Distinct neural ensemble response statistics are associated with recognition and discrimination of natural sound textures

Xiu Zhai[a,b], Fatemeh Khatami[b,c,1], Mina Sadeghi[a,1], Fengrong He[b], Heather L. Read[b,d,e], Ian H. Stevenson[b,d,e], and Monty A. Escabí[a,b,d,e,2]

[a]Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269; [b]Biomedical Engineering, University of Connecticut, Storrs, CT 06269; [c]Bioengineering Department, School of Engineering, University of the Pacific, Stockton, CA 95211; [d]Psychological Sciences, University of Connecticut, Storrs, CT 06269; and [e]Connecticut Institute for Brain and Cognitive Sciences, University of Connecticut, Storrs, CT 06269

Edited by Robert J. Zatorre, Montreal Neurological Institute, McGill University, Montreal, QC, Canada, and accepted by Editorial Board Member Thomas D. Albright August 21, 2020 (received for review March 25, 2020)

The perception of sound textures, a class of natural sounds defined by statistical sound structure such as fire, wind, and rain, has been proposed to arise through the integration of time-averaged summary statistics. Where and how the auditory system might encode these summary statistics to create internal representations of these stationary sounds, however, is unknown. Here, using natural textures and synthetic variants with reduced statistics, we show that summary statistics modulate the correlations between frequency organized neuron ensembles in the awake rabbit inferior colliculus (IC). These neural ensemble correlation statistics capture high-order sound structure and allow for accurate neural decoding in a single trial recognition task with evidence accumulation times approaching 1 s. In contrast, the average activity across the neural ensemble (neural spectrum) provides a fast (tens of milliseconds) and salient signal that contributes primarily to texture discrimination. Intriguingly, perceptual studies in human listeners reveal analogous trends: the sound spectrum is integrated quickly and serves as a salient discrimination cue while high-order sound statistics are integrated slowly and contribute substantially more toward recognition. The findings suggest statistical sound cues such as the sound spectrum and correlation structure are represented by distinct response statistics in auditory midbrain ensembles, and that these neural response statistics may have dissociable roles and time scales for the recognition and discrimination of natural sounds.

natural sounds | auditory textures | sound statistics | neural coding | perception

What makes a sound natural, and what are the neural codes that support recognition and discrimination of real-world natural sounds? Although it is known that the early auditory system decomposes sounds along fundamental acoustic dimensions such as intensity and frequency, the higher-level neural computations that mediate natural sound recognition are poorly understood. This general lack of understanding is in part attributed to the structural complexity of natural sounds, which is difficult to study with traditional auditory test stimuli, such as tones, noise, or modulated sequences. Such stimuli can reveal details of the neural representation for relatively low-level acoustic cues, yet they don't capture the rich and diverse statistical structure of natural sounds. Thus, they cannot reveal many of the computations associated with higher-level sound properties that facilitate auditory tasks such as natural sound recognition or discrimination. A class of stationary natural sounds termed textures, such as the random sounds emanating from a running stream, a crowded restaurant, or a chorus of birds, have been proposed as alternative natural stimuli which allow for manipulating high-level acoustic structure (1). Texture sounds are composed of spatially and temporally distributed acoustic elements that are collectively perceived as a single source and are defined by their statistical features. Identification of these natural sounds has been proposed to be mediated through the integration of time-averaged summary statistics, which account for high-level structures such as the sparsity and time-frequency correlation structure found in many natural sounds (1–3). Using a generative model of the auditory system to measure summary statistics from natural texture sounds, it is possible to synthesize highly realistic synthetic auditory textures (1). This suggests that high-order statistical cues are perceptually salient and that the brain might extract these statistical features to build internal representations of sounds.

Although neural activity throughout the auditory pathway is sensitive to a variety of statistical cues such as the sound contrast, modulation power spectrum, and correlation structure (4–12), how sound summary statistics contribute toward basic auditory tasks such as recognition and discrimination of sounds is poorly understood. Furthermore, it is unclear where along the auditory pathway summary statistics are represented and how they are reflected in neural activity. The inferior colliculus (IC) is one candidate midlevel structure for representing such summary statistics. As the principal midbrain auditory nucleus, the IC

Significance

Being able to recognize and discriminate natural sounds, such as from a running stream, a crowd clapping, or ruffling leaves, is a critical task of the normal functioning auditory system. Humans can easily perform such tasks, yet they can be particularly difficult for the hearing impaired and they challenge our most sophisticated computer algorithms. This difficulty is attributed to the complex physical structure of such natural sounds and the fact they are not unique: they vary randomly in a statistically defined manner from one excerpt to the other. Here we provide evidence that the central auditory system is able to encode and utilize statistical sound cues for natural sound recognition and discrimination behaviors.

