose, we used the time-resolved d' measure, introduced previously (*39*). The analysis relates the fraction of hits, correct rejections and false alarms that occur in a particular time window relative to stimulus onset to compute d' (see **Figure 4A** for illustration, and Methods for Details). In agreement with the above-mentioned study, we observed a significant increase in the detectability of vocalizations as the prior exposure to the textural noise grew longer in duration (**Figure 4B;** Kruskal-Wallis test, p=0.005), suggesting an adaptation to or learning of the stimulus statistics as a basis for more accurate target detection.

We hypothesized that the improvement in detectability might be related to an adaptation of the cortical responses to the background noise which would lead to a “pop-out” of the vocalization-related activity. Reconstruction of the presented sounds from the neural recordings allowed us to observe changes in the encoded information over time and has been used to demonstrate adaptation to background noise in human subjects (*44*). Using a similar approach, we reconstructed sound cochleograms by estimating linear filters from the activity of acoustically responsive cells (see **Figure 4C** for an illustration of this process, see Methods for details; (*45*, *46*)). Briefly, neural responses used for reconstructing the sounds were pooled across animals and selected (395 out of 1353 neurons) based on the signal-to-noise (SNR) of their

spectrotemporal receptive fields (STRF). Data from passive listening sessions were used to avoid interference from licking artifacts. Reconstruction quality was defined as the correlation between the original and reconstructed cochleograms.

We divided the set of stimuli according to vocalization onset times and quantified the quality of reconstruction for vocalization epochs. In agreement with the behavioral data above, the vocalization epochs were reconstructed progressively better as the duration of exposure to the textural background sound increased (**Figure 4D**; Kruskal-Wallis test, $p=4.00\times10^{-4}$). While the background reconstruction appears in *absolute* terms better than the vocalization reconstruction, this might be a consequence of the reconstruction kernels being trained on texture-only data (see Discussion).

To validate this outcome using an alternative decoding method, we trained a binary support vector machine (SVM) classifier to distinguish the population activity vector at each time point as originating either from vocalization or from background-only intervals (**Figure 4E**). Subsequently, we assessed the classifier's accuracy in correctly identifying population activity during the vocalization epoch (spanning a total of 2 seconds) as a function of vocalization onset time (**Figure 4F)**. A two-way ANOVA was performed to examine the effect of time within the 2 second-long vocalization epoch and vocalization onset time of the trial. The analysis revealed a significant main effect of time within the vocalization epoch ($F=1.64$, $p=4.66\times10^{-8}$), reflecting the adaptation to the repeated presentation of vocalizations. Importantly, we also observed a stronger main effect of vocalization onset time ($F=97.07$, $p=7.72\times10^{-159}$): Analogous to the reconstruction analysis previously outlined, this approach also demonstrated a gradual improvement in classification accuracy as the duration of prior exposure to sound textures within a trial increased.

Time-dependent improvements in both detection performance and the neural coding of the vocalizations themselves suggest that the auditory system of the mouse can integrate information over a timescale of multiple seconds to optimize the detection of an acoustic target stimulus in noise. These results align with the conclusions from human uECoG data that showed an improved encoding of the acoustic features of speech following an adaptation period after a switch in acoustic statistics (*44*).

Behavior and neural activity improve at high cross-frequency correlations

Previous studies have shown that the auditory system is sensitive to long-term statistical regularities present in sound textures (*18*, *30*). In particular, cross-frequency correlations have been shown to significantly improve the texture identification performance in human subjects,

suggesting that this statistical dimension carries important information about natural sound textures. Having established that animals can benefit from prolonged exposure to the textures in our paradigm, we varied the strength of marginal, pairwise correlations between frequency bands while keeping the marginal statistics of sound envelopes the same. The latter ensured that vocalizations could not be detected more easily based on sound properties other than CFCs. As the primary auditory cortex has been suggested to be involved in the processing of cross-frequency modulations (*47*, *48*), it constituted a natural first target to study the sensitivity of the auditory system to the correlational structure of the sound textures. By probing CFCs we also step beyond the consideration of marginal statistics only (*39*).

At a behavioral level, we observed that the animals were able to detect vocalizations that were embedded in high CFC textures more readily than those embedded in low CFC sounds (**Figure 5A**). This was true when the data were analyzed at the recording and the individual animal level (at recording level, Wilcoxon signed rank test, n=112, $p=4.64\times10^{-11}$, Cohen's d=0.78) at the animal level, Wilcoxon signed rank test, n=7, p=0.0156, Cohen's d=1.48). The neural population exhibited significantly less adaptation to high CFC textures, when aligned to trial start (**Figure 5B**, norm. $\Delta FR_{High\text{-}Low}$=5.9%; n=2065 matched units; Wilcoxon signed rank test, $p=1.79\times10^{-19}$), although the effect size (Cohen's d=0.066) was much smaller than one might have expected due to the increased variability of the time marginal. The correlation between neurons was also higher in high CFCs compared to low CFCs (0.008 vs 0.011, respectively, p=0.0047, Wilcoxon signed ranks test). Pupil diameter also exhibited small but consistent divergence between the two conditions after the initial dilation related to sound onset (**Figure 5B**; $\Delta Pupil\varnothing_{High\text{-}Low}$=0.002, 2-way ANOVA, for CFC level, df=1, F=233.8, $p=8.63\times10^{-53}$, Cohen's d=0.002), likely reflecting the larger fraction of trials in which a reward was received (see above).

Population response aligned to vocalization onset exhibited a significantly larger response to vocalizations when they were embedded in high CFC textures (**Figure 5C** top; onset response (0-1 s): norm. $\Delta FR_{High\text{-}Low}$=10.1% ($d$=0.11, $p=2.6\times10^{-30}$); overall (0-2 s): $\Delta FR_{High\text{-}Low}$=5.7% ($d$=0.068, $p=8.3\times10^{-14}$); Wilcoxon signed rank test, n=1895 vocalization responsive cells, see Methods for details). The pupil response showed a significant and substantial increase in the high CFC as well (**Figure 5C,** bottom; average $\Delta Pupil\varnothing_{High\text{-}Low}$=0.018, $F=8.64\times10^{4}$, p=0, Cohen's d=0.18). If the neural response was normalized to the pre-vocalization period, the difference between low and high CFC was qualitatively and quantitatively similar as well as statistically significant (onset response (0-1 s): norm. $\Delta FR_{High\text{-}Low}$=5.4% ($d$=0.05, $p=5.65\times10^{-15}$); overall (0-2 s): $\Delta FR_{High\text{-}Low}$=5.9% ($d$=0.066,$p=1.79\times10^{-19}$)).

Together, this demonstrates that vocalization detection is aided by high CFCs of the background textures, as evidenced by improved behavior and higher vocalization related neural responses, in accordance with studies on comodulated noise (*49*). Importantly, an arousal increase, as indicated by the pupil dilation, is unlikely to lead to increased performance, but has been shown to reduce accuracy (*50*).

To understand the potential difference in information content of neural responses that underlies the behavioral performance we again used reconstruction analysis as well as support vector machine (SVM) classification on neural data. Reconstruction of background-only and vocalization epochs (i.e. vocalization + background) revealed that high CFC backgrounds were reconstructed better (**Figure 5D;** Wilcoxon signed rank test, n=64 stimuli per condition, $p=1.85\times10^{-4}$, Cohen's *d*=0.63) while there were no significant differences in the reconstruction of vocalization epochs across these two types of background texture (**Figure 5E;** Wilcoxon signed rank test, n=72 per condition, p=0.862, Cohen's *d*=0.085). This was somewhat surprising, given that the detection performance was better in the context of high CFC sounds. Since reconstruction analysis is based on a subset of cells with the highest SNR of STRF, we hypothesized that the information relevant to behavior might be carried by neurons not included in reconstruction. To test the effect of excluding units with low SNR in STRFs, we trained a binary SVM classifier to classify the population activity vector for each time point as either vocalization (1) or no vocalization (0, i.e. background only). We averaged the probability of classification as vocalization across all 10 repetitions of vocalizations to obtain the classifier performance for each background condition (**Figure 5F**). Here, we find a significant difference that was aligned with the animal's behavior, with the classifier having a higher probability of detecting vocalizations when they were embedded in the high CFC background (2-way ANOVA, for CFC, df=1, F=6.44, p=0.011; for time, df=199, F=5.14, $p<\epsilon$, Cohen's *d*=0.147). This result could either highlight a role for cells that do not have sharply defined STRFs, or the ability of the SVM to decode based on nonlinear relations between cells, in representing sounds, or conspecific vocalizations in particular, in A1.

PV cells play a specific role in the processing of cross-frequency correlations

It has previously been suggested (reviewed in (*51*, *52*) that inhibitory interneurons could play an important role in highlighting a target sound in the context of an acoustic background, however, that this remained underexplored in the context of complex sounds. To study the contribution of inhibition to target detection in our paradigm involving complex textural sounds, we

312 optogenetically manipulated the activity of inhibitory PV and SST interneurons. Optogenetic
313 manipulation was performed in 30% of the trials. We titrated the light amplitude to limit the
314 reduction in firing rate to ~25% in cells that were not directly activated by light stimulation, in order
315 to remain within a physiological range of network activity, i.e. to avoid completely silencing the
316 response.

317 To analyze the outcomes of optogenetic manipulation, we classified recorded units as PV
318 if their activity was significantly increased as a result of light stimulation (**Figure 6A;** (*53*)). While
319 PV units with significant light response had spike shapes that were reminiscent of fast-spiking
320 (FS) cells reported in the literature (sometimes also referred to as narrow spiking cells), we did
321 not observe one-to-one correspondence between PV and FS cells **(Figure 6B)**. Therefore, we
322 used these three partially overlapping groups of neurons to characterize the population response.
323 The time course of sound-evoked response was similar across these cell types, but the average
324 response of PV cells within the first 0.5s of sound presentation was significantly stronger than
325 the other groups (Kruskal-Wallis test, p=0.001, **Figure 6C**). Light stimulation increased the sound-
326 evoked response for PV cells (median=8.88%, Wilcoxon signed rank test, $p=3.54\times10^{-5}$), and
327 decreased it for both FS cells (median =-3.84%, Wilcoxon signed rank test, p=0.0196) and non-
328 FS cells (median=-11.85%, Wilcoxon signed rank test, $p=1.16\times10^{-189}$, **Figure 6D**).

329 The effect of optogenetic manipulation of PV was found to have opposing effects,
330 depending on the type of background texture (**Figure 6E-F**). For high CFC sounds, activating PV
331 cells impaired the behavioral performance (Figure 6E, at the recording level, Wilcoxon signed
332 rank test, p=0.0025; **Figure 6F**, at the animal level, Wilcoxon signed rank test, p=0.0313), while
333 conversely performance under low CFC sounds was significantly improved both at the animal
334 level (**Figure 6F,** Wilcoxon signed rank test, p=0.0156) and the recording level (**Figure 6E**,
335 Wilcoxon signed rank test, p=0.0062) as a result of this manipulation. Coarse features of the cell-
336 type specific PSTHs were comparable across low and high CFC conditions, suggesting that the
337 behavioral outcomes were not captured in any straightforward manner by population averages
338 (**Figure 6G**). Therefore, to better understand the changes in the information content of neural
339 responses, we used the SVM classifier to analyze the neural responses during the passive
340 listening condition. As before, SVM evaluations mirrored the behavioral results analyzed at the
341 animal level, where light stimulation improved the detection performance under low CFC
342 background (2-way ANOVA; for light state, df=1, F=221.62, $p=3.54\times10^{-34}$; for time, df=199,
343 F=3.99, $p=2.63\times10^{-21}$) and impaired the performance under high CFC condition (2-way ANOVA;
344 for light state, df=1, F=89.55, $p<\epsilon$; for time, df=199, F=1.62, $p=3\times10^{-4}$, **Figure 6H, I**).

Although the asymmetric effect of PV stimulation suggests a specific role played by PV interneurons in the processing of textures with high CFC, the outcome may still be due to the nonspecific effect of inhibition itself, rather than inhibition by PV cells per se. To evaluate this, we performed the same experiments on mice that expressed ChR2 in SST interneurons. While analyses involving no optogenetic stimulation trials showed comparable results (**Supplementary Figure 3C,F**), at neither behavioral nor animal levels, the SST dataset showed any asymmetric behavioral effect of optogenetic activation observed in the PV dataset (**Supplementary Figure 3D,E**). As was the case in the collection of the PV dataset, we titrated the intensity of the light to achieve a 25% reduction in population average response (compare to **Figure 6G** for non-FS cells) and used the maximum light intensity when the optical titration results were not conclusive. While this strategy led to effective control over population response in the PV dataset (Figure 6G, right column), we observed that the optical stimulation of SST cells (median=1.33%, Wilcoxon signed rank test, $p=0.677$) led to a weaker but significant inactivation of FS (median=-2.93%, Wilcoxon signed rank test, $p=3.31\times10^{-5}$) and non-FS cells (median=-6.93%, Wilcoxon signed rank test, $p=1.54\times10^{-51}$) (**Supplementary Figure 3B**). Therefore, even though differences between SST and PV datasets are strongly suggestive of cell-type specific roles for PV interneurons, we cannot fully exclude the possibility that the effect may be due to the overall level of inhibition.

Neural coding transitions from samples to statistics within the trial

Lastly, we analyzed whether the neural responses contain more information about the specific realization or its long-term statistics, and in particular the contribution of different cell-types. As the statistical estimation appears to improve over time in a given trial, we hypothesized that the neural coding would transition from the specific realization to textural statistics over time. Extending an analysis in the visual system (*54*), we used a nested ANOVA to partition the total variance of neural responses into texture class and realization components: Texture classes were defined by their marginal and correlation statistics, while realizations corresponded to different segments of the synthesized sound clip belonging to a given texture class. Nested ANOVA can model the hierarchical relationship between realization and texture classes, and allows us to evaluate how sensitive the cells are to these features of the sound textures. Our goal in this analysis was two-fold: 1) how the explained variance (VE) for each component changed over time during the adaptation period, and 2) how the length of the analysis window affected the VE.

In order to address this question in a principled manner, we used moving average responses and performed nested ANOVA for each cell and time point, using different window

377 sizes, which varied from 30 ms to 1000 ms in length **(Figure 7A**). We performed nested ANOVA
378 separately on the three groups of cells (PV, FS, non-FS) introduced earlier. We then averaged
379 the VE across cells to obtain the VE for the respective cell type population. While the population
380 VE remained constant over the time window analyzed here (**Figure 7B-D,** left column), extending
381 the sliding window length increased the texture class VE without improving the realization VE
382 (**Figure 7B-D,** right column). With smaller sliding window averaging, VE by realization was greater
383 than that by texture class for all cell types. However, when the responses were averaged over a
384 longer period of time, the VE due to texture class increased significantly. Two-way ANOVA with
385 cell type and window lengths used as factors for variance components identified cell type to be a
386 significant factor for realization VE (two-way ANOVA, $F=38.81$, $p=1.76\times10^{-17}$), but not for texture
387 class VE (two-way ANOVA, $F=1.99$, $p=0.136$). Window length had a significant effect for both
388 sources of variance (two-way ANOVA, $F=362.1$, $p<\epsilon$, for texture VE; $F=66.35$, $p=4.07\times10^{-55}$ for
389 realization VE). These results show that the sensitivity of the population response in the auditory
390 cortex varies based on the timescale of the analysis used and highlight the importance of the
391 choices that are made in analyzing the sound responses where the temporal aspect is very critical.
392 They also suggest that primary auditory cortical cells may be using temporal multiplexing when
393 representing the noise, instead of converging on time-averaged statistics alone (*55*).
394 To further characterize the response variability, we performed two-way ANOVA using
395 marginal and CFC statistics as factors, because these variables are not independent of one
396 another (hence, not nested). Similar to the analysis above, the time course of VE by each of these
397 factors did not show a strong pattern within the window analyzed here (**Figure 7E-G**, left column).
398 However, in agreement with the previous analysis, average VE by both marginal and CFC
399 components increased with longer moving average window lengths (**Figure 7E-G**, right column).
400 Interestingly, both cell type (two-way ANOVA, for marginal statistics, $F=11.73$, $p=8.19\times10^{-6}$; for
401 CFC statistics, $F=4.74$, $p=0.0088$) and window size (two-way ANOVA, for marginal statistics,
402 $F=114.26$, $p=2.36\times10^{-94}$; for CFC statistics, $F=63.32$, $p<\epsilon$) significantly affected the amount of VE
403 by both types of texture statistics, suggesting some level of specialization in the type of statistics
404 represented by different groups of cells.

Discussion

We investigated the neural encoding of vocalizations in the context of textural background sounds in the auditory cortex of behaving mice. Both the behavioral performance and the corresponding neural encoding of the vocalization improved with exposure to the textural stimulus in individual trials, indicating a rapid, within trial adaptation to the textural statistics. Sound textures characterized by higher correlations across frequency channels (CFC) enabled the mice to more easily detect the vocalizations compared to the context of low CFC textures. Activation of parvalbumin (PV) interneurons exhibited an asymmetrical impact on behavioral performance for high and low CFCs, which was mimicked by the fidelity of the vocalizations' neural activity, suggesting a specific role for PV cells in the processing of spectrotemporal correlations. These results contribute to our understanding of the neural basis for the well-documented ability of humans to comprehend conversations under adverse acoustic conditions, and in particular recognize and deal with sounds whose structure is only constrained statistically, such as many naturally occurring environmental sounds.