Author contributions: X.Z., M.S., H.L.R., I.H.S., and M.A.E. designed research; X.Z., F.K., M.S., F.H., and M.A.E. performed research; X.Z., F.K., M.S., I.H.S., and M.A.E. analyzed data; X.Z., I.H.S., and M.A.E. wrote the paper; and F.H. assisted with neural data acquisition.

Competing interest statement: H.L.R. has ownership interest in Elemind Technologies, Inc. and this private company did not sponsor this research.

This article is a PNAS Direct Submission. R.J.Z. is a guest editor invited by the Editorial Board.

[1]F.K. and M.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: escabi@engr.uconn.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2005644117/-/DCSupplemental.

receives highly convergent brainstem inputs with varied sound selectivities. Neurons in the IC are selective over most of the perceptually relevant range of sound modulations and neural activity is strongly driven by multiple high-order sound statistics (4–7, 10). In previous work, we showed the correlation statistics of natural sounds are highly informative about stimulus identity and they appear to be represented in the correlation statistics of auditory midbrain neuron ensembles (4). Correlations between neurons have also been proposed as mechanisms for pitch identification (13) and sound localization (14). This broadly supports the hypotheses that high-order sound statistics are reflected in the response statistics of neural ensembles and that these neural response statistics could potentially subserve basic auditory tasks.

Here using natural and synthetic texture sounds, we test the hypothesis that statistical structure in natural texture sounds modulates the response statistics of neural ensembles in the IC of unanesthetized rabbits, and that distinct neural response statistics have the potential to contribute toward sound recognition and discrimination behaviors. By comparing the performance of neural decoders with human texture perception, we find that place rate representation of sounds (neural spectrum) accumulates evidence about the sounds on relatively fast time scales (tens of milliseconds) exhibiting decoding trends that mirror those seen for human texture discrimination. High-order statistical sound cues, by comparison, are reflected in the correlation statistics of neural ensembles, which require substantially longer evidence accumulation times (>500 ms) and follow trends that mirror those measured for human texture recognition. Collectively, the findings suggest that spectrum cues and accompanying place rate representation (neural spectrum) may contribute surprisingly little toward the recognition of auditory textures. Instead, high-order statistical sound structure is reflected in the distributed patterns of correlated activity across IC neural ensembles and such neural response structure has the potential to contribute toward the recognition of natural auditory textures.

Results

Natural Sound Texture Statistics Modulate Neural Correlation but Not Neural Spectrum Statistics. To determine how natural sound statistics influence the response statistics of neural ensembles in IC, we first characterized several key statistics from an auditory

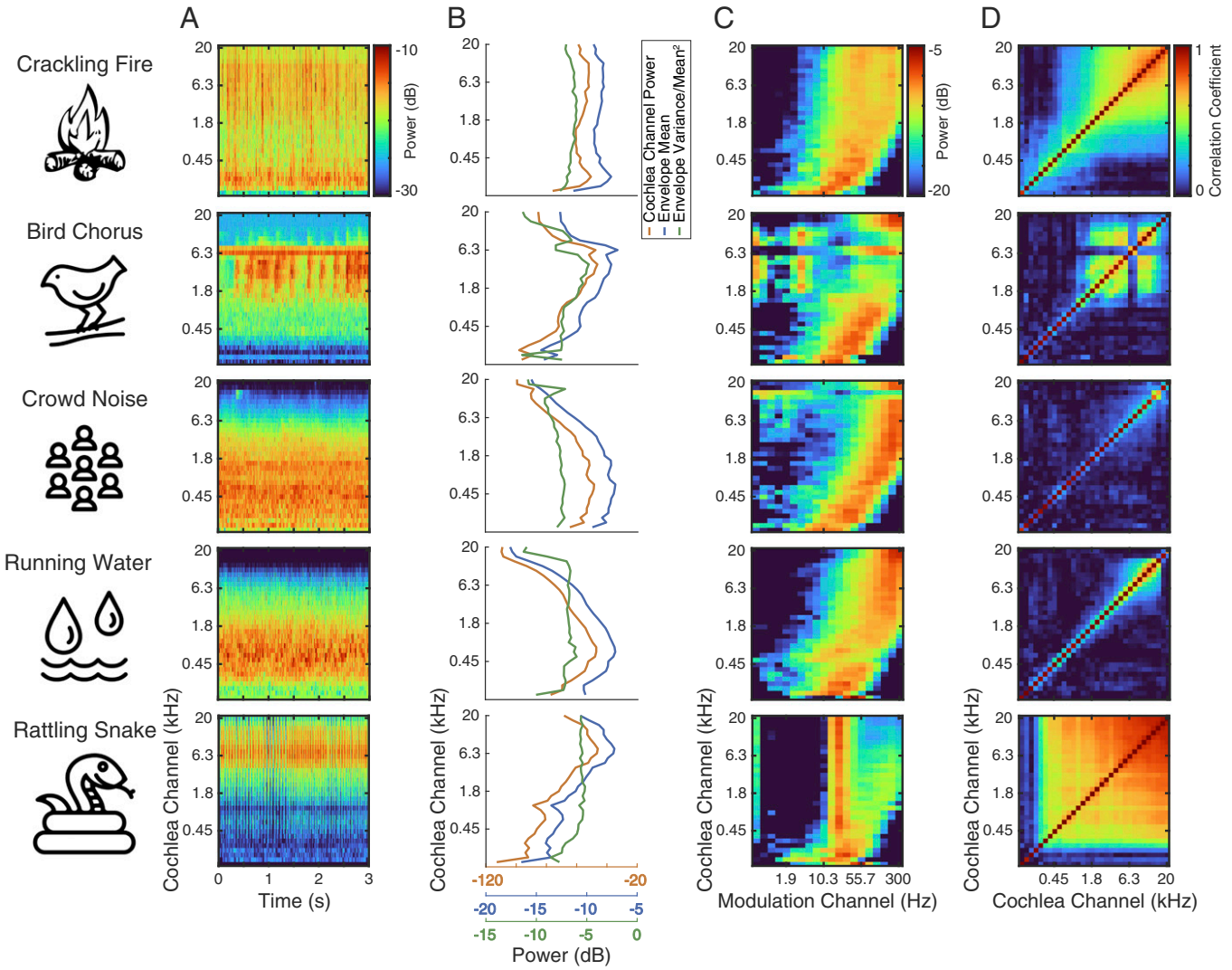

Fig. 1. Sound summary statistics measured from five natural sound textures. (*A*) Cochlear model spectrograms. Color scale indicates the time-varying envelopes decomposed from frequency-organized cochlear filters. (*B*) Cochlear channel power and envelope marginal moments (mean and variance/mean^2). (*C*) Modulation power spectrum. The power of each modulation band is normalized by the variance of the corresponding cochlear envelope and plotted as a function of modulation frequency (Hz). (*D*) Cochlear cross-band envelope correlations.

model representation for five natural sound recordings. These include sounds from a crackling fire, bird chorus, outdoor crowd, running water, and a rattling snake (*SI Appendix* and Sounds S1–S5). We then used texture synthesis (1) to generate synthetic sound variants with perturbed low- and high-order statistics (*Methods*). The selected sound textures have distinct spectral and temporal properties, and they each show diverse structures in the measured statistics (Fig. 1). The synthetic sound variants were generated by sequentially imposing the cochlear channel power statistics (i.e., power spectrum, Spec condition; Fig. 1 *A* and *B*, *SI Appendix*, and Sounds S6–S10), cochlear channel marginal statistics (+Mar condition; envelope mean, variance, and skew; Fig. 1*B*, *SI Appendix*, and Sounds S11–S15), the modulation power of each cochlear channel (+MPS condition; Fig. 1*C*, *SI Appendix*, and Sounds S16–S20), and the correlation structure between cochlear and modulation channels (+Corr condition; Fig. 1*D*, *SI Appendix*, and Sounds S21–S25). As illustrated, there are marked differences in both low- and high-order statistical cues across the different natural sounds. For instance, the crowd and water sounds have relatively similar power spectra (Fig. 1 *A* and *B*) and modulation power (Fig. 1*C*), and the correlations between frequency channels are weak, as reflected in the diagonalized correlation matrix (Fig. 1*D*). The rattling snake sound, by comparison, has a power spectrum that is biased toward higher frequencies (Fig. 1 *A* and *B*). Furthermore, the cochlear channels are highly correlated across frequencies (Fig. 1*D*) with modulation power concentrated at about 20 Hz (Fig. 1*C*), which reflects the coherent periodicity of the broadband rattling sound. Collectively, the differences in statistical structure for these natural textures could differentially drive neural responses in IC and may contribute toward recognition or discrimination of these sounds.