Behavioral performance and adaptation to sound textures

Sound textures are characterized by their stable long-term statistics, and initial studies have shown that humans can utilize and estimate different types of statistics (*18*, *30*). These studies found that sampling a sound texture longer enhances the ability to distinguish between different texture classes, while at the same time compromising the ability to distinguish samples drawn from the same statistics, consistent with a neural encoding of statistics. Similarly, longer exposure to sound textures improved the ability to detect changes in statistics (*39*, *56*), phoneme recognition (*44*), and vowel discrimination (*57*). The process of estimation of statistical properties, which is believed to underlie these findings, occurs over multiple seconds, requires no conscious effort, and depends on certain stimulus features such as stationarity (*58*). In our study, texture statistics were changed on every trial, and thus also required the animals to adapt to these statistics within a given trial to optimize their performance. We find that the performance in detecting a vocalization inside texture noise also improves with exposure duration, and is thus consistent with an estimation process (Fig. 4). However, we did not observe plateauing in performance over time, previously observed in human studies (*39*). This discrepancy could be explained by task difficulty, where integration over longer periods might remain advantageous for more demanding tasks.

Neural adaptation to and the encoding of sound textures

The neural encoding of sound textures has not received a lot of attention in previous studies, in particular not in the context of behavior and the detection of embedded target sounds, despite its paramount importance in real-world hearing. Here, we demonstrate that the encoding of vocalizations improves with time into the trial/stimulus, which we quantify through SVM decoding and stimulus reconstruction, respectively, while the quality of the encoding of the background texture remains unchanged (**Figure 4**). From previous human and neural modeling studies (*11*, *44*), we would have expected to find a reduced fidelity of encoding of the background texture, in conjunction with an improved encoding of the foreground. While we find the generally expected adaptation in firing rate (*5*) to the texture stimulus in line with (*44*), we do not find a specific buildup of invariance against the background. Within the limits of the current task, this is hence not directly in line with the general idea that highlighting of relevant information - which we find to be the case - is necessarily accompanied with a reduction of information about background sounds (*46*), at least not in mouse A1. Higher level areas might show different degrees and properties of adaptation to different stimulus contexts (*59*). In addition to the species difference to (*44*), in our case the reconstruction model was not trained on clean speech neural data (not available in our case), but on textural noise responses, which probably improved the overall reconstruction quality for textures at the expense of vocalizations.

On the other hand, the observation that vocalization reconstructions improved over time is consistent with the jointly observed, improved behavioral detection performance, as well as with previous studies into the effect of prolonged exposure to masking noise that showed a stronger signal-evoked response (*48*), improved reconstruction of the speech signal (*44*) and improved detection and recognition of foreground targets in the context of textural noise (*22*).

The most fundamental form of adaptation that occurs in the sensory systems is firing rate adaptation, where the firing rate of neurons to an ongoing stimulus is reduced over time (*60–62*). Across the auditory system the time-scale of this adaptation ranges from 10s of milliseconds (*63*, *64*) to 100s of milliseconds (Fig. 3/5), but does not extend to multiple seconds as observed presently in the within trial performance improvement. Further, response magnitude and subsequent adaptation to the sound in the primary auditory cortex are known to be sensitive to variables such as task engagement (*65–67*) and attention (*68*). In agreement with previous studies, we observed differences in the magnitude of sound-evoked responses between active and passive listening conditions, as well as trial outcomes. These differences were also observable in pupil diameter over time, and likely reflect the aggregate nature of metrics like average population response and pupil dilation (*69*).

472 Further, we observed that A1 responses carry information about different features of the sound
473 at different timescales (**Figure 7**), consistent with the dual timescale encoding of sound observed
474 in previous studies on the perception of sound textures (*39*, *70*). At faster timescales, the variance
475 of all cell types analyzed here is better explained by the realization (i.e. fine temporal details of
476 the sound clip), whereas at the scale of multiple hundred milliseconds, neural responses were
477 better explained by the texture identity, which is a higher-level, time-averaged feature of the
478 sound. This is consistent with the idea that the neural responses may carry information at different
479 timescales, a phenomenon known as temporal multiplexing (*55*, *70–73*). The existence of neural
480 codes on different timescales in A1 can in turn underlie the duration dependence of texture class
481 and exemplar differentiation (*30*), if suitably integrated by higher order areas.

482 Previous work on the sensitivity of different auditory regions to texture statistics in the
483 anesthetized rat has shown that while the inferior colliculus (IC) is remarkably sensitive to most
484 classes of statistics in the McDermott and Simoncelli model (*18*), A1 does not reliably respond to
485 changes introduced to ongoing texture statistics (*24*, *26*). In contrast, we find in the awake mouse
486 that (in particular for long time windows) texture type can explain ~20% of the variance (Fig. 7),
487 and substantially more than different realizations with similar statistics. On the one hand his could
488 be due to the difference in state, i.e. awake vs. anesthetized, or species, i.e mouse vs. rat. In
489 addition, the stimulus design also differed, where Peng et al. studied changes between different
490 textures, while we focussed on the neural encoding of textures in separate trials. Overall, the
491 strong adaptation of cells in A1 to ongoing background sounds (*20*, *23*, *44*) could also limit the
492 ability to decode texture type in both studies, which, however, might be a reflection of a noise
493 reduction strategy in the cortex. In this context, the constant ability to reconstruct the spectrogram
494 over time (Fig. 4D) is puzzling, and may suggest a greater coding efficiency with longer
495 presentation duration.

496 A recent, highly insightful study on active sound detection in noise (*29*), closely linking
497 experimental data to a normative model. Their primary manipulation - switches between low and
498 high contrast textures - led to two different time-courses of adaptation in both behavioral
499 performance and neuronal sensitivity. However, their contrast change does not map directly onto
500 our low/high CFC paradigm, as only low CFC textures are used, limiting a direct comparison of
501 the result. In particular, differences in integration windows (max. 1 vs. 6 seconds), timing
502 (expected vs. unexpected context duration) and behavioral performance measure (percent
503 correct vs. d') complicate a direct comparison. Framing the present paradigm in their normative
504 framework extended to include correlations and longer timescales would be required to fully
505 compare and reconcile the observed adaptation dynamics between the two paradigms.

506

507 *Sensitivity to cross-frequency correlations*

508 On the behavioral level, we observed higher detection performance in high CFC textures
509 (**Figure 5**), which aligned with results on the neuronal level, where we observed slightly stronger
510 adaptation to low CFC stimuli and significantly stronger vocalization-evoked neural response in a
511 high CFC context. The presently observed effect size in the neural response is small, but
512 comparable to previous studies (*47*, *74*), which also highlighted a dependence on the relative
513 level of masker and target, which was set to a fixed relation in the present study. The improved
514 performance for high CFC stimuli cannot be explained by faster convergence statistical properties
515 for these stimuli, which we checked directly on the level of the stimuli for different excerpt
516 durations (see **Supplementary Figure 4**). More importantly, we find a slightly better encoding of
517 vocalizations for high CFC textures, and a strong, asymmetric effect of increased PV activity on
518 the encoding of vocalizations for high vs. low CFC backgrounds. This generally aligns with
519 previous work that demonstrated that the average frequency bandwidth of neurons across
520 successive stages of auditory processing tends to increase from auditory nerve fibers to A1 (*75*).
521 Due to this feature, cells in the cortex might be well suited for integrating information across
522 different frequency bands.

523 Given the pronounced comodulation in the presently used high CFC sounds, the
524 behavioral and encoding improvements could relate to comodulation masking release (CMR),
525 which describes the facilitation of narrow-band sound detection when different frequency bands
526 in the background noise are comodulated in level (*32*). The latter is considered to provide a
527 spectral cue for grouping different parts of the spectrogram which the auditory system could use
528 to segregate non-comodulated target sounds from the background (*76*–*78*). Previous studies
529 have shown that there is evidence on the level of neural responses on multiple levels of the
530 auditory system, ranging from the cochlear nucleus (*79*) to the auditory cortex (*47*, *48*, *80*, *81*).
531 While a rich literature exists on CMR in humans, the limited number of studies in animals have
532 typically not studied behavior and neural encoding together, in particular not in conjunction with
533 cell-type specific modulation as done here (*25*, *47*, *48*, *74*, *81*, *82*), which has left the concrete
534 neural basis of CMR unresolved. Importantly though, the present study only has limited relevance
535 to CMR: while the behavioral and neural improvement in the high CFC condition is closely related
536 to CMR, our main results, i.e. (i) integration of statistics over time, and (ii) role of PV interneurons,
537 are either general to both low and high CFC (i) or contribute constructively to the low CFC (i.e.
538 non CMR) condition (ii). Therefore, they contribute to the understanding of listening in noise, even
539 in conditions when no comodulation of close or distance flanking bands exists.

Another factor contributing to the improved performance in high CFC background could be related to “dip listening”, i.e. the ability to catch acoustic “glimpses”, when the level of noise momentarily drops (*83*). In the high CFC case, the SNR of the vocalization relative to the background fluctuated more, and high SNR moments could have been used to improve detection performance. It is, however, noteworthy that modulated noise does not always aid the recognition of signals in a complex auditory scene, as it can also transiently capture the attention and thus interfere with target recognition (*84*, *85*).

The role of inhibitory neurons in processing background noise

We characterized the role that PV, SST, FS and non-FS cells may play in the processing of naturalistic noise in the primary auditory cortex of the mouse. Identifying cell-types is crucial for understanding the underlying processing, e.g. hearing in complex auditory scenes with competing sound sources (*15*, *48*, *86*). As there remains debate on whether opto-tagging (parvalbumin, PV, and somatostatin, SST) or spike parameters (*87–89*) is the best approach for classification of cell-types, we opted for using both in parallel (**Figure 6**).

While the cellular basis of hearing in the complex auditory scene is only starting to be explored, various fundamental aspects of it such as surround suppression, adaptation to repeated sounds, and the role played in attentional control have been studied previously (see (*52*) for a detailed review). These particular aspects of sound processing are important in real-life audition because natural sound sources usually differ from basic sounds used in auditory experiments by wider bandwidths and more spectral overlap. This poses a challenging problem which requires to filter out interfering noise and selectively attend to relevant sounds. (*14*) showed that while wideband noise drives both SST and PV interneurons, only inactivating SST cells leads to a change in surround suppression observed in excitatory cells. Stimulus history modulates the inhibitory effect of SST cells, while PV cells were shown to provide the same amount of inhibition regardless of the stimulus history (*9*).

At first glance it may seem that this contrasts with our results, namely that only the activation of PV cells influenced the behavioral and the neural activity. However, upon close inspection the results are compatible: the result from (*14*) emphasizes the role of SST cells in integrating *broad-band noise*, and conversely a lack of a role of PV cells for the same task. This aligns with our result, as we find a beneficial role for integrating *narrow-band* noise, as in the low CFC case (see Fig. 8 for a graphical summary). As noted above, high CFC sounds had stronger but randomly timed amplitude modulation, in contrast to the relatively flat time course of low CFC sounds. It makes sense that this might not have been detectable in the aforementioned study,

where narrow-band context was not assessed. Conversely, in the *broad-band noise* case, we find PVs to have a detrimental effect, which we interpret as a shift in the balance of activation between different cell-types, which might interfere with the SST's beneficial effect on broad-band integration. To our surprise, direct modulation of SST neurons did not lead to a behavioral difference in either of the CFC conditions, which raises the question whether the results from (*14*) generalize to textural stimuli. The aforementioned study used basic white noise, while here we used naturalistic textural sounds with greater internal predictability, both in time and across frequencies. Consequently, the improved performance during PV stimulation in low CFC sounds suggests a more direct role in temporal coding, as in the low CFC case within-channel temporal predictability is greater, which aligns with previous findings on the role of PV cells in temporal coding (*86*). The local co-tuning of PV cells with excitatory cells has been suggested to reduce the noise in the neural activity (*86*), which presently translates to subtracting the recent, narrow-band noise. This can form the basis for improved signal-to-noise for the embedded vocalization, and thus provide the basis for the improvement in behavior and neural encoding. Slightly reduced adaptation and improved reconstruction quality for high CFC sounds suggest that in the context of our task, the primary role of PV mediated inhibition may be to shape the population response in order to serve something akin to temporal envelope tracking or neural entrainment (*90*–*92*).

Limitations and future directions

We presently chose to increase the activity of PV/SST cells (similar to other studies; (*93*, *94*), in order to understand their influence on the encoding of textural sounds and subsequent behavioral decisions. Instead one could have silenced these cell-types, which would be more consistent with the classical lesioning approach that removes a part of a system to understand its functional contribution (*95*). While optogenetic modulation using light-activated Cl^- (GtACR, NpHR, etc.) or H^+ (Arch) channels (see (*96*) for overview) would have allowed for this in a transient and reversible manner, it is not clear whether this would have been superior, as it risks moving the system outside of its physiological range (*97*) which complicates interpretation in highly interconnected systems (*98*). In our perturbational approach, we instead aimed to modulate the average activity of the non-PV/SST cells in a limited manner and thus keep the system close to its physiological range. Further, training of the SVM and reconstruction algorithms was always balanced for correlation, light stimulation, and vocalization base frequencies, and should therefore not have biased the quality of decoding towards either high or low CFCs. The observed, interpretable changes in behavior and neural activity for the high/low CFC stimuli seem to validate this approach, however, we would like to point out that it is inherently limited, as the modulation is not

608 specific to individual cells, and depends on the penetration depth of the light. Future studies
609 should improve on this technique by the use of more specific stimulation techniques such as
610 Spatial Light Modulators (SLM) alongside optical imaging (as in e.g. (*99*, *100*)).

611 Optogenetic manipulation of PV and SST cells make a case for a specific role of PV-
612 mediated inhibition in the processing of high CFC noise (see **Figure 6**). However, the behavioral
613 difference between PV and SST may arise from different levels of inhibition achieved in each
614 case. While we used identical light delivery in both cases, there are factors that could have limited
615 the amount of inhibition in the SST case: (1) Physiologically, PV cells form inhibitory networks,
616 wile SST cells avoid each other and form disinhibitory circuits with other interneuron types, which
617 may limit the degree of inhibition achievable (*101*). (2) More PV cells tend to be located in
618 superficial layers than SST (*102*), which simplifies their stimulation in particular in chronic
619 experiments, where scarring may reduce the amount of light that penetrates deeply enough.
620 Future studies utilizing more local light delivery or deeper penetration using 2-photon stimulation
621 could further resolve this question.

622 The current recordings were targeted to the primary auditory cortex, and there is clearly a
623 need to study the corresponding activity on different levels of the system, starting from the inferior
624 colliculus to higher stages of processing, in particular the parietal cortex which has previously
625 been demonstrated to be involved in the integration of sensory evidence (*103*). Chronic widefield
626 and 2-photon optical imaging using more recent calcium indicators (*104*), which provide high
627 temporal resolution and excellent signal-to-noise ratio, combined with targeted optical stimulation
628 are likely the best techniques for further advancing this research, as they allow to resolve the
629 spatial layout of the system with little to no undesired perturbation.

630 Lastly, while the present set of stimuli was already large, reaching the time limits of working
631 with behaving mice, future studies should address additional dimensions in the space of stimulus
632 statistics. First of all, we only studied two cross-frequency correlation levels, and resolving this
633 more finely, or introducing other naturalistic CFC patterns would be very important. In addition,
634 introducing different temporal correlation widths would be another interesting direction to
635 demonstrate the involvement of PV interneurons in temporal integration. Each of these questions
636 would have to be addressed in separate studies to obtain sufficient data for the specific stimulus
637 contrast, comparable to the low/high contrast studied by (*29*).

Methods

Data presented in this study are collected from 7 PV::ChR2 (4 males, 3 females; age at experiment mean = 10.2 weeks, range = 9-13 weeks) and 3 SST::ChR2 (3 males; age at experiment mean=10 weeks, range=9-12 weeks) mice obtained from F1 generation of crossing between either PV-Cre (JAX stock #017320) or SST-IRES-Cre (JAX stock #013044) with Ai32 (JAX stock #024109). 4 animals (2×PV::Ai32 and 2×SST::Ai32) were used for hearing ability tests (see below). Animals were housed in a reversed light/dark cycle environment (12h light/dark), inside the cages where food pellets were available *ad libitum*, while water access was limited to motivate animals to learn the behavioral task (see Details below in Animal Training). Most animals were housed in pairs separated by a perforated transparent plexiglass wall that allowed limited interaction between animals while protecting the implant from damage and allowing per-animal monitoring of water intake. All animal experiments were performed under animal experimentation project 2017-0041 approved by the Centrale Commissie Dierproeven (CCD) and supervised by the Animal Welfare Body of the Radboud University Medical Center.

Surgery

Animals underwent two surgeries with a training period (6-8 weeks) between them. In the first surgery, a cement base holding a headpost was implanted for head-fixed training, while in the second surgery, the electrodes were implanted (see details below). For both surgeries, analgesic treatment began 1 day before the surgery and continued for 3 days post-surgery using oral intake of an analgesic diluted in water (Rimadyl, 5 mg per 100 ml). Animals had *ad libitum* access, with the concentration estimated based on their normal water intake. At the beginning of each surgery, mice received an additional dose of Rimadyl (5 mg/kg, s.c.) and in the second surgery, a dose of dexamethasone (2 mg/kg, s.c.) was used to reduce brain swelling.

In the first surgery, a plastic headpost for head fixation was implanted on the contralateral cranial surface of mice that were 6-8 weeks of age**.** Implantation was performed under isoflurane anesthesia (1.5-2%) while continuously monitoring breathing rate and body temperature. The animal was placed on a thermal blanket and the eyes were covered with eye ointment (Ophtosan, ASTFarma). After shaving the head and prior to making the incision, 0.1ml of Lidocaine/Bupivacaine was administered subcutaneously using a fine syringe needle. After removing the skin over the cranium, the bone was cleaned from tissue using hydrogen peroxide (H_2O_2, 3%). Next, a 3D-printed, plastic headpost (material: polycarbonate, Prusa) was attached

669 to the cranial surface using dental cement (Super-Bond C&B Kit, Sun Medical), positioned
670 approximately midway between bregma and lambda and close to the lateral ridge, leaning
671 laterally at a ~50º angle from the midline. In a subset of animals, a round glass plate was placed
672 over the auditory cortex for intrinsic optical imaging of the auditory cortex (see below).

673 In the second surgery, the mice were implanted with electrode arrays in the primary auditory
674 cortex (A1), whose location was estimated by either stereotactic coordinates (2.8 mm from
675 Bregma, 5.9 mm laterally from sagittal suture; based on (*105*) alone or together with the help of
676 intrinsic optical imaging (see below). A small craniotomy of ~1×2 mm in size was opened and a
677 small amount of biocompatible silicone gel (Dura-Get, Cambridge NeuroTech) was placed on the
678 craniotomy to minimize brain swelling. Electrode arrays (see Details in the section
679 Electrophysiology below) were inserted orthogonal to the brain surface with the prongs aligned
680 along an axis that maximized the coverage of topographically organized frequency space (based
681 on the orientation of area A1 in the map of (*105*)). A ground screw was attached contralaterally,
682 anterior to the headpost, and connected to the ground cable of the recording probe using
683 conductive epoxy (CW2400, Chemtronics).

684 **Intrinsic optical imaging**

685 In a subset of animals (7/10), intrinsic optical imaging of the auditory cortex was performed to
686 locate the primary auditory cortex and improve targeting area A1. During the implant surgery, the
687 mouse was lightly anesthetized with isoflurane (~1%). Pure tones and noise bursts were
688 presented (70 dB SPL) as probe stimuli to activate the auditory cortex. The surface of the brain
689 was monitored optically through a glass window (ø=3 mm, CS-3R, Harvard Apparatus) placed
690 onto a thin layer of cement, centered on the region of the skull that was estimated to be above
691 the auditory cortex based on stereotactic coordinates. The brain surface was monitored through
692 the cement and skull using an sCMOS camera (CS2100M-USB, Thorlabs), receiving light from a
693 combination of a 4x objective (N4X-PF, Nikon) and a tube lens (TTL-100A, Thorlabs). The
694 recording area underneath the glass was first illuminated with a collimated LED (525 nm, M525L4,
695 Thorlabs) to obtain a reference image emphasizing the blood vessels. During functional imaging,
696 illumination was switched to 625 nm (M625L4, Thorlabs) and imaged at a rate of 20 fps. The focal
697 plane was placed roughly 100 µm below the vessel level.

698 The acquired movies were analyzed by a sequence of normalizing and filtering steps: (1) reduce
699 image resolution by averaging over 4x4 blocks, (2) normalize frames for differences in global

700 illumination based on the average intensity outside the craniotomy, (3) baseline correct frames
701 for each pixel by dividing them by the average intensity in the pre-stimulus period, (4) reduce the
702 temporal sampling rate to 10 fps by averaging consecutive frames, (5) average over collected
703 trials, and (6) perform spatial filtering by convolving with a 2D Gaussian kernel (SD = 2 pixels).
704 The resulting response maps were then obtained at the minimal response after the stimulus onset,
705 indicating the time of the largest activity-dependent change in blood flow, volume, and
706 oxygenation in the observed brain regions (*106*). The resulting spatial pattern was then compared
707 with previously established maps of the auditory cortex (*105*) to identify the most likely location of
708 A1 in relation to the local vasculature.

709 **Electrophysiology**

710 All recordings were performed using silicon probes featuring 64 channels (NeuroNexus) of various
711 geometries ($N_{prongs}xN_{sites}$), specifically two 8x8 (A8x8-Edge-5mm-50-150-177, A8x8-Edge-5mm-
712 100-200-177) and two 4x16 (A4x16-Poly2-5mm-23s-200-177, A4x16-2.5mm-tet/lin-300/125-333-
713 121/177-PC). Data was digitized using a combination of 64 channel headstages (Intan
714 Technologies) connected to an OpenEphys recording system (Open Ephys Inc.). The data at a
715 sampling rate of 30 kS/s, band-pass filtered between 0.3-7.0 kHz, and then spike-sorted to obtain
716 single and multiunit responses using autoSortC function included inside the code repository
717 provided along with this article. Electrophysiological recordings were collected over a period of 4-
718 8 weeks, depending on the signal quality. Electrodes were advanced weekly (in steps of 125-
719 250 μm) to cover the entire depth of the cortex over the whole recording period. At the end of the
720 experimental period, animals were disimplanted, and intact probes were cleaned (Tergazyme)
721 and reused.

722 **Optogenetic stimulation**

723 Cell-type specific manipulation of PV interneurons was carried out by photostimulating blue-light
724 activated opsins (ChR2) 470 nm light was delivered using an LED light source (Optogenetics-
725 LED, Prizmatix, maximum light power ~10 mW). Dual fiber optic cannulas (manufactured by
726 Prizmatix) were inserted into the custom-made implant case through tunnels on both sides that
727 lead to the brain surface. The end of the optical fiber was located ~1-1.5 mm above the brain
728 surface. For determining the level of light on a given day, short recording sessions were carried
729 out, where we delivered 100 ms light pulses of varying light intensities, separated by a pause of
730 random duration (1-1.5 s). These sessions were used to opto-tag the cells that were responding

731 directly to the photostimulation. Additionally, the power of the light delivered at each recording
732 was titrated based on these sessions to induce a 25% decrease in population (with PV cells
733 excluded) firing rate as a result of activation of PV interneurons. The goal of this titration procedure
734 was to modulate the influence of PV interneurons in a controlled manner, but not completely
735 silence the auditory cortex, which we expected would abolish the behavior (*29*). During the main
736 experiment, light was provided at the same intensity as above, during a randomly selected but
737 fixed set of trials. If the titration experiment failed to record the expected inverse relationship
738 between the light intensity and population firing rate (often due to scar tissue that developed over
739 time), the light intensity was set to the maximum possible by the device. The light was turned on
740 at the same time as the sound with a ramp of intensity at the beginning and the end of the
741 stimulation to reduce the stimulation artifact.

742 **Pupillometry**

743 The left eye of head-fixed animals was recorded using a high-speed camera (GV-5240CP-NIR-
744 GL, IDS Imaging Development Systems) under collimated infrared (740 nm) light illumination
745 (WLS-22-A WheeLED, Mightex) delivered through a light guide. Images were captured at 50 fps
746 and compressed to MPEG-4 files (quality ratio=0.5) before being further processed. Pupil
747 diameter was estimated using a custom script utilizing pose-estimation tool, DeepLabCut (*107*).
748 The model was trained to estimate positions of 6 points around the pupil perimeter as well as the
749 center of the pupil. These points were then used to compute multiple radius and diameter
750 estimates for each frame. To account for mislocalized points, the final pupil diameter was
751 computed as the median of all measurements. Time points, when an animal blinked, were omitted
752 from the analysis, as the marker values were not available for the corresponding frames.

753 **Facial motion analysis**

754 We conducted a hearing ability test on untrained mice (n=4; 2×PV::ChR2 and 2×SST::ChR2)
755 using the same setup as previously described. The mice were exposed to a randomized sequence
756 of pure tones, while their facial expressions were recorded. The experiment involved presenting
757 pure tones of five different frequencies (2, 4, 8, 16, and 32 kHz) and a silent stimulus as a control
758 condition. Each tone was played for 2 seconds with a random pause between tones lasting from
759 4 to 6 seconds. The mice were exposed to each tone 50 times during the experiment. The videos
760 were recorded and subsequently subjected to singular value decomposition to reduce their
761 dimensionality, as detailed elsewhere (*38*). To estimate the hearing ability of the mice for each

frequency, we identified the peak of the average motion energy time series from the first motion component (**Supplementary Figure 5**). We performed only one session per animal to prevent habituation to the experimental context.

Acoustic stimuli

Acoustic stimuli consisted of a background texture and repeated, synthetic mouse vocalizations. Sound textures were synthesized using the Sound Texture Synthesis Toolbox (*18*) adapted to our particular use case. To generate sounds that vary mainly along a limited set of statistical dimensions of interest, we first measured the respective statistics in a set of naturally occurring sounds provided by (*18*). Using measurements from different natural sounds, we synthesized artificial sound textures that had specific subband means ("frequency marginals") and variances as well as particular cross-frequency correlations (see Fig. 1D-G for a visualization, and **Supplementary Figure 1** for spectrograms of all stimuli). Subband means and variances were modified together, referred to as a *base*. For instance, the subband means came from “Bubbling_water.wav” (which is also referred to as the frequency marginal), the subband variances came from “Drumroll.wav” and the cross-frequency correlation values in the high correlation case were taken from “Jogging_on_gravel.wav”. The complete list of source sounds to constrain different properties of the textures is provided in **Table 1** and are provided in the code and data repository (due to size restrictions for the supplementary audio). In total, we created 16 different textures, which comprised all combinations of the constrained properties, i.e. 4 bases (denoted *1.1, 1.2*, *2.1, 2.2*), 2 cross-frequency correlations (low/high CFC), and 2 realizations. The bases X.1 and X.2 differed only by random, amplitude-balanced changes in their subband means, which mimicked the differences we used in a previous study (*39*). This effectively introduced large (1.X vs. 2.X) and small (X.1 vs X.2) differences in the marginals. As in the study of (*18*), the CFCs are constrained as marginal (time-average) values between pairs of frequency channels. The average pairwise subband correlation in the low CFC case was 0.2, and in the high CFC case 0.8, inherited from the correlation structure in their respective source sounds.

We thus generated a well-controlled set of acoustic textures, whose statistical properties were constrained to be naturalistic based on the statistics of real sounds, however, for which all combinations existed. In particular, this allows us to consider the changes in cross-frequency correlations to be rather general, because it was applied over 4 different combinations of subband means and variances with 2 realizations for each. Importantly, the hearing range of C57Bl/6 mice at this age range is about 2-40 kHz (*108*, *109*), i.e. substantially higher than in humans. In order to avoid using all the stimulus energy in low frequencies that are inaudible to the mouse, we

applied the statistical properties along the frequency axis in the range 4-64kHz, using the same number of frequency channels (30) as in the statistical analysis of the original sound. The combination of the statistical properties from different sources, and the shift in frequency had the useful effect that the textures were only naturalistic, as opposed to naturally occurring and thus overall novel for the mice and required them to estimate their statistics in the experiment.

Texture synthesis from these statistics was carried out using the above-mentioned toolbox from (*18*). Briefly, this process entailed modifying a randomly generated Gaussian white noise sample to eventually match a set of prescribed statistics, using a gradient descent based line search. This process was iterative and was terminated when the resulting sound was sufficiently matched for the specified statistic. More details on how textures are synthesized from specified statistical values can be found in (*18*).

Mouse vocalizations were modeled as inverted 'u' shaped vocalizations, also referred to as chevron-type by some authors, which are one of the most common types of mouse vocalizations (*33*, *34*, *110*). These vocalizations started with an ascending frequency line, which reached a peak and then decayed more slowly and to a lower frequency than the initial frequency, as is typical for these vocalizations in mice (although a lot of variability exists). Vocalization base frequencies were varied pseudorandomly between 5 frequency locations ([2,4,8,16,32] kHz, see **Supplementary Figure 5C** for marginal spectra of the vocalizations) to ensure animals listened to the presented sounds broadly instead of focusing on a particular frequency range. The sound level of the vocalizations was adapted to a fixed dB difference relative to the subband marginal for each texture, to guarantee a constant degree of difficulty independent of the combination of texture and vocalization frequency. Each vocalization had a frequency range of 0.2 octaves. The parametric design of our background textures (see above) guaranteed that the subband marginal properties (mean and variance) were matched between low and high CFC sounds and therefore any changes in vocalization detection performance had to rely on differences in CFC. In particular, the matched levels between low and high CFC stimuli also did not provide an advantage in relation to the vocalizations, which would randomly be masked strongly or little in high CFC, while more consistently over time in the low CFC case, but to same degrees on average.

To understand the effect of exposure time on the detection of vocalizations, they were presented randomly at 9 linearly spaced onset times from 0 to 6 s. Vocalizations were delivered as a sequence of 10 repetitions at each trial, with an individual vocalization lasting 0.1 s with pauses between adjacent vocalizations lasting 0.1 s (in total, 1.9 s vocalization period). The texture sound continued after the last vocalization for another second. Recent research has

828 demonstrated that mice show a preference for such regularly spaced sequences, while caring
829 less about the precise shape of the vocalizations (*111*).
830 Synthesized sounds were converted to an analog output voltage (USB-6351, National
831 Instruments), amplified (A-9030, Onkyo) and delivered at a sampling rate of 250 kHz using a
832 high-fidelity speaker (T250D, Fostex) that had been calibrated before the experiment to output
833 the same sound intensity within 5 dB within the range of 2-100 kHz using an inverse impulse
834 response measured via an ultrasonic microphone (CM16/CMPA, Avisoft).

835 **Animal training**

836 Experiments were carried out on head-fixed animals inside a sound-proofed booth. The booth
837 was constructed from 3 cm composite wood and the inside walls, floor, and ceiling were covered
838 with acoustic foam (thickness: 5 cm, black surface Basotect Plan50, BASF). The acoustic foam
839 shields against external noises above ~1 kHz with a sound absorption coefficient >0.95 (defined
840 as the ratio between absorbed and incident sound intensity), which corresponds to >26 dB of
841 shielding in addition to the shielding provided by the booth itself. Sound delivery, data acquisition,
842 and control of the peripherals were performed by custom written software in MATLAB via a data
843 acquisition card (USB-6351, National Instruments).
844 The purpose of the detection task we used was to teach animals to respond to
845 vocalizations by licking the water source. Licks were automatically registered when the animal's
846 tongue broke an infrared beam connecting the infrared (IR) light source to the detector located
847 opposite it. The training procedure started with 1-3 days of accommodation to the setup. During
848 this period, animals got used to being head-fixed and rewarded from the water delivery system.
849 At the initial stage of the experimental training, sounds were generated using a set of parameters
850 that made the task easy for the animal. This primarily involved using shorter vocalization onset
851 times, longer vocalization epochs, and higher signal-to-noise ratio (SNR) in the generation of
852 sounds. The difficulty of the task was gradually increased to the level where the vocalizations
853 became non-trivial to detect and potential alternative strategies for carrying out the task did not
854 yield successful outcomes for the animal. To motivate animals to engage in the behavioral task,
855 the total water amount available to any given animal was limited to 1.5 mL/day by an automatic
856 water delivery system that tracked the amount of water delivered in the cage and during the
857 experiment.
858 Licking before vocalization onset was considered a *false alarm* (FA) and resulted in a
859 penalty waiting period (7 seconds). Licking during the vocalization period was considered a *hit*
860 and was rewarded with 5-7 μL water by opening a solenoid valve (#003-0120-900, Parker).

861 Licking after the last vocalization but before the end of the trial were considered a *miss* (see Fig. 3
862 for an analysis of these licks). The choice behavior of the animal was tracked over the whole
863 training period to verify that the animal was indeed responding to the vocalizations, rather than
864 using some other task-irrelevant cues. The animals were considered ready for electrode
865 implantation once they displayed a consistently high correlation between first lick time and
866 vocalization onset, showing that their responses were often time-locked to vocalization onset.
867 During the neural recording sessions, the active condition was always run before the passive
868 condition to avoid frustrating the thirsty animals and leading to behavioral extinction.