Though human listeners are perceptually sensitive to statistical cues in natural sound textures, there is little evidence on how neural responses to these statistics may contribute to texture perception. Here, we used the synthetic and original texture sounds as stimuli to determine whether sound statistics modulate the response statistics of neural ensembles in the unanesthetized rabbit IC ($n = 4$ animals, 29 penetration sites were included for analysis). Fig. 2 demonstrates the neural response statistics measured from a representative penetration site for a synthetic bird sound (synthesized variant that includes all of the statistics) using an analog representation of multiunit activity (aMUA) (*Methods*). From the response neurogram (Fig. 2*B*, average activity across response trials) we estimated the stimulus-driven neural correlation statistics by correlating the recorded aMUA signals from all electrode channel pairs across independent response trials (*Methods*). Fig. 2 *F* and *G* illustrate different degrees of correlated neural activity from example recording channel pairs in response to the synthetic bird chorus texture sound. The neural responses from channels 7 and 8, for instance, show a significantly correlated temporal signature (Fig. 2*F*; $P < 0.01$, Fisher z-transform test) which is reflected in the pointwise scatterplot and ultimately in the measured correlation coefficient ($r = 0.89$) and stimulus driven correlation ($c = 0.286$). Similarly, channel 7 shows a significant although weaker correlation with channels 11 ($r = 0.33$, $c = 0.11$) and 16 ($r = 0.16$, $c = 0.05$). Since our recordings from multiple spatially separated electrode channels follow the tonotopic ordering of the IC (Fig. 2*A*, frequency response areas cover a frequency range of approximate 0.5 to 10 kHz for this penetration site), we refer to the cross-channel correlations at zero lag as spectral correlations (Fig. 2*C*). Conceptually, this metric captures the degree to which distinct neural recording channels are temporally synchronous with one another (4, 15), analogous to the model-based channel correlation (16) (Fig. 1*D*). We also estimated the neural response correlations across time for each neural recording channel which we refer to as the temporal correlations (Fig. 2*D* and *Methods*). As previously shown, the temporal correlations capture the stimulus-driven temporal response pattern for each individual channel (4) and are closely related to the sound modulation power spectrum (MPS, related via a Fourier transform) (1). Thus, although the temporal correlations are computed directly in the time domain, they are mathematically equivalent to the power spectrum of the neural responses, and thus can be thought of as a neural equivalent of the MPS. Both the spectral and temporal correlations shown here are "stimulus"-driven correlations, where "noise" correlations on single trials have been removed by trial shuffling (*Methods*). Finally, for each recording location, we also measured the neural spectrum statistic assessed by computing the average response amplitude from each electrode channel over time and across response trials (Fig. 2*E* and *Methods*). This neural response statistic closely resembles the model-based sound spectrum which is widely used to measure the frequency composition of sounds.

If higher-order sound statistics provide meaningful cues for identifying sounds, neural responses should reflect and vary systematically with statistical variation of natural sounds. As seen in the example of Fig. 1, high-order natural sound statistics vary markedly for each natural texture, which could provide useful information about the identity of the sound. The measured neural correlations and neural spectrum vary markedly across the five texture sounds and across synthetic variants with different statistics as seen for the penetration site of Fig. 2 (Fig. 3; additional examples in *SI Appendix*, Fig. S1). In general, spectral (Fig. 3*A*) and temporal (Fig. 3*B*) correlations to different natural sound textures are highly diverse and reflect stimulus-dependent structures. For example, in the synthetic fire sound containing the full set of statistics (Fig. 3 column of +Corr condition), spectral correlations are extensive with the envelopes of both nearby and distant electrode channels showing correlated activity. By comparison, the neural correlations to the crowd sound are localized to neighboring channels with relatively low frequencies, and the response to the snake sound exhibits strong correlated activity between channels with best frequencies above ~1 kHz. Temporal correlations also show distinct, stimulus-dependent patterns. The temporal correlations of the bird sound, for instance, show a broad/slow component at high frequencies, while the correlations of fire, crowd, and water sounds are narrow/fast and show little frequency selectivity. In contrast to all four other sounds, the snake sound exhibits periodic correlations for high frequency channels that reflect the periodic structure of snake rattling at ~20 Hz (~50-ms period; Fig. 3 *B*, *Bottom*). The neural spectra (Fig. 3*C*) show a somewhat lower amount of diversity across sounds. The fire, bird, and snake sounds induce the strongest activity for high-frequency channels, while crowd and water more strongly drive low-frequency channels. Thus, neural response correlations reflect statistical structure that can potentially distinguish each of the natural texture sounds.

In addition to the differences in response statistics observed for each texture sound, neural correlation statistics in the IC also varied systematically with the sound statistics that were included in the perturbed texture variants. Adding sound statistics to the synthetic variants increases the strength of spectral and temporal correlations between neural responses (Fig. 3 *A* and *B*). Moreover, the patterns of neural correlations change and become increasingly similar to the original sound responses as statistics are added (Fig. 3 *A* and *B*). The same is not true for the neural spectrum, which is largely unaffected by adding high-order statistics to the synthetic sound variants and consistently resembles the response to the original sound (Fig. 3*C*). To quantify this effect, we calculated two indices that compare the neural correlations and spectrum for the reduced and original sound variants: a cross-validated similarity index (SI) and strength ratio (SR) (*Methods*). In the example penetration site, the SI and SR of the spectral (Fig. 3*D*) and temporal (Fig. 3*E*) correlations for

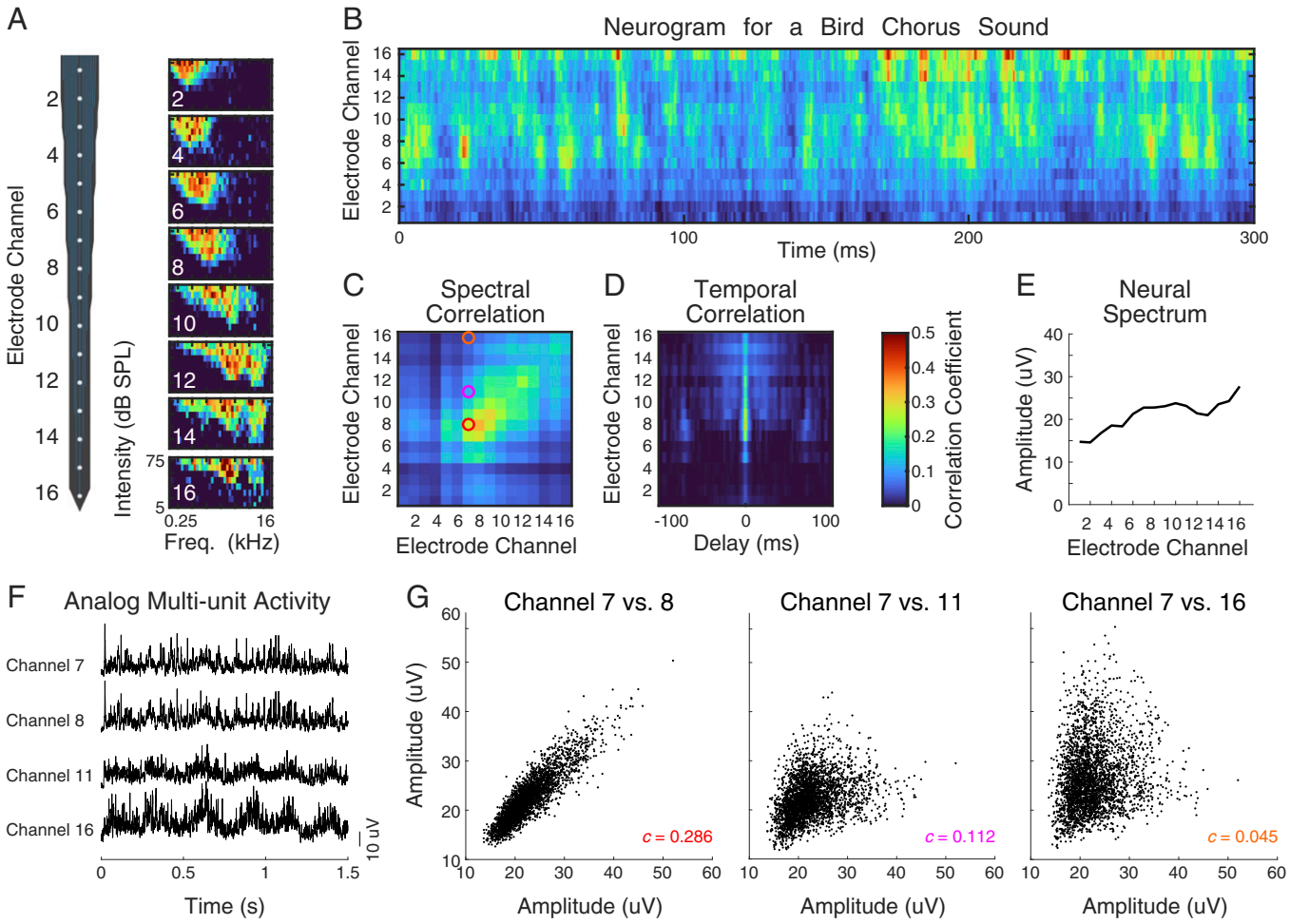

Fig. 2. Measuring neural response statistics from auditory midbrain ensembles. (*A*) Linear electrode array used for neurophysiological recordings. Frequency response areas are shown for eight of the recording channels in a representative tonotopically organized inferior colliculus (IC) penetration site (red indicates high activity and blue indicates low activity). (*B*) Segment of aMUA neural signals across 16 recording channels (red indicates strong response and blue indicates weak response) for a synthetic bird chorus sound texture (full statistics condition; *Methods*). (*C* and *D*) Stimulus-driven spectral and temporal neural ensemble correlations obtained from the aMUA. Spectral correlations are obtained by correlating the aMUA signals for different channels while temporal correlations are obtained by computing autocorrelations of individual channels (*Methods*). (*E*) The neural spectrum is calculated from the aMUA by computing the mean response for each of the 16 channels. (*F*) aMUA signals for four example recording channels from *B*. (*G*) Scatterplots using the aMUA signals in *F* for three different channel pairs illustrate different degrees of correlation. The corresponding pixels in the spectral correlation matrix shown in *C* have the stimulus-driven correlation coefficients of 0.286, 0.112, and 0.045, respectively. SPL, sound pressure level.