869 **Data analysis**

870 **Analysis of neural activity.** Neural recordings were bandpass-filtered (0.1-7000 Hz, Butterworth
871 filter, 4th order) and spike-sorted automatically under human control using a custom-written spike-
872 sorter (AutoSort) in Matlab. Sorting was performed per prong of the recording array to better
873 disambiguate neurons visible on multiple electrodes. AutoSort applies a standard sequence of
874 steps, i.e. detection of spikes based on a negative size threshold per recording channel, collection
875 of waveforms from all electrodes on a prong, dimensionality reduction (PCA) to 6 dimensions,
876 followed by clustering (hierarchical clustering with Ward distance) and semiautomatic fusion of
877 clusters based on intra- and between cluster distances. Subsequent analysis only worked on
878 spike timing.
879 Peristimulus-time histograms (PSTHs) were constructed with different temporal
880 alignments, specifically to texture sound onset, vocalization onset and lick times (see Fig. 3 and
881 5). Different PSTHs were also constructed for passive and active conditions, as well as different
882 trial outcomes. PSTHs represent the average, normalized firing rate across a set of neurons, i.e.
883 to account for different baseline firing rates, each neuron contributed to these PSTHs normalized
884 by its spontaneous rate, collected across all conditions. In particular this retains differences
885 between active and passive conditions in spontaneous and driven activity. Normalization per
886 alignment can also be informative to highlight relative changes, which we provide in
887 **Supplementary Figure 2**. In the analysis for Fig. 5C, we focussed on vocalization responsive
888 cells, which were identified via a Wilcoxon signed ranks test ($p<0.05$) between their onset and
889 sustained response.
890 **Unit classification.** To classify units we ran short optical stimulation-only recordings every day
891 prior to sessions involving sound delivery. We used light-evoked response within the 30 ms of
892 light onset and classified units as PV/SST interneurons when the light-evoked response of the
893 unit was significantly higher than the pre-light period. We used a relatively long time window

894 (30 ms) for classification to avoid misclassifying units due to light stimulation artifacts. However,
895 shorter classification windows produced comparable results (**Supplementary Figure 6**).
896 Qualitatively, our classification results were in agreement with previous studies that also
897 characterized spike parameters such as spike width and repolarization time (*112–114*). We
898 observed that while PV interneurons had features commonly described as fast or narrow spiking
899 (FS), not all cells with narrow spikes crossed the threshold mentioned. Therefore we additionally
900 trained a binary classification model to classify a cell either as FS or not, based on its spike
901 parameters, and analyzed the responses of this group of cells in parallel to optically identified
902 cells.
903 **Behavioral data analysis.** For time-sensitive d' analysis we combined response times and trial
904 outcomes from all trials for a given animal. Then with 0.75 s time steps, and 1.9 s window size we
905 estimated the d' for each time point using the approximation d'(t)=Z(HR(t)) - Z(FAR(t)), where Z(p)
906 is the inverse of the Gaussian cumulative distribution function (CDF), HR(t) corresponds to the
907 number of hit and correct rejection trials within the time window, whereas FAR(t) corresponds to
908 the number of false alarms. For all other behavioral analyses hit rate was estimated either at the
909 level of session or animal was used. The hit rate for a given condition was simply the ratio of hit
910 trials to total trials within a given condition. For tracking the training process, we shuffled the
911 reaction times for a given session and estimated the mean shuffled hit rate using 100
912 randomizations. The successful training session was defined as the one with a hit rate > $\mu_{shuffled}$ +
913 $2\sigma_{shuffled}$.
914 **Stimulus reconstruction.** To quantify the information in the neural population we performed
915 stimulus reconstruction from population response under passive recording conditions using the
916 "flat prior reconstruction" method described earlier (*46*). In contrast to the "optimal prior
917 reconstruction" method from the same study the flat prior method removes the correlations from
918 the input in the STRF estimation step and therefore does not have an inherent advantage for
919 sounds with higher cross-frequency correlations, which would have confounded out comparison
920 of reconstructions across the different levels of CFCs. Specifically, first, a subset of units (395 out
921 of 1353) was selected based on the signal-to-noise ratio (SNR) of their STRF. Linear
922 reconstruction kernels were computed using training data that encompassed the background only
923 section of the balanced set of stimuli. Kernels were estimated this way 20 times using different
924 sets of stimuli, randomly selected from the training data and the final reconstruction kernels used
925 in the analysis was the average across these randomizations. Reconstruction kernels were then
926 projected onto the neural responses from the test dataset in order to obtain reconstructed

927 spectrograms. The quality of reconstruction was quantified using point-wise correlation between
928 the original and reconstructed spectrograms.
929 **Vocalization decoding with support vector machine (SVM) analysis.** A binary SVM classifier
930 was used for classifying a given population vector as belonging to the vocalization (1) or
931 background epoch (0). The model was trained using a subset of data balanced for correlation,
932 light stimulation, and vocalization base frequencies. Spike counts for each time point and the unit,
933 SC(t,u), from the training dataset, were divided into SC(t_{VOC}, u) and SC(t_{BKG}, u) parts
934 corresponding to data coming from vocalization and background epochs. The trained model was
935 then 10-fold cross-validated. The model was used for labeling population vectors belonging to the
936 vocalization epoch of the test dataset. Since vocalizations were repeated multiple times during
937 the trial (See Acoustic stimulus section), the probability of being classified as vocalization was
938 averaged across different repetitions to yield a single mean time series of 200 ms length.
939 However, statistical tests (Kruskal-Wallis test) between different classification probabilities were
940 performed on unaveraged, raw probabilities outputted by the SVM classifier.
941 **Factorization of variance with nested ANOVA.** The variance of unit responses into texture
942 class and realization components was performed as described in detail earlier (*54*). For this
943 analysis, we averaged binned spike counts for each cell using moving to mean with window sizes
944 ranging from 30 ms to 1000 ms. The resulting neural response matrix for each cell had dimensions
945 of n(stimuli)×n(time bins) but the endpoints of the matrices were discarded, yielding shorter
946 matrices when the window size was larger. Spike counts were normalized to the baseline period
947 before the sound onset. To analyze the variance of the cell's response at a given time point,
948 nested ANOVA was performed. Realizations (2 per class) were excerpts taken from different parts
949 of the same clip. Each clip belonged to a certain texture class (8 in total) defined by its marginal
950 statistics and cross-frequence correlation. The realization variable was nested under the texture
951 class. Following (*54*) we used the variance explained (VE) by the ANOVA model to understand
952 the effect of window size, time course, and cell type on the analysis. Population VE by each
953 variable was obtained by averaging across cells. To summarize the relationship between the
954 window size and VE, we further averaged the population VE across time points.
955 **Convergence of Statistical Properties.** To check whether low and high CFC stimuli differed in
956 the variability of contained statistics in shorter duration samples, we ran excerpts with durations
957 ranging from 100 ms to 10 s through the estimation routines provided by (*18*), using the function
958 measure_texture_stats. We used 10 samples per duration and realization. The resulting statistics
959 were collected and next the standard deviation across the estimates was computed for the

envelope marginals, envelope variance, cross-frequency correlations and modulation power in each frequency band. The results of this analysis are shown in **Supplementary Figure 4**.

Figures and Tables

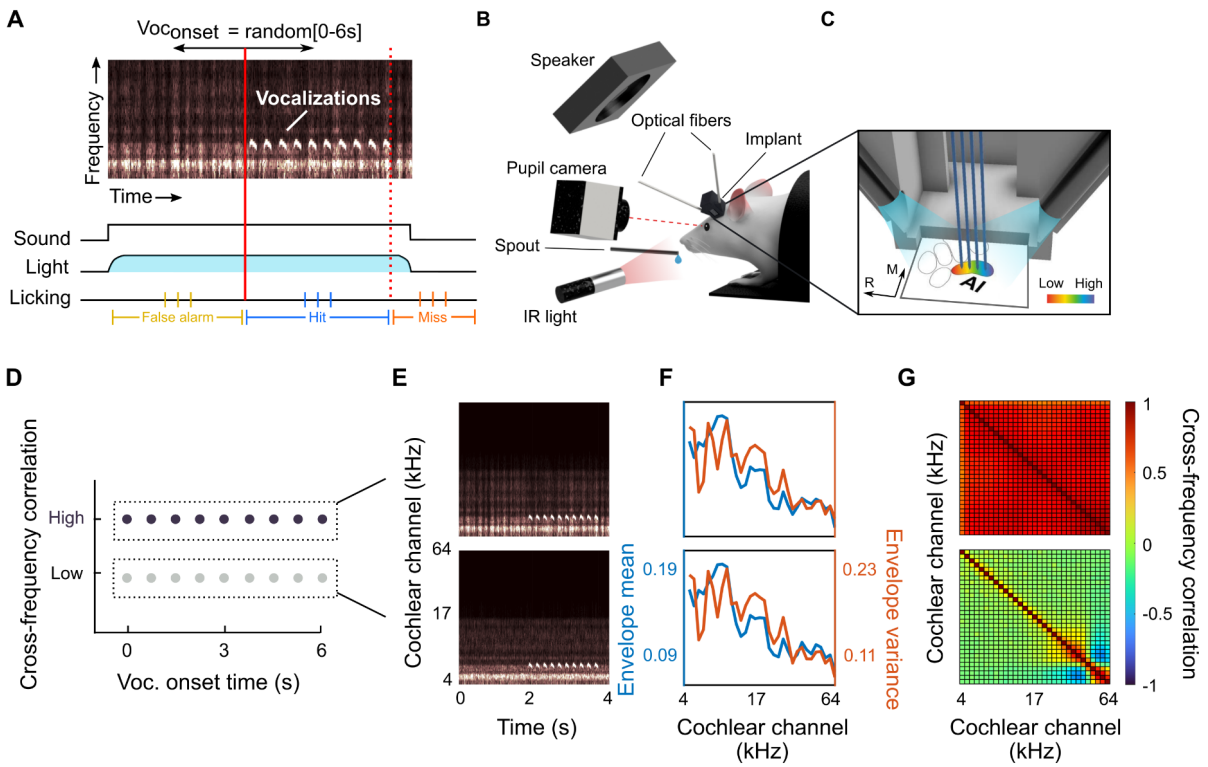


Figure 1. Neural recordings from the auditory cortex in mice actively detecting vocalizations masked by textural noise.

A Mice were head-fixed during the training and neural recording sessions. This allowed them to be in the same position w.r.t. the speaker, placed centrally in front of the animal. During all sessions, the left eye was recorded while illuminated with infrared light. During neural recording sessions, neural activity was recorded from the left auditory cortex using microelectrode arrays with 64 Ch. and the brain surface could be illuminated with light (470 nm) to activate PV interneurons.

B Electrodes were inserted at the primary auditory cortex such that they spanned along the tonotopic gradient in A1. The light was delivered bidirectionally through optical fibers that were inserted into the sides of the microdrive case.

C In each trial the mouse listened for vocalizations embedded in a sound texture. The frequency and onset timing (0-6 s) of vocalizations were randomly assigned to different trials. In a subset of trials (30%, randomly selected) PV interneurons were slightly activated (~5% change in FR) for the duration of the sound. The mice signaled that they detected a vocalization by licking, which led to a water reward if it occurred in the response window (0-1.9 s after vocalization onset), or no water if the lick occurred before (false alarm, FA), too late or not at all (miss). FAs were punished with a time-out of 7 s.

D We tested the influence of cross-frequency correlations (CFCs) by using textures with low and high CFCs for multiple vocalization onset times.

E Spectrograms of high and low cross-frequency correlation sounds. Sounds were matched for their marginal envelope mean and variance (**F**) but not for cross-frequency correlations (**G**). See Supplementary Figure 1 for all stimuli and Supplementary Figure 4C for all vocalizations.

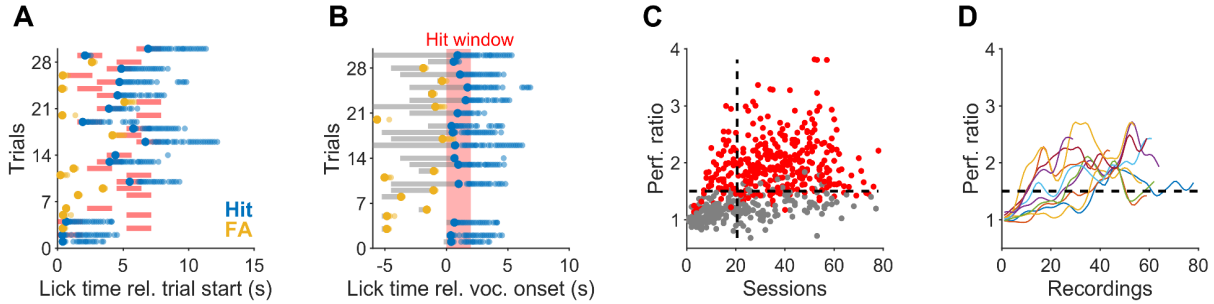


Figure 2. Mice learn to detect vocalizations embedded in textural noise.

A After training the mice learned to respond to the onset of the vocalization. The mice continued to lick (blue dots) during and beyond the response window (red line). False alarms (orange dots) tended to occur close to sound onset with typically only one or a few licks.

B Same data as in **A**, now aligned to vocalization onset (at 0, red line), showing that mice respond after the vocalization onset, indicating that they respond to the vocalization.

C Progress in training over time as a function of performance ratio (real / shuffled hit rate) across all animals and recordings. Red dots indicate recordings where the animal performed above the criterion which was set to be $2\times\sigma_{rnd}$ above the mean shuffled hit rate. Gray dots, on the other hand, correspond to recordings where the criterion was not met. Linear discriminant analysis revealed that animals tend to perform consistently above chance level after ~22 sessions and once the performance ratio exceeds 1.5.

D Same data as shown in (C), but session-binned for each animal individually. The black dashed line corresponds to the performance criterion derived in (C). Animals exhibit variability in terms of their learning rate but on average cross the performance criterion within ~22 recordings (SD=12.43), confirming the analysis in **C**.

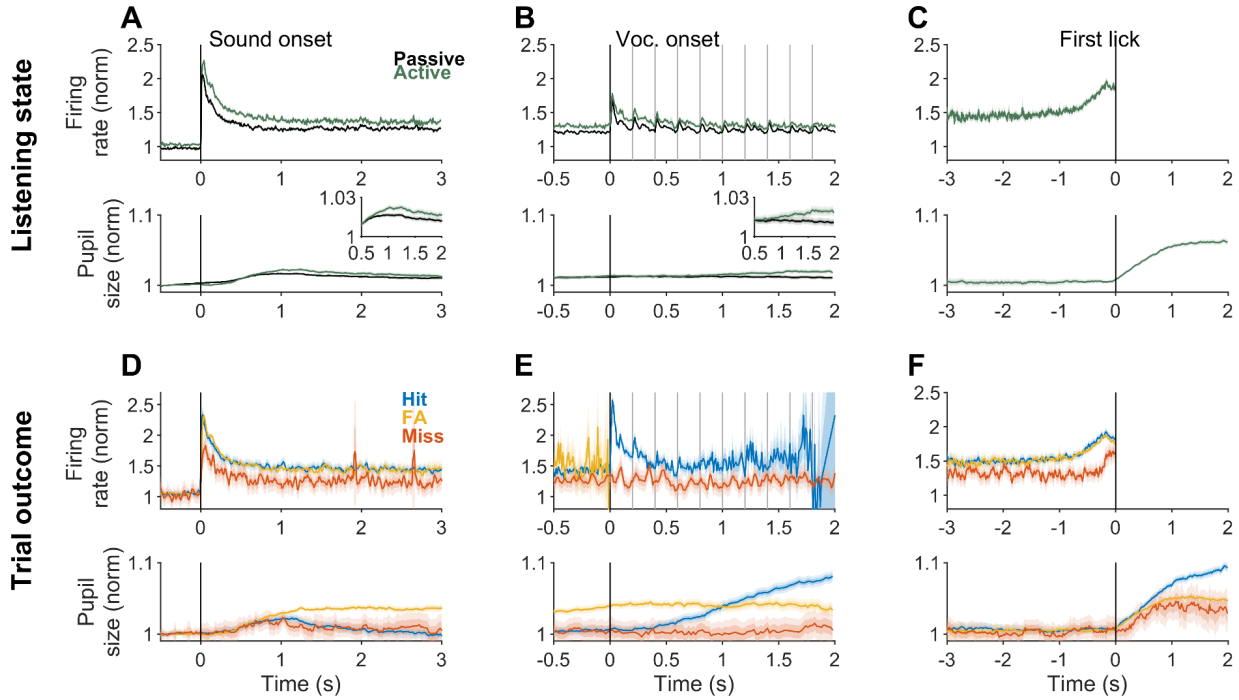


Figure 3. Neural population response and pupil diameter reflect behavioral variables reliably

Normalized firing rate (top panels) and pupil size (bottom panels) in active (n=12333 trials) and passive (n=19238 trials) listening conditions (A-C), and for different trial outcomes (D-F), aligned to sound onset (left column), vocalization onset (middle column), and first lick time (right column).

A Active engagement in the task is characterized by elevated average firing rate as well as larger pupil diameters both when the responses are aligned to sound onset from trials where no vocalizations occurred before 3 s. The inset shows a zoomed vertical axis.

B Same as (A), aligned to vocalization onset.

C Neural population response shows a ramping prior to the first lick, which may be indicative of cognitive processes leading up to the decision to lick.

D Sound-evoked population response was higher in hit and FA trials than in miss trials. Pupil diameter remained elevated in FA trials, likely reflecting the arousal after early licks in these trials.

E Similar to (D), vocalization-evoked neural responses were stronger in hit trials compared to miss trials. Pupil dynamics primarily reflect different time points when the animal starts licking the spout. In FA trials licking starts before vocalization onset, in hit trials ~0.5 s after, and in miss trials, the lick happens after vocalization offset.

F While neural responses show ramping up prior to the first lick (top), pupil diameter increases after the lick reflecting arousal, which is heightened by reward in hit trials (blue).

Neural and pupil data normalized to spontaneous averages per cell and animal, respectively (see Methods for Details and Supplementary Figure 2 for alignment-based normalization).

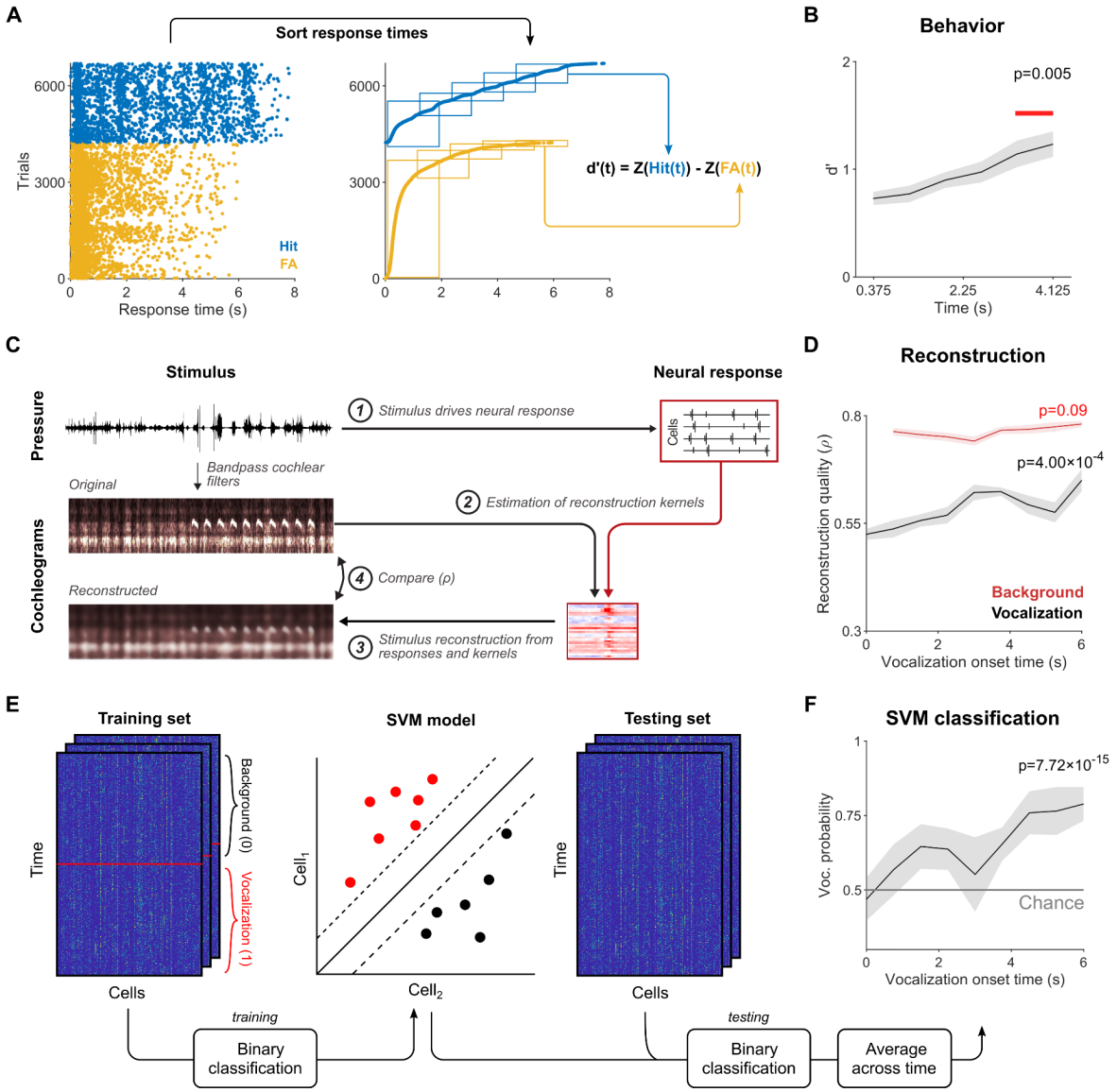


Figure 4. Detection performance and neural encoding of vocalizations improves with texture exposure.

A We quantified detection performance using a time-resolved d' measure, based on response times from all hit and FA trials (left) sorted in time (right). For each time bin, *t*, the proportion of hit and FA trials is computed. The difference between *Z(p)*, the inverse Gaussian cumulative distribution function for each type of trial, provides an estimate of d' for each time bin.

B The detection performance showed a significant increase over time into the trial (Kruskal-Wallis test, p=0.005, N=7 mice), i.e. with additional exposure to the texture. The red line indicates the time points after which d' becomes significantly different from the first time bin. Gray hull indicates 1 SEM.

C Reconstruction of sound stimuli was carried out by using a subset of stimulus-response pairs to estimate linear reconstruction filters. These filters were then used to reconstruct the original stimuli based on recorded neural responses. The quality of reconstructions was quantified by correlation between original and reconstructed spectrograms.

D The reconstruction quality of vocalization (black) significantly increased relative to the vocalization onset time in the trial, while the reconstruction quality of the background (red) epoch did not significantly change (Kruskal-Wallis test, p-values shown in figure, hulls indicate 1 SEM).

E A support vector machine (SVM) classifier was trained on a subset of data to perform the classification of the population activity vector as either vocalization (1) or background (0). The test set was then evaluated and binned across vocalization onset time bins as in **D**.

F The SVM applied to the population vectors from the vocalization period showed a significant increase with vocalization onset time, consistent with both reconstruction analysis (D) and behavioral results (B). Flat line represents chance-level performance.

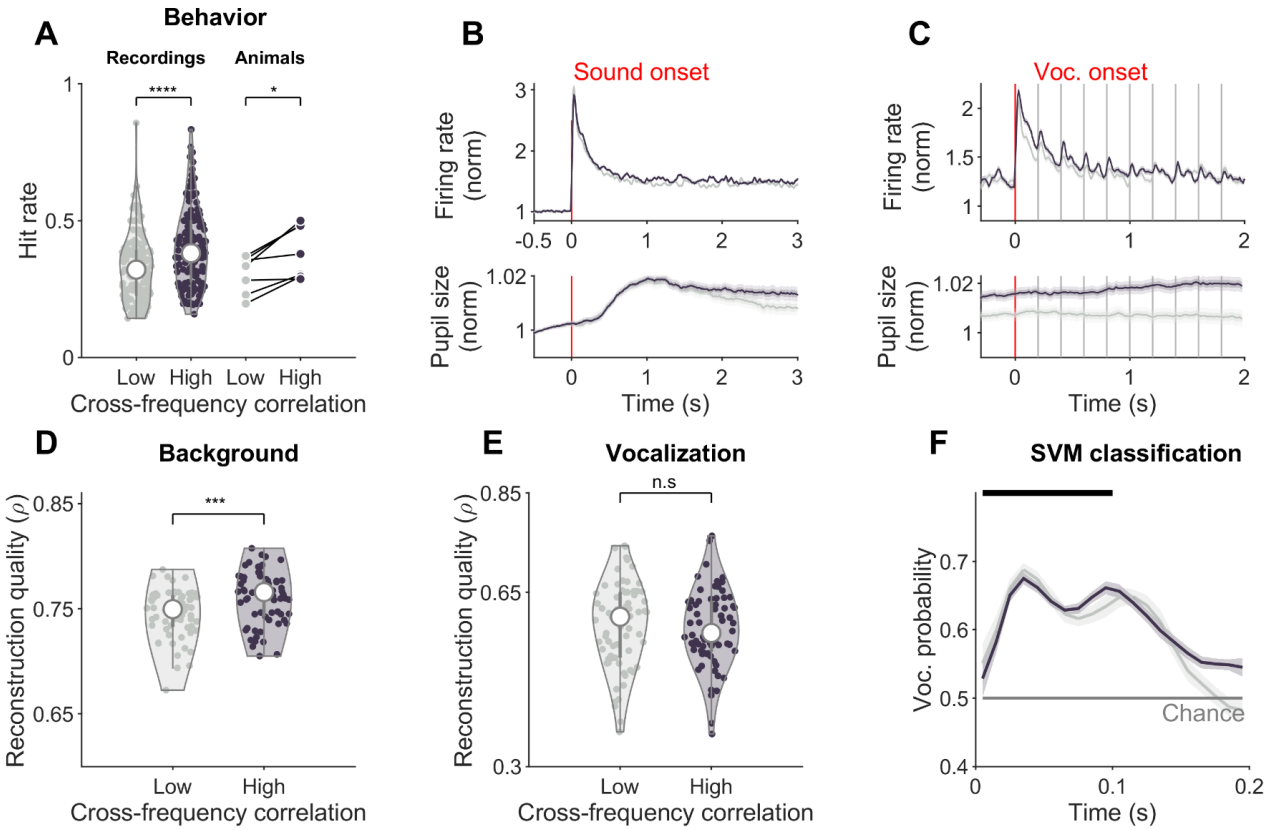


Figure 5. Detection of vocalizations is easier in the context of high cross-frequency correlations.

A Behavioral performance across cross-frequency correlation (CFC) conditions was quantified as hit rates and analyzed at the recording and animal levels. At both levels, animals were significantly better at detecting vocalizations embedded in high CFC backgrounds than low CFC backgrounds. White dots on the violin plots correspond to the median.

B The average PSTH (top) to the sound textures in active trials shows a small but significant difference, with high CFC adapting less than low CFC. The pupil response diverges after ~1 s, indicating heightened arousal under high CFC conditions. Curves include only trials where vocalization has not yet been presented and FAs are only included up to the lick.

C Average PSTH in response to vocalizations show a significantly greater response to high CFCs compared with low CFCs, consistent with comodulation masking studies. Again the pupil also dilates more in high CFC trials. Only responsive cells are included here.

D Quality of background reconstruction for high and low cross-frequency correlation textures shows that high CFC background-only epoch was better reconstructed, suggesting that the temporal details of texture were represented by the cells in A1.

E Vocalization epoch, on the other hand, did not significantly differ in terms of reconstruction quality.

F In the context of high CFC backgrounds, vocalizations were better recognizable, using an SVM based decoder, which was trained to label the population activity vector for a given time point as either vocalization or background. Black line represents the duration of a single vocalization presentation; solid gray line represents chance-level performance for the SVM classifier.

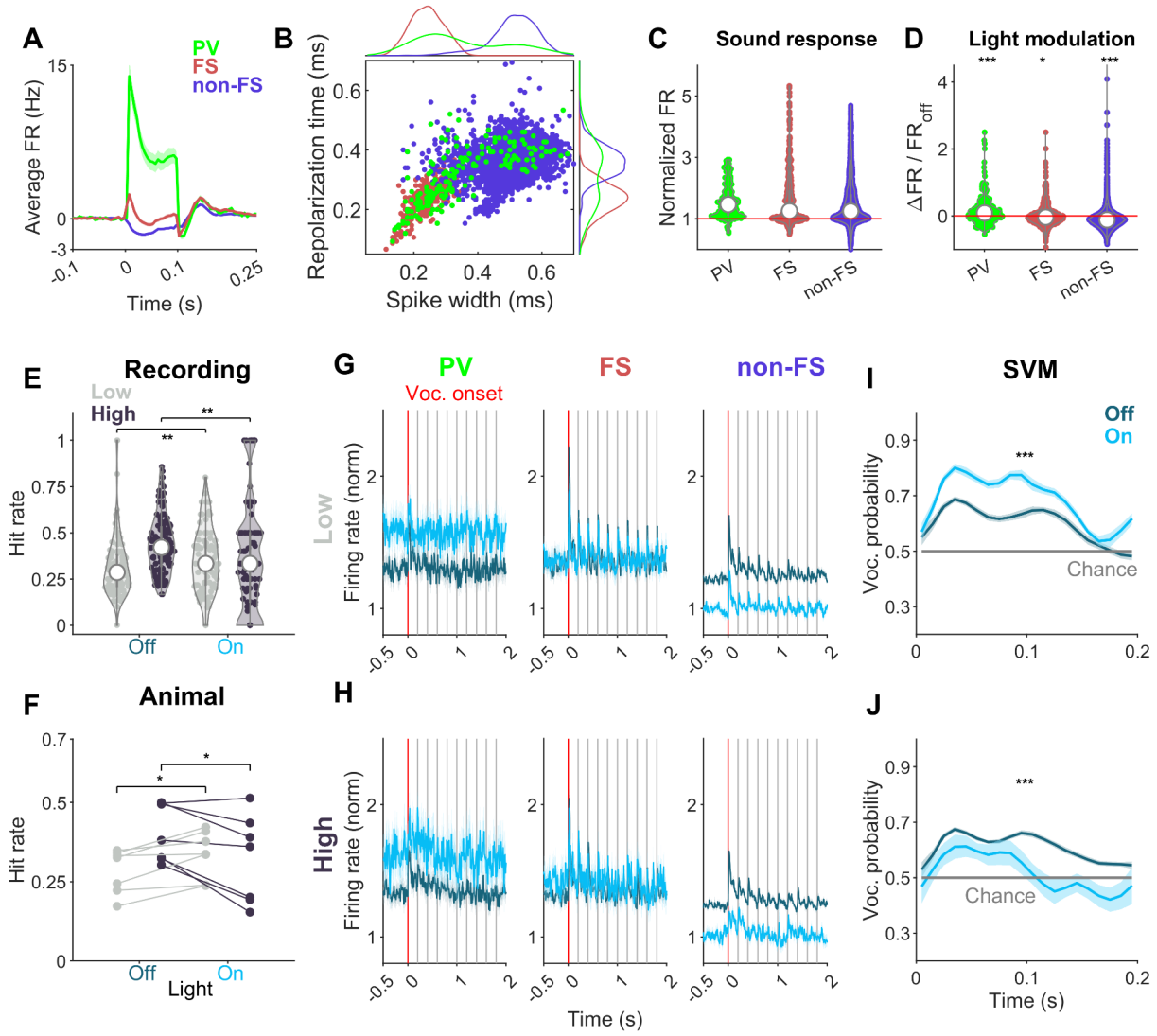


Figure 6. Optogenetic manipulation of PV cells has an asymmetric effect on behavior and neural encoding of vocalizations.

A Average firing rate change in parvalbumin interneurons (PV, green; n=182), fast-spiking, (FS, red, n=439), and non-FS cells (violet, n=3741) in opto-tagging sessions.

B Distribution of spike parameters shows PV cells tend to have narrower spike widths, though spike parameters do not uniquely identify PV cells.

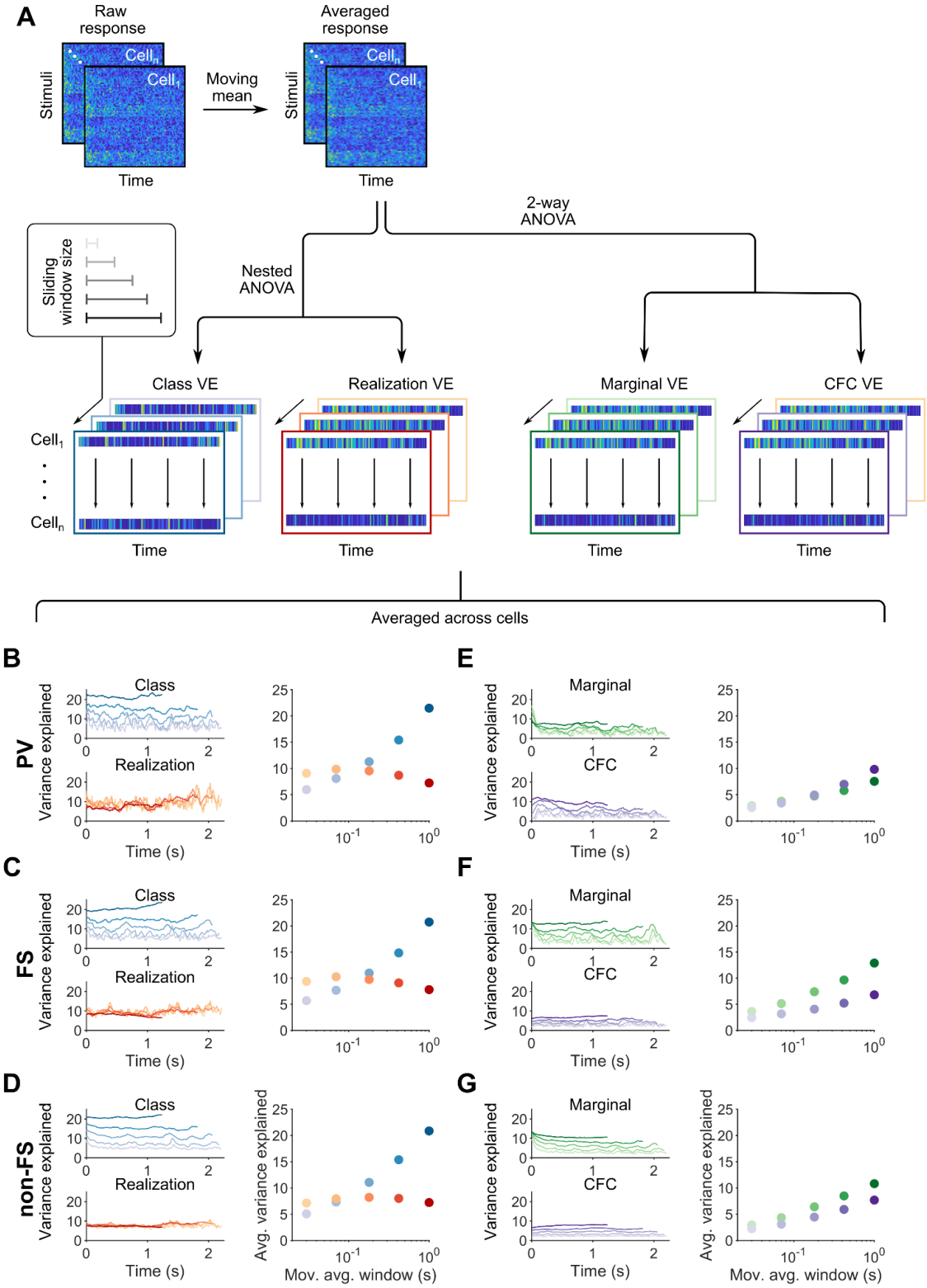
C All cell groups show significant sound-evoked responses relative to spontaneous firing rate (red), within 0.5s of texture onset.