fire, bird, and snake sounds increase as more statistics are included in the sound variants. Thus, the correlation strength increases upon adding high-order statistics while the correlation pattern converges on that of the original sound. By comparison, for crowd and water sounds, the neural correlations do not change substantially with the included sound statistics, indicating that the neural correlation statistics are similar to the response of original sounds even when only the sound spectrum is imposed. Such lack of sequential change for these sounds reflects the fact that the crowd and water sounds are both well characterized by relatively low-level acoustic structure. They have minimal acoustic correlations and are relatively random over time (Fig. 1). By comparison, although the neural spectrum shows a fair amount of diversity across sound textures, it is very stable regardless of which statistics are included in the synthetic variants (Fig. 3*F*). Consequently, the measured SI and SR of the neural spectrum for this recording site are near constant for all five sounds.

Similar trends were observed across the neural population indicating that high-order statistics of the sounds strongly influence the neural ensemble activity (Fig. 3 *G*–*I*; for individual sounds, see *SI Appendix*, Fig. S2). Neural correlations changed systematically and converge on that of the original sound upon adding sound statistics, whereas the neural spectrum is relatively stable and resembles the original sound condition regardless of which statistics are included in the reduced sounds (Fig. 3 *G*–*I*). Averaged across all five sounds and penetration sites, the SI of spectral correlations shows a significant increase from 0.26 ± 0.26 (mean ± SD), when only the sound spectrum is included (i.e., Spec condition), to 0.78 ± 0.14, when the marginals, modulation power, and correlations are included (i.e., +Corr condition) in the synthetic sound stimuli ($P < 0.05$, paired t test comparing each condition against Spec; corrected for multiple comparisons, applied to this and all subsequent t tests). Similarly, the SI of temporal correlations increases systematically from −0.04 ± 0.11 to 0.58 ± 0.09 ($P < 0.05$, paired t test). The SR of the neural correlations statistics also systematically increases upon adding statistics to the synthetic variants (spectral = 0.55 ± 0.10 to 0.97 ± 0.13; temporal = 0.50 ± 0.16 to 0.90 ± 0.12; $P <$ 0.05, paired t test). These results differ from those for the neural spectrum, which exhibits only a slight increase in SI (from 0.84 ± 0.10

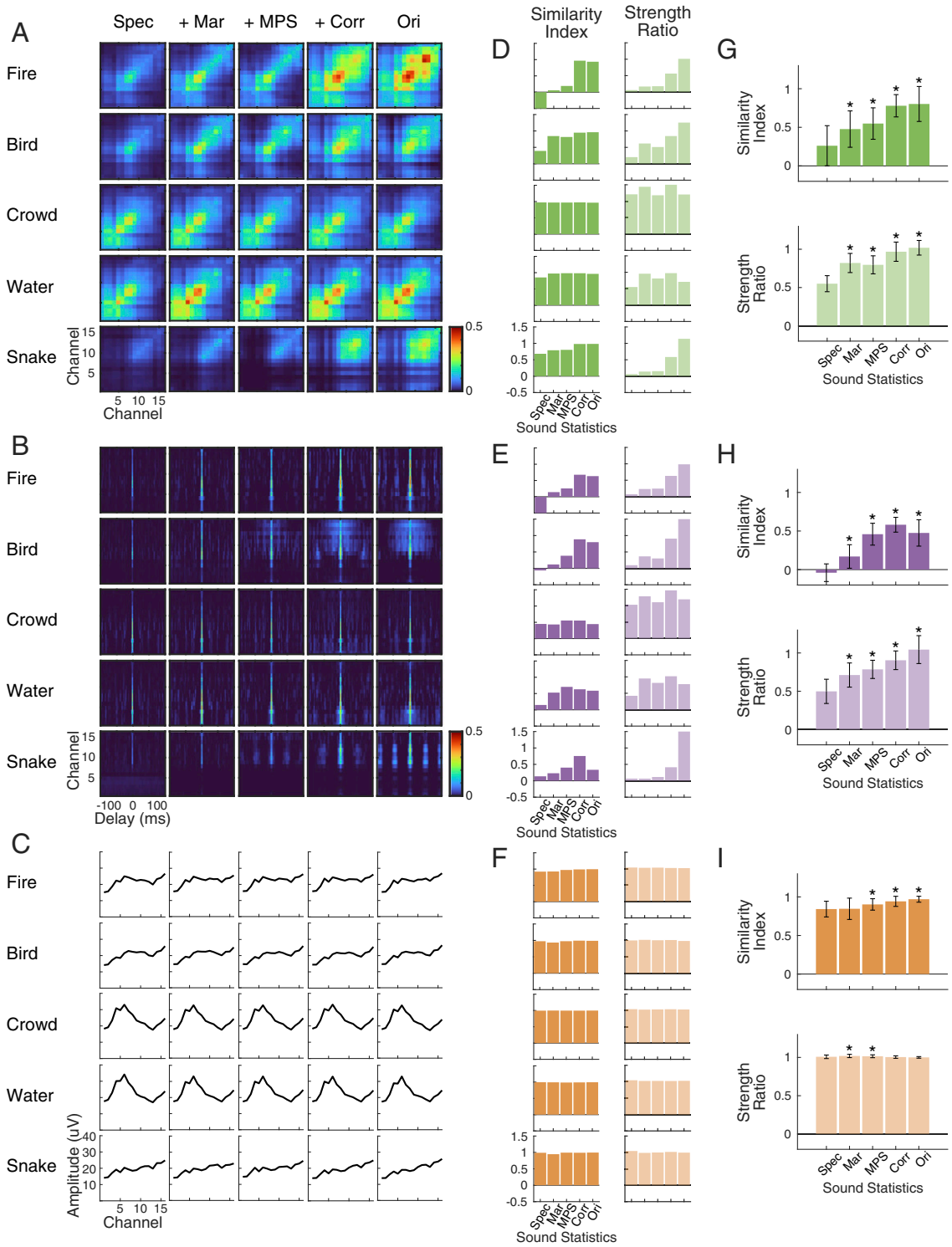

Fig. 3. Neural correlations and neural spectrums for synthetic variants and original sounds. (*A*) Spectral correlations for the IC penetration site shown in Fig. 2 (same site for *B*–*F*). (*B*) Temporal correlations for the same recording site. Spectral and temporal correlations are shown for the original sound (Ori) and synthetic variants with different statistics included (Spec, +Mar, +MPS, +Corr). Neural correlation matrices show distinct structures and unique patterns across the five tested sounds, while for each sound, such stimulus-dependent correlations converge upon adding sound statistics. (*C*) Neural spectrums. Although the neural spectrums show diversity across sounds, they do not change significantly with sound statistics included. (*D*–*F*) Similarity index and strength ratio as a function of sound statistics for three response metrics. Comparisons are made between the response of a synthetic sound variant and that of the original sound. For spectral (*D*) and temporal (*E*) correlations, both similarity index and relative strength increase upon adding sound statistics to synthetic variants of fire, bird, and snake sounds. For neural spectrums (*F*), the similarity index and relative strength are relatively constant with different statistics for all sounds. (*G* and *H*) Average similarity index and strength ratio as a function of sound statistics across $n = 29$ penetration sites ($n = 4$, 9, 8, and 8 from four animals). Averaged across five sounds, the similarity index and strength ratio of spectral (*G*) and temporal (*H*) correlations increase with statistics, indicating the correlation matrices converge to the original sound structure upon adding sound statistics. The neural spectrum (*I*) similarity index and strength ratio do not show a significant change with added sound statistics. Error bars, SD. Asterisks indicate a statistical significance increase when compared to the Spec condition ($P < 0.05$, paired *t* test, corrected for multiple comparisons). Spec: cochlear spectrum; Mar: cochlear marginals; MPS: modulation power spectrum; Corr: correlation; Ori: original in this and all subsequent figures.

for Spec to 0.94 ± 0.06 for +Corr; $P < 0.05$, paired t test) and where the SR remains near constant as statistics are added (~1, all SD < 0.02; not significant [N.S.], paired t test). Overall, these findings demonstrate that neural correlations in IC become stronger and their pattern converges on that of the original sound textures upon adding high-order statistics while the neural spectrum is substantially less variable and is largely unaffected by high-order sound statistics.