D Optical stimulation led to a significant increase in average sound-evoked response in PV cells, while FS and non-FS cells were inhibited to different extents.

E/F Hit rates observed for low and high CFC under the light on and off conditions. Behavioral performance degrades under high CFC conditions but improves under low CFC conditions. Points correspond to different recordings (**E**, n=189) and animals (**F**, n=7).

G/H Vocalization-onset-aligned PSTH (normalized to pre-sound activity) grouped according to cell types and background texture conditions (top/bottom row: low vs. high CFC; Light on trials: light blue, Light off trials: dark blue). Vertical gray lines indicate successive repetitions of vocalizations. Coarse features of PSTHs such as steady-state response level reflect the light reaction of each cell type shown in **D**.

I/J We used the same SVM classifier as in Fig. 5E/F. The probability of being classified as a vocalization epoch was averaged over multiple presentations of vocalizations to yield a single time series for each condition. Despite having similar PSTH traces, the vocalization classification reflected the asymmetric effect of optical stimulation on behavioral performance (E/F). Gray line indicates change level performance. Error hulls represent 1 SEM.


A
Raw response
Cell$_n$
Cell$_1$
Stimuli
Time
Moving mean
Averaged response
Cell$_n$
Cell$_1$
Stimuli
Time
2-way ANOVA
Sliding window size
Nested ANOVA
Class VE
Realization VE
Marginal VE
CFC VE
Cell$_1$
Cell$_n$
Time
Time
Time
Time
Averaged across cells
B
C
D
E
F
G
PV
FS
non-FS
Class
Realization
Marginal
CFC
Variance explained
Time (s)
Avg. variance explained
Mov. avg. window (s)

1091 **Figure 7. Sensitivity of different cell types to sound texture features**

1092 **A** The total variance of a cell's response to sound textures was partitioned into contributions from different
1093 variables defining background sounds. We first averaged the cell's response over time (first 2.25 s of texture
1094 only) using five logarithmically spaced, sliding window lengths. The averaged response of the cell was
1095 analyzed using either nested ANOVA (B-D) or two-way ANOVA at each time point. Variance explained (VE)
1096 was averaged across cells to evaluate change in VE over time as the animal listened to the sound, and to
1097 understand the effect of moving window size.
1098 **B-D** For shorter window lengths, realization VE exceeded class VE, whereas for longer windows, the
1099 opposite trend was observed. Nested ANOVA was conducted to partition the variability attributed to texture
1100 class and realization. The analysis was performed separately for PV (B; n=49), FS (C; n=192), and non-FS
1101 (D; n=1081) cells. All cell groups exhibited an increase in explained variance (VE) concerning texture class
1102 as window lengths increased, while their time courses remained relatively constant over the 2.25 s of
1103 response analyzed (left). To assess the influence of window length on variance component magnitudes,
1104 we calculated the average for each component over time (right). Notably, realization VE (yellow-red)
1105 displayed significant differences across cell types, whereas texture class VE (blues) did not (see text for
1106 statistics). Additionally, window length significantly affected both components.
1107 **E-G** Impact of marginal and CFC statistics on responses from various cell types. Consistent with B-D, we
1108 observed a progressive increase in VE by all cell types as the window size increased (2-way ANOVA). Both
1109 window size and cell type significantly affected the VE of both types of texture statistics, highlighting
1110 differences among the features that different cell types are sensitive to.

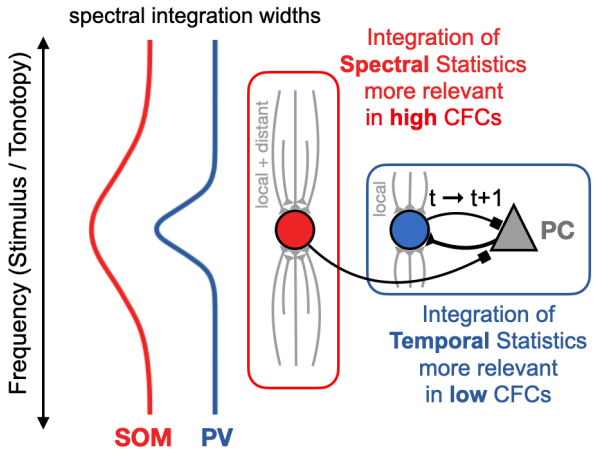


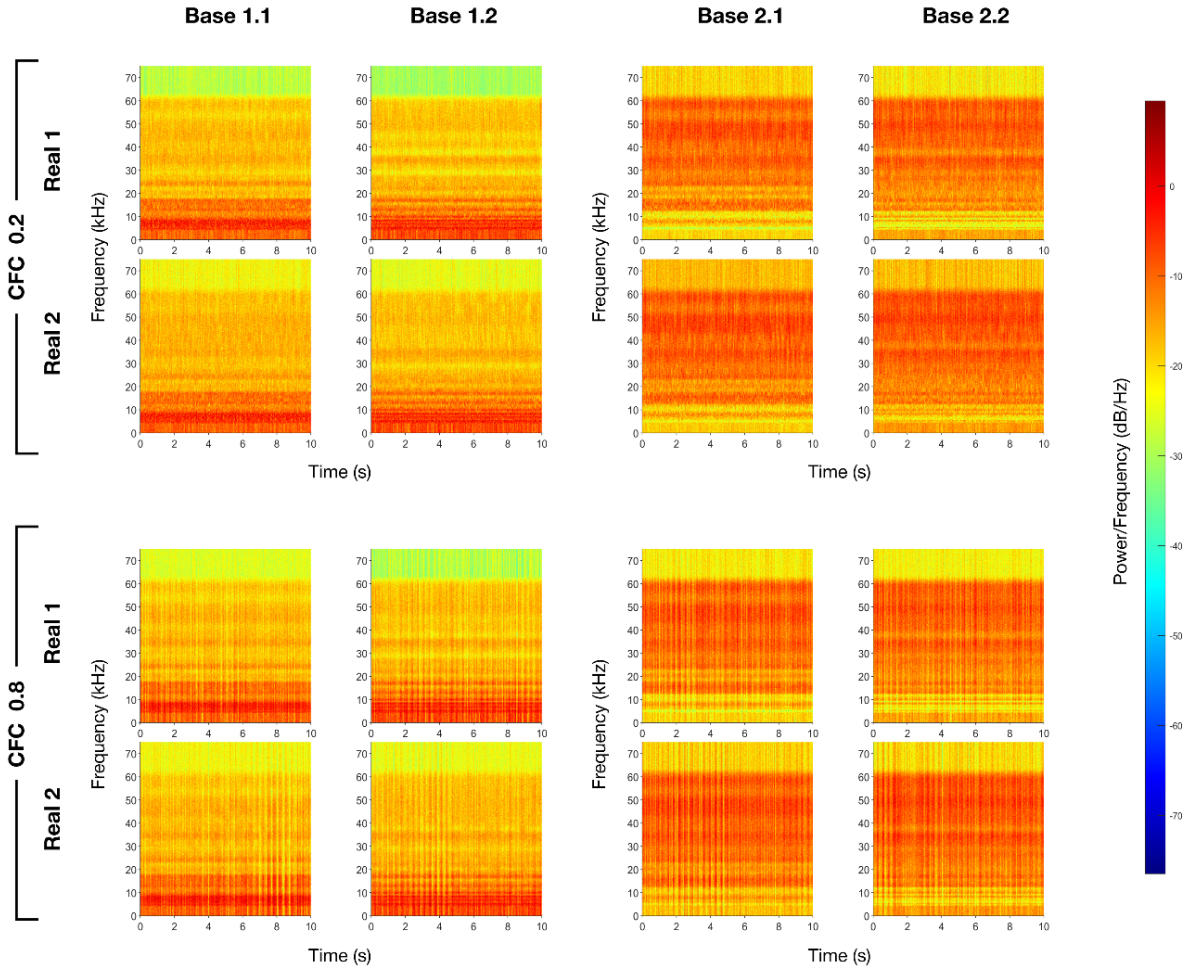
Figure 8: Putative roles of PV and SOM cells in the integration of spectral and temporal statistics

The activity of PV and SOM interneurons both modulate the activity of pyramidal cells in the cortex. However, PV and SOM interneurons differ in their spectral integration width, with SOM cells integrating more widely than PV cells. This would allow SOM cells to contribute to the integration of spectral statistics, e.g. in the case of high CFCs (Lakunina et al. 2020). As presently shown, the activation of PV cells leads to improved encoding of vocalization in low CFCs (and vice versa). Together with the known reciprocal connections between PV and PC, this suggests that they contribute more to narrowband, temporal integration of statistics. Gray connections indicate a mixture of intracortical and thalamocortical inputs.

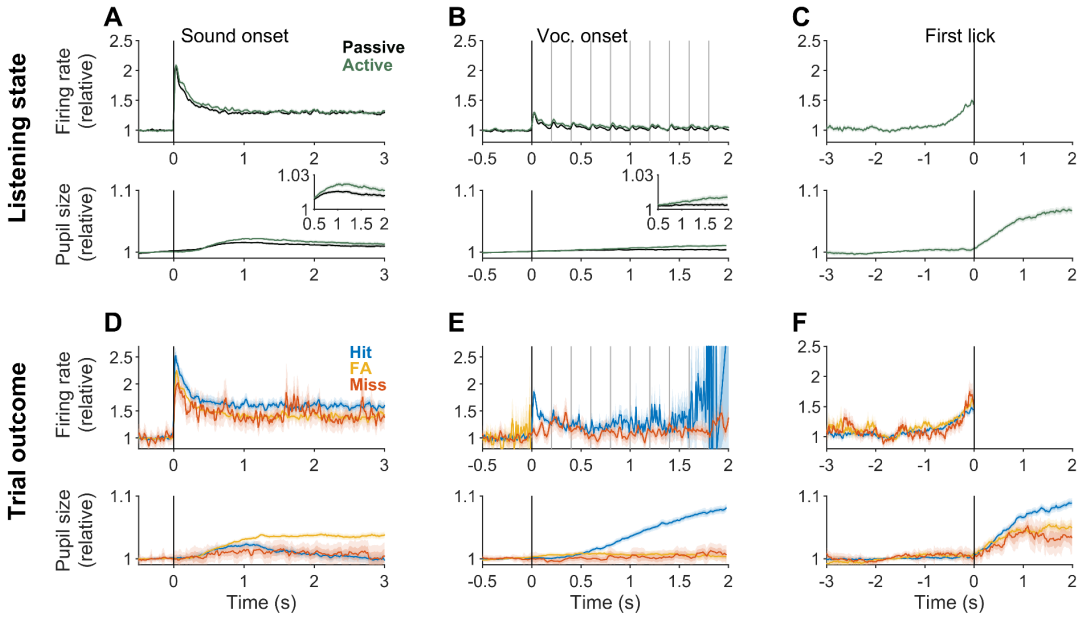
Texture Label	Source sound
Base 1.1/1.2	Subband means: Bubbling_water.wav Subband variances: Drumroll.wav
Base 2.1/2.2	Subband means: Shaking_coins.wav Subband variances: Crowd_noise.wav
Low CFC	Cross-frequency correlations: Fast_running_river.wav
High CFC	Cross-frequency correlations: Jogging_on_gravel.wav

Table 1. List of source sounds for statistical values imposed on white noise.

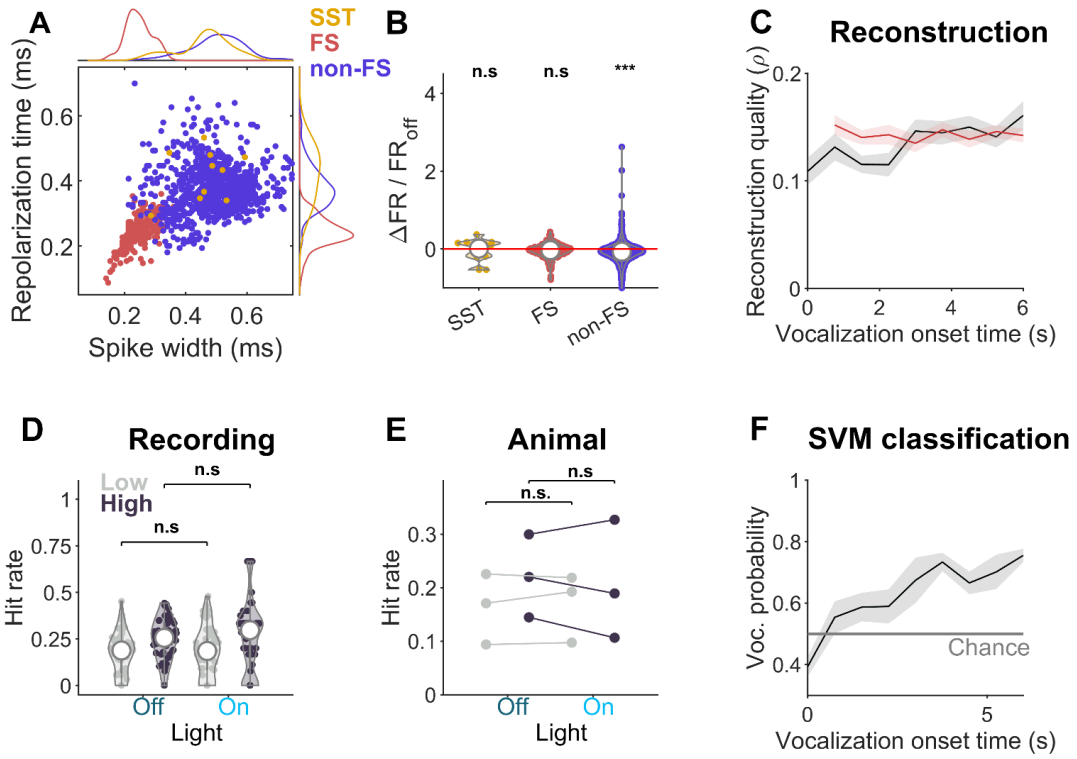
Supplementary Information



Supplementary Figure 1. Spectrograms of background textures used in the experiment. In total, 16 distinct background textures were generated and used in the experiment. The textures were created using two different base frequency configurations (Base 1 and Base 2, each with two variants: 1.1, 1.2, 2.1, and 2.2, columns), defined by their envelope means and variances. Bases 1 and 2 were chosen to cover low and high frequencies, in order to activate a large set of neurons. The textures were further differentiated by their cross-frequency correlation (CFC, top/bottom rows), with either high (0.8) or low (0.2) average correlation across frequency bands. For each combination of CFC and base configuration, two independent realizations (Real 1 and Real 2, rows within CFCs) were generated.



Supplementary Figure 2. Unit responses normalized by experimental condition. Similar to Figure 3 in the main text, but here the response was normalized separately for each condition. For sound and vocalization onset-aligned responses, the 0.5 s time window prior to the respective event was used to normalize the data. For first-lick aligned responses, the time window from -2 to -1.5 s prior to the lick was used to normalize the data.



Supplementary Figure 3. Manipulation of SST cells does not affect behavioral performance.

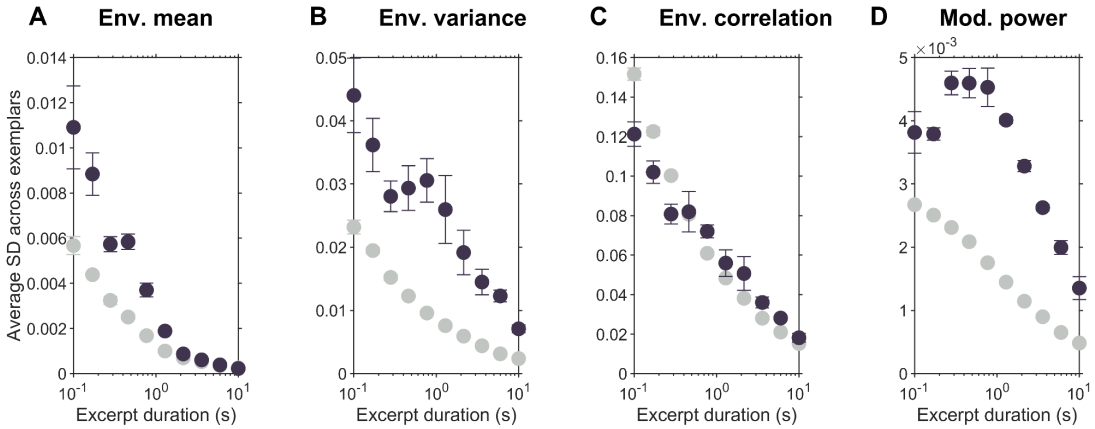
A Unlike PV cells shown in Figure 6, SST cells (n = 12) were found to overlap primarily with non-FS cells.

B Activation of SST cells does not affect the sound-evoked activity of SST (median = 1.67%, Wilcoxon signed rank test, p = 0.791) and FS cells (median = 1.35%, Wilcoxon signed rank test, p = 0.752), but significantly reduces that of non-FS cells (median = -3.82%, Wilcoxon signed rank test, p < 0.001).

C Reconstruction quality of vocalization (light gray) and background (red) epoch is quantified for stimuli with different vocalization onset times. The thick black line represents average reconstruction quality across sounds with identical vocalization onset times.

D-E Activation of SST cells does not affect the behavioral performance for low or high CFC backgrounds, neither at recording (n = 60 recordings; Wilcoxon signed rank test, p = 0.697 for low CFC; p = 0.416 for high CFC sounds) nor at animal levels (n = 3; Wilcoxon signed rank test, p = 0.750 for low CFC; p = 0.500 for high CFC).

F The probability of the population vector being classified as coming from the vocalization epoch gradually increased in SST dataset as well. Solid gray line represents chance-level performance for the SVM classifier.



Supplementary Figure 4. Statistics time-courses for low and high CFC textures.

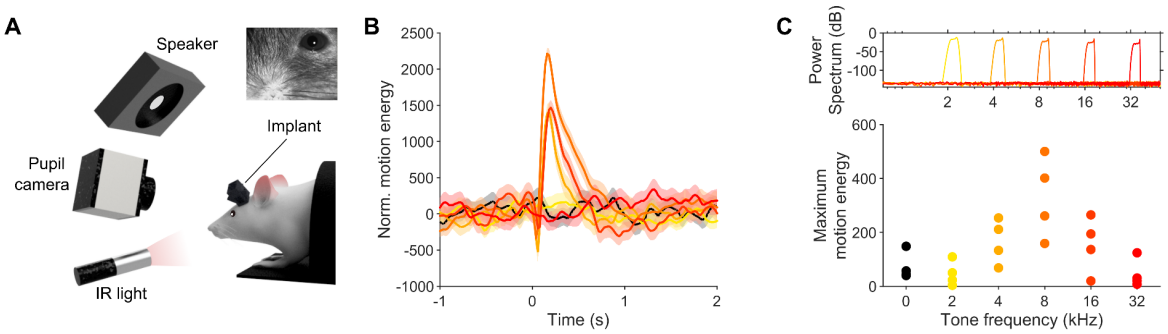
We used the toolset of (*18*) to analyze the variability in statistical properties estimated from different lengths of data, ranging from 100 ms to 10 s, i.e. covering the range of durations used in our experiment. Importantly, the values below are the S.D. across 10 samples, i.e. are not the value for the statistic itself, mirroring the analysis in Fig. 1C, (*25*) Error bars show SEM across textures.

A The marginals of the envelopes per frequency channel converged after ~1 s, with low CFC textures exhibiting a faster decay than high CFC textures, which is consistent with the greater variability across time in the latter.

B The variance of the envelope per frequency channel converged more slowly than the marginals, with CFC textures converging much more slowly and exhibiting more residual variability.

C The correlation across frequency channels exhibited also slower convergence of estimation accuracy with duration, which however, did not markedly differ between low and high CFCs.

D The modulation power in the envelopes surprisingly exhibited a non-monotonic decay, for high CFC stimuli only, suggesting that new time-scales of modulation came 'into sight', when the duration increased to around 500 ms, before then converging. Low CFC textures instead showed a slow decay that was comparable in speed to envelope variance and correlation.

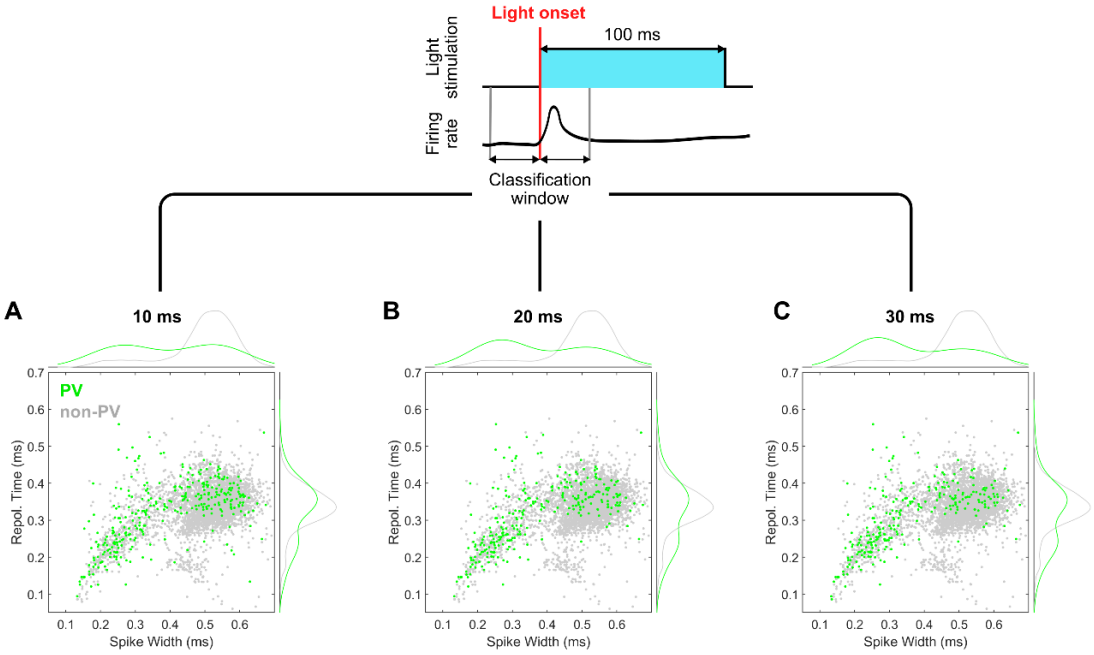


Supplementary Figure 5. Quantifying hearing in untrained animals through facial motion analysis.

A Untrained animals (n = 4) were exposed to a randomized sequence of pure tones (t = 2 s), while the animal's face was recorded (inset).

B Recorded video data was subjected to dimensionality reduction via singular value decomposition (SVD). The motion energy of the first component, normalized to different tone frequencies (from yellow to red), is shown here, aligned to the tone onset for an example session.

C Top: Envelope mean of vocalization sounds showing power across different frequencies (2-32 kHz) for vocalizations used in the task. Bottom: The maximum response of the first component to different frequencies demonstrates that the animal's response to certain frequencies is reduced to the level of the control condition of silence (black dots), indicating that frequencies on the edges of hearing ability (2 and 32 kHz) are no longer heard as well as frequencies between 4-16 kHz.



Supplementary Figure 6. Classification of light-sensitive neurons using different temporal windows.

Individual neurons were classified as PV (green) or non-PV (gray) based on their responses to light stimulation. Classification was performed by statistically comparing firing rates before and after light onset during repeated stimulation trials. Panels A, B, and C show the distribution of classified neurons when using classification windows of 10 ms, 20 ms, and 30 ms respectively around the light onset. While the number of identified PV neurons decreases with longer classification windows (A: n=336; B: n=293; C: n=264), the distribution of spike parameters (repolarization time vs. spike width) remains quite similar across all temporal windows. 78% of cells classified as PV cells with a 30 ms classification window are also classified as such using shorter window lengths (in contrast to 61% and 71% classified using 10 and 20 ms, respectively). Density plots along the axes show the distribution of each parameter for both cell types.

References

1. H. B. Barlow, “Sensory Communication (ed. Rosenblith, W. A.)” ((MIT Press, Cambridge, MA)., 1961), pp. 217–234.

2. A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* **36**, 181–204 (2013).

3. R. P. N. Rao, D. H. Ballard, Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* **2**, 79–87 (1999).

4. I. Winkler, S. L. Denham, I. Nelken, Modeling the auditory scene: predictive regularity representations and perceptual objects. *Trends Cogn Sci* **13**, 532–540 (2009).

5. B. D. B. Willmore, A. J. King, Adaptation in auditory processing. *Physiological Reviews*, doi: 10.1152/physrev.00011.2022 (2022).

6. E. C. Cherry, Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America* **25**, 975–979 (1953).

7. J. C. Middlebrooks, J. Z. Simon, A. N. Popper, R. R. Fay, Eds., *The Auditory System at the Cocktail Party* (Springer International Publishing, Cham, 2017; http://link.springer.com/10.1007/978-3-319-51662-2)vol. 60 of *Springer Handbook of Auditory Research*.

8. H. J. Kennedy, M. G. Evans, A. C. Crawford, R. Fettiplace, Fast adaptation of mechanoelectrical transducer channels in mammalian cochlear hair cells. *Nat Neurosci* **6**, 832–836 (2003).

9. R. G. Natan, W. Rao, M. N. Geffen, Cortical Interneurons Differentially Shape Frequency Tuning following Adaptation. *Cell Rep* **21**, 878–890 (2017).

10. D. Pérez-González, M. Malmierca, Adaptation in the auditory system: an overview. *Frontiers in Integrative Neuroscience* **8** (2014).

11. A. J. E. Kell, J. H. McDermott, Invariance to background noise as a signature of non-primary auditory cortex. *Nat Commun* **10**, 3958 (2019).

12. D. M. Schneider, S. M. N. Woolley, Sparse and Background-Invariant Coding of Vocalizations in Auditory Scenes. *Neuron* **79**, 141–152 (2013).

13. H. K. Kato, S. K. Asinof, J. S. Isaacson, Network-Level Control of Frequency Tuning in Auditory Cortex. *Neuron* **95**, 412-423.e4 (2017).

14. A. A. Lakunina, M. B. Nardoci, Y. Ahmadian, S. Jaramillo, Somatostatin-Expressing Interneurons in the Auditory Cortex Mediate Sustained Suppression by Spectral Surround. *J. Neurosci.* **40**, 3564–3575 (2020).

15. A. A. Lakunina, N. Menashe, S. Jaramillo, Contributions of Distinct Auditory Cortical Inhibitory Neuron Types to the Detection of Sounds in Background Noise. *eNeuro* **9**,

ENEURO.0264-21.2021 (2022).

16. A. M. Lesicko, C. F. Angeloni, J. M. Blackwell, M. De Biasi, M. N. Geffen, Corticofugal regulation of predictive coding. *eLife* **11**, e73289 (2022).

17. E. A. K. Phillips, C. E. Schreiner, A. R. Hasenstaub, Cortical Interneurons Differentially Regulate the Effects of Acoustic Context. *Cell Reports* **20**, 771–778 (2017).

18. J. H. McDermott, E. P. Simoncelli, Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron* **71**, 926–940 (2011).

19. O. Scheuregger, J. Hjortkjær, T. Dau, Identification and Discrimination of Sound Textures in Hearing-Impaired and Older Listeners. *Trends in Hearing* **25**, 23312165211065608 (2021).

20. N. C. Rabinowitz, B. D. B. Willmore, A. J. King, J. W. H. Schnupp, Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. *PLOS Biology* **11**, e1001710 (2013).

21. N. C. Rabinowitz, A. J. King, Auditory Perception: Hearing the Texture of Sounds. *Current Biology* **21**, R967–R968 (2011).

22. J. M. Hicks, J. H. McDermott, Noise schemas aid hearing in noise. *Proc Natl Acad Sci U S A* **121**, e2408995121 (2024).

23. N. Mesgarani, S. V. David, J. B. Fritz, S. A. Shamma, Mechanisms of noise robust representation of speech in primary auditory cortex. *Proceedings of the National Academy of Sciences* **111**, 6792–6797 (2014).

24. A. P. Mishra, F. Peng, K. Li, N. S. Harper, J. W. H. Schnupp, Sensitivity of neural responses in the inferior colliculus to statistical features of sound textures. *Hearing Research* **412**, 108357 (2021).

25. R. Narayan, V. Best, E. Ozmeral, E. McClaine, M. Dent, B. Shinn-Cunningham, K. Sen, Cortical interference effects in the cocktail party problem. *Nat Neurosci* **10**, 1601–1607 (2007).

26. F. Peng, N. S. Harper, A. P. Mishra, R. Auksztulewicz, J. W. H. Schnupp, Dissociable roles of the auditory midbrain and cortex in processing the statistical features of natural sound textures. bioRxiv [Preprint] (2023). https://doi.org/10.1101/2023.04.12.536666.

27. S. Souffi, C. Lorenzi, C. Huetz, J.-M. Edeline, Robustness to Noise in the Auditory System: A Distributed and Predictable Property. *eNeuro* **8**, ENEURO.0043-21.2021 (2021).

28. X. Zhai, F. Khatami, M. Sadeghi, F. He, H. L. Read, I. H. Stevenson, M. A. Escabí, Distinct neural ensemble response statistics are associated with recognition and discrimination of natural sound textures. *Proc Natl Acad Sci U S A* **117**, 31482–31493 (2020).

29. C. F. Angeloni, W. Młynarski, E. Piasini, A. M. Williams, K. C. Wood, L. Garami, A. M.

Hermundstad, M. N. Geffen, Dynamics of cortical contrast adaptation predict perception of signals in noise. *Nat Commun* **14**, 4817 (2023).

30. J. H. McDermott, M. Schemitsch, E. P. Simoncelli, Summary statistics in auditory perception. *Nat Neurosci* **16**, 493–498 (2013).

31. J. M. Festen, R. Plomp, Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *J Acoust Soc Am* **88**, 1725–1736 (1990).

32. J. W. Hall, M. P. Haggard, M. A. Fernandes, Detection in noise by spectro-temporal pattern analysis. *The Journal of the Acoustical Society of America* **76**, 50–56 (1984).

33. J. J. Heckman, R. Proville, G. J. Heckman, A. Azarfar, T. Celikel, B. Englitz, High-precision spatial localization of mouse vocalizations during social interaction. *Sci Rep* **7**, 3017 (2017).

34. M. V. Segbroeck, A. T. Knoll, P. Levitt, S. Narayanan, MUPET—Mouse Ultrasonic Profile ExTraction: A Signal Processing Tool for Rapid and Unsupervised Analysis of Ultrasonic Vocalizations. *Neuron* **94**, 465-485.e5 (2017).

35. R. C. Wilson, A. Shenhav, M. Straccia, J. D. Cohen, The Eighty Five Percent Rule for optimal learning. *Nat Commun* **10**, 4646 (2019).

36. M. J. McGinley, S. V. David, D. A. McCormick, Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron* **87**, 179–192 (2015).

37. V. Breton-Provencher, M. Sur, Active control of arousal by a locus coeruleus GABAergic circuit. *Nat Neurosci* **22**, 218–228 (2019).

38. C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, K. D. Harris, Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, eaav7893 (2019).

39. Y. Boubenec, J. Lawlor, U. Górska, S. Shamma, B. Englitz, Detecting changes in dynamic and complex acoustic environments. *eLife* **6**, e24910 (2017).

40. R. G. O'Connell, P. M. Dockree, S. P. Kelly, A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nat Neurosci* **15**, 1729–1735 (2012).

41. S. Jaramillo, A. M. Zador, The auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nat Neurosci* **14**, 246–251 (2011).

42. E. Zagha, J. C. Erlich, S. Lee, G. Lur, D. H. O'Connor, N. A. Steinmetz, C. Stringer, H. Yang, The Importance of Accounting for Movement When Relating Neuronal Activity to Sensory and Cognitive Processes. *J. Neurosci.* **42**, 1375–1382 (2022).

43. A. Verdier, N. Dominique, D. Groussard, A. Aldanondo, B. Bathellier, S. Bagur, Enhanced perceptual task performance without deprivation in mice using medial forebrain bundle stimulation. *Cell Reports Methods* **2** (2022).

44. B. Khalighinejad, J. L. Herrero, A. D. Mehta, N. Mesgarani, Adaptation of the human auditory cortex to changing background noise. *Nat Commun* **10**, 2509 (2019).

45. C. Keine, R. Rübsamen, B. Englitz, Signal integration at spherical bushy cells enhances representation of temporal structure but limits its range. *eLife* **6**, e29639 (2017).

46. N. Mesgarani, S. V. David, J. B. Fritz, S. A. Shamma, Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex. *Journal of Neurophysiology* **102**, 3329–3339 (2009).

47. I. Nelken, Y. Rotman, O. Bar Yosef, Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* **397**, 154–157 (1999).

48. J. Sollini, P. Chadderton, Comodulation Enhances Signal Detection via Priming of Auditory Cortical Circuits. *J Neurosci* **36**, 12299–12311 (2016).

49. J. L. Verhey, D. Pressnitzer, I. M. Winter, The psychophysics and physiology of comodulation masking release. *Exp Brain Res* **153**, 405–417 (2003).

50. P. R. Murphy, J. Vandekerckhove, S. Nieuwenhuis, Pupil-Linked Arousal Determines Variability in Perceptual Decision Making. *PLOS Computational Biology* **10**, e1003854 (2014).

51. M. Aizenberg, L. Mwilambwe-Tshilobo, J. J. Briguglio, R. G. Natan, M. N. Geffen, Bidirectional Regulation of Innate and Learned Behaviors That Rely on Frequency Discrimination by Cortical Inhibitory Neurons. *PLOS Biology* **13**, e1002308 (2015).

52. F. Studer, T. R. Barkat, Inhibition in the auditory cortex. *Neuroscience & Biobehavioral Reviews* **132**, 61–75 (2022).

53. S. Q. Lima, T. Hromádka, P. Znamenskiy, A. M. Zador, PINP: A New Method of Tagging Neuronal Populations for Identification during In Vivo Electrophysiological Recording. *PLOS ONE* **4**, e6099 (2009).

54. C. M. Ziemba, J. Freeman, J. A. Movshon, E. P. Simoncelli, Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences* **113**, E3140–E3149 (2016).

55. S. Panzeri, N. Brunel, N. K. Logothetis, C. Kayser, Sensory neural codes using multiplexed temporal scales. *Trends in Neurosciences* **33**, 111–120 (2010).

56. U. Górska, A. Rupp, Y. Boubenec, T. Celikel, B. Englitz, Evidence Integration in Natural Acoustic Textures during Active and Passive Listening. *eNeuro* **5**, ENEURO.0090-18.2018 (2018).

57. S. M. Town, K. C. Wood, J. K. Bizley, Signal processing in auditory cortex underlies degraded speech sound discrimination in noise. *bioRxiv*, 833558 (2019).

58. R. McWalter, J. H. McDermott, Adaptive and Selective Time Averaging of Auditory Scenes. *Current Biology* **28**, 1405-1418.e10 (2018).

59. M. López Espejo, S. V. David, A sparse code for natural sound context in auditory cortex. *Current Research in Neurobiology* **6**, 100118 (2024).

60. M. S. Malmierca, S. Cristaudo, D. Pérez-González, E. Covey, Stimulus-specific adaptation in the inferior colliculus of the anesthetized rat. *The Journal of Neuroscience* **29**, 5483–5493 (2009).

61. N. Ulanovsky, L. Las, I. Nelken, Processing of low-probability sounds by cortical neurons. *Nat Neurosci* **6**, 391–398 (2003).

62. N. Ulanovsky, L. Las, D. Farkas, I. Nelken, Multiple Time Scales of Adaptation in Auditory Cortex Neurons. *J Neurosci* **24**, 10440–10453 (2004).

63. A. M. Taberner, M. C. Liberman, Response Properties of Single Auditory Nerve Fibers in the Mouse. *Journal of Neurophysiology* **93**, 557–569 (2005).

64. S. Tolnai, O. Hernandez, B. Englitz, R. Rübsamen, M. S. Malmierca, The medial nucleus of the trapezoid body in rat: spectral and temporal properties vary with anatomical location of the units. *European Journal of Neuroscience* **27**, 2587–2598 (2008).

65. H. K. Kato, S. N. Gillet, J. S. Isaacson, Flexible Sensory Representations in Auditory Cortex Driven by Behavioral Relevance. *Neuron* **88**, 1027–1039 (2015).

66. D. P. Knudsen, T. Q. Gentner, Active recognition enhances the representation of behaviorally relevant information in single auditory forebrain neurons. *Journal of Neurophysiology* **109**, 1690–1703 (2013).

67. G. H. Otazu, L.-H. Tai, Y. Yang, A. M. Zador, Engaging in an auditory task suppresses responses in auditory cortex. *Nat Neurosci* **12**, 646–654 (2009).

68. P. Yao, J. Shen, L. Chen, S. Ge, Q. Xiong, Cortical ensemble activity discriminates auditory attentional states. *Molecular Brain* **12**, 80 (2019).

69. D. Saderi, Z. P. Schwartz, C. R. Heller, J. R. Pennington, S. V. David, Dissociation of task engagement and arousal effects in auditory cortex and midbrain. *eLife* **10**, e60153 (2021).

70. M. Berto, E. Ricciardi, P. Pietrini, N. Weisz, D. Bottari, Distinguishing fine structure and summary representation of sound textures from neural activity. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.03.17.484757.

71. F. García-Rosales, M. J. Beetz, Y. Cabral-Calderin, M. Kössl, J. C. Hechavarria, Neuronal coding of multiscale temporal features in communication sequences within the bat auditory cortex. *Commun Biol* **1**, 1–14 (2018).

72. C. Kayser, M. A. Montemurro, N. K. Logothetis, S. Panzeri, Spike-Phase Coding Boosts and Stabilizes Information Carried by Spatial and Temporal Spike Patterns. *Neuron* **61**, 597–608 (2009).

73. K. M. M. Walker, J. K. Bizley, A. J. King, J. W. H. Schnupp, Multiplexed and Robust Representations of Sound Features in Auditory Cortex. *J. Neurosci.* **31**, 14565–14576 (2011).

74. S. B. Hofer, G. M. Klump, Within- and Across-Channel Processing in Auditory Masking: A Physiological Study in the Songbird Forebrain. *J. Neurosci.* **23**, 5732–5739 (2003).

75. E. L. Bartlett, S. Sadagopan, X. Wang, Fine frequency tuning in monkey auditory cortex and thalamus. *J Neurophysiol* **106**, 849–859 (2011).

76. T. Dau, S. Ewert, A. J. Oxenham, Auditory stream formation affects comodulation masking release retroactively. *The Journal of the Acoustical Society of America* **125**, 2182–2188 (2009).

77. S. A. Goldman, T. Baer, B. C. J. Moore, Effects of the selective disruption of within- and across-channel cues to comodulation masking release. *The Journal of the Acoustical Society of America* **130**, 2866–2873 (2011).

78. J. H. Grose, J. W. Hall, L. Mendoza, Perceptual organization in a comodulation masking release interference paradigm: exploring the role of amplitude modulation, frequency modulation, and harmonicity. *J Acoust Soc Am* **97**, 3064–3071 (1995).

79. D. Pressnitzer, R. Meddis, R. Delahaye, I. M. Winter, Physiological Correlates of Comodulation Masking Release in the Mammalian Ventral Cochlear Nucleus. *J. Neurosci.* **21**, 6377–6386 (2001).

80. S. M. A. Ernst, S. Uppenkamp, J. L. Verhey, Cortical representation of release from auditory masking. *Neuroimage* **49**, 835–842 (2010).

81. L. Las, E. A. Stern, I. Nelken, Representation of Tone in Fluctuating Maskers in the Ascending Auditory System. *J. Neurosci.* **25**, 1503–1513 (2005).

82. U. Langemann, G. M. Klump, Signal detection in amplitude-modulated maskers. I. Behavioural auditory thresholds in a songbird. *European Journal of Neuroscience* **13**, 1025–1032 (2001).

83. J. F. Culling, M. A. Stone, “Energetic Masking and Masking Release” in *The Auditory System at the Cocktail Party*, J. C. Middlebrooks, J. Z. Simon, A. N. Popper, R. R. Fay, Eds. (Springer International Publishing, Cham, 2017; https://doi.org/10.1007/978-3-319-51662-2_3)*Springer Handbook of Auditory Research*, pp. 41–73.

84. E. M. Kaya, M. Elhilali, Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience* **8** (2014).

85. B. J. Kwon, C. W. Turner, Consonant identification under maskers with sinusoidal modulation: Masking release or modulation interference? *The Journal of the Acoustical Society of America* **110**, 1130–1140 (2001).

86. J. C. Nocon, H. J. Gritton, N. M. James, R. A. Mount, Z. Qu, X. Han, K. Sen, Parvalbumin neurons enhance temporal coding and reduce cortical noise in complex auditory scenes. *Commun Biol* **6**, 1–14 (2023).

87. G. A. Ascoli, L. Alonso-Nanclares, S. A. Anderson, G. Barrionuevo, R. Benavides-Piccione, A. Burkhalter, G. Buzsáki, B. Cauli, J. DeFelipe, A. Fairén, D. Feldmeyer, G. Fishell, Y. Fregnac, T. F. Freund, D. Gardner, E. P. Gardner, J. H. Goldberg, M.

Helmstaedter, S. Hestrin, F. Karube, Z. F. Kisvárday, B. Lambolez, D. A. Lewis, O. Marin, H. Markram, A. Muñoz, A. Packer, C. C. H. Petersen, K. S. Rockland, J. Rossier, B. Rudy, P. Somogyi, J. F. Staiger, G. Tamas, A. M. Thomson, M. Toledo-Rodriguez, Y. Wang, D. C. West, R. Yuste, The Petilla Interneuron Nomenclature Group (PING), Petilla terminology: nomenclature of features of GABAergic interneurons of the cerebral cortex. *Nat Rev Neurosci* **9**, 557–568 (2008).

88. R. Yuste, M. Hawrylycz, N. Aalling, A. Aguilar-Valles, D. Arendt, R. Armañanzas, G. A. Ascoli, C. Bielza, V. Bokharaie, T. B. Bergmann, I. Bystron, M. Capogna, Y. Chang, A. Clemens, C. P. J. de Kock, J. DeFelipe, S. E. Dos Santos, K. Dunville, D. Feldmeyer, R. Fiáth, G. J. Fishell, A. Foggetti, X. Gao, P. Ghaderi, N. A. Goriounova, O. Güntürkün, K. Hagihara, V. J. Hall, M. Helmstaedter, S. Herculano-Houzel, M. M. Hilscher, H. Hirase, J. Hjerling-Leffler, R. Hodge, J. Huang, R. Huda, K. Khodosevich, O. Kiehn, H. Koch, E. S. Kuebler, M. Kühnemund, P. Larrañaga, B. Lelieveldt, E. L. Louth, J. H. Lui, H. D. Mansvelder, O. Marin, J. Martinez-Trujillo, H. M. Chameh, A. N. Mohapatra, H. Munguba, M. Nedergaard, P. Němec, N. Ofer, U. G. Pfisterer, S. Pontes, W. Redmond, J. Rossier, J. R. Sanes, R. H. Scheuermann, E. Serrano-Saiz, J. F. Staiger, P. Somogyi, G. Tamás, A. S. Tolias, M. A. Tosches, M. T. García, C. Wozny, T. V. Wuttke, Y. Liu, J. Yuan, H. Zeng, E. Lein, A community-based transcriptomics classification and nomenclature of neocortical cell types. *Nat Neurosci* **23**, 1456–1468 (2020).

89. H. Zeng, What is a cell type and how to define it? *Cell* **185**, 2739–2755 (2022).

90. R. Drullman, J. M. Festen, R. Plomp, Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America* **95**, 1053–1064 (1994).

91. J. Obleser, C. Kayser, Neural Entrainment and Attentional Selection in the Listening Brain. *Trends in Cognitive Sciences* **23**, 913–926 (2019).

92. R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, M. Ekelid, Speech Recognition with Primarily Temporal Cues. *Science* **270**, 303–304 (1995).

93. C. K. Kim, A. Adhikari, K. Deisseroth, Integration of optogenetics with complementary methodologies in systems neuroscience. *Nat Rev Neurosci* **18**, 222–235 (2017).

94. J. P. Rickgauer, K. Deisseroth, D. W. Tank, Simultaneous cellular-resolution optical perturbation and imaging of place cell firing fields. *Nat Neurosci* **17**, 1816–1824 (2014).

95. A. R. Vaidya, M. S. Pujara, M. Petrides, E. A. Murray, L. K. Fellows, Lesion Studies in Contemporary Neuroscience. *Trends in Cognitive Sciences* **23**, 653–671 (2019).

96. J. S. Wiegert, M. Mahn, M. Prigge, Y. Printz, O. Yizhar, Silencing Neurons: Tools, Applications, and Experimental Constraints. *Neuron* **95**, 504–529 (2017).

97. M. Jazayeri, A. Afraz, Navigating the Neural Space in Search of the Neural Code. *Neuron* **93**, 1003–1014 (2017).

98. E. Jonas, K. P. Kording, Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology* **13**, e1005268 (2017).

99. A. Forli, D. Vecchia, N. Binini, F. Succol, S. Bovetti, C. Moretti, F. Nespoli, M. Mahn, C. A.

Baker, M. M. Bolton, O. Yizhar, T. Fellin, Two-Photon Bidirectional Control and Imaging of Neuronal Excitability with High Spatial Resolution *In Vivo*. *Cell Reports* **22**, 3087–3098 (2018).

100. A. M. Packer, L. E. Russell, H. W. P. Dalgleish, M. Häusser, Simultaneous all-optical manipulation and recording of neural circuit activity with cellular resolution in vivo. *Nat Methods* **12**, 140–146 (2015).

101. C. K. Pfeffer, M. Xue, M. He, Z. J. Huang, M. Scanziani, Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nat Neurosci* **16**, 1068–1076 (2013).

102. L. Ouellet, E. de Villers-Sidani, Trajectory of the main GABAergic interneuron populations from early development to old age in the rat primary auditory cortex. *Frontiers in Neuroanatomy* **8** (2014).

103. B. W. Brunton, M. M. Botvinick, C. D. Brody, Rats and humans can optimally accumulate evidence for decision-making. *Science* **340**, 95–98 (2013).

104. S. Grødem, I. Nymoen, G. H. Vatne, F. S. Rogge, V. Björnsdóttir, K. K. Lensjø, M. Fyhn, An updated suite of viral vectors for in vivo calcium imaging using intracerebral and retro-orbital injections in male mice. *Nat Commun* **14**, 608 (2023).

105. H. Tsukano, M. Horie, R. Hishida, K. Takahashi, H. Takebayashi, K. Shibuki, Quantitative map of multiple auditory cortical regions with a stereotaxic fine-scale atlas of the mouse brain. *Sci Rep* **6**, 22315 (2016).

106. A. Grinvald, A. Shmuel, I. Vanzetta, E. Shtoyerman, D. Shoham, A. Arieli, “Intrinsic signal imaging in the neocortex” in *Yuste, R., Lanni F., Konnerth, A. (Eds.) Imaging Neurons* (1999).

107. A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, M. Bethge, DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat Neurosci* **21**, 1281–1289 (2018).

108. S. Romero, A. E. Hight, K. K. Clayton, J. Resnik, R. S. Williamson, K. E. Hancock, D. B. Polley, Cellular and Widefield Imaging of Sound Frequency Organization in Primary and Higher Order Fields of the Mouse Auditory Cortex. *Cerebral Cortex* **30**, 1603–1622 (2020).

109. Q. Zhang, H. Liu, J. McGee, E. J. Walsh, G. A. Soukup, D. Z. Z. He, Identifying MicroRNAs Involved in Degeneration of the Organ of Corti during Age-Related Hearing Loss. *PLOS ONE* **8**, e62786 (2013).

110. M. L. Scattoni, S. U. Gandhy, L. Ricceri, J. N. Crawley, Unusual Repertoire of Vocalizations in the BTBR T+tf/J Mouse Model of Autism. *PLOS ONE* **3**, e3067 (2008).

111. C. Perrodin, C. Verzat, D. Bendor, Courtship behaviour reveals temporal regularity is a critical social cue in mouse communication. *eLife* **12** (2023).

112. A. K. Moore, M. Wehr, Parvalbumin-Expressing Inhibitory Interneurons in Auditory Cortex

Are Well-Tuned for Frequency. *J. Neurosci.* **33**, 13713–13723 (2013).

113. T. Olsen, A. R. Hasenstaub, Offset Responses in the Auditory Cortex Show Unique History Dependence. *J Neurosci* **42**, 7370–7385 (2022).

114. C. Trainito, C. von Nicolai, E. K. Miller, M. Siegel, Extracellular Spike Waveform Dissociates Four Functionally Distinct Cell Classes in Primate Cortex. *Current Biology* **29**, 2973-2982.e5 (2019).

115. A. Alishbayli, *Processing of Statistically Defined Sounds in the Auditory Cortex* (S.l. : s.n., 2024; https://repository.ubn.ru.nl/handle/2066/301602)

Acknowledgements: We would like to express our gratitude to Amber van der Stam, Dionne Lenferink, and Tom van Iersel for their assistance with animal training. BE also acknowledges valuable discussions at the Kavli Conference on Statistical Learning in the Brain at UCSB, supported by NSF Grant No. PHY-1748958 and the Gordon and Betty Moore Foundation Grant No. 2919.02 to the Kavli Institute for Theoretical Physics (KITP). The present work is based on Chapter 2 of the first author's PhD thesis *Processing of Statistically Defined Sounds in the Auditory Cortex* (*115*).

Funding: This work was supported by funding to BE from the NWO VIDI grant (016.VIDI.189.052) and the NWO ALW Open grant (ALWOP.146).

Author contributions:

Conceptualization:	AA, KP, BE
Methodology:	AA, KP, PvH, BE
Investigation:	AA, KP, PvH, BE
Software:	AA, KP, PvH, BE
Data curation:	AA, KP, PvH, BE
Validation:	AA, KP, PvH, BE
Supervision:	AA,BE
Resources:	AA, BE
Formal analysis:	AA
Visualization:	AA, BE
Writing—original draft:	AA, BE
Writing—review & editing:	AA, KP, PvH, BE
Project administration:	BE
Funding acquisition:	BE

Competing Interests: The authors declare they have no competing interest.

Data and Materials Availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The code used for generating the figures in this manuscript are available in the Zenodo Repository at **https://doi.org/10.5281/zenodo.14963093**