The Contribution of Neural Response Correlation and Spectrum Statistics to Recognition and Discrimination of Texture Sounds. Given that the neural response pattern and strength is strongly modulated by the high-order texture statistics, we next explored how the response spectrum- and correlation-based neural codes can contribute to recognition and discrimination of sound textures. We used a single-trial neural decoder (*Methods*) to determine whether the neural correlations and spectrum in IC could allow sound textures to be identified and discriminated. Each of the classifiers was specifically designed to take into account task-specific structure and the information that would be available to a listener during a texture recognition or discrimination task (*Methods*). In the recognition task, a naïve Bayes classifier was trained with the neural spectrums or correlations of the five original sounds (obtained with a 1,000-ms window). The classifier was required to identify the delivered sound using single-response trials of a variable duration (62.5 to 1,000 ms in half-octave steps). Fig. 4*A* shows the cross-validated (half of the data were used for training the model and the other half for validation; *Methods*) classification performance for the example penetration site shown previously (Fig. 3) when the stimuli were synthetic sounds that included all statistics (+Corr condition, applied to Fig. 4 *A–D*; results for all other statistic conditions are shown in *SI Appendix*, Fig. S3). At short durations, the performance of neural correlation-based classifiers ranges from ~0 to 90% for different sounds, although the average performance is above chance (32.2%, 29.2%, and 32.2% for spectrotemporal, spectral, and temporal, respectively, at 62.5 ms; chance is 20%). For most sounds, classifier performance increases with sound duration. An exception to this, was the temporal classifier performance for the snake sound, which was below chance. This misclassification was likely due to the fact that the rattling frequency between the first and second half of the data were different (~15 vs. 25 Hz) which affected the cross-validation results (*Methods*). In contrast to the correlation-based classifiers, where performance tends to improve with sound duration, the neural spectrum classifier shows relatively stable and high performance even for short sound durations (67.0% at 62.5 ms).

Similar results were observed across all penetration sites (Fig. 4*C*). The performance of all classifiers is above chance and increases with sound duration. Among the correlation-based classifiers, the spectrotemporal classifier shows the highest performance (increase from 34.6 to 79.3% with duration), followed by the spectral (30.0 to 69.7%) and the temporal (31.5 to 67.3%) classifiers. In contrast, the neural spectrum classifier has more stable and higher performance (52.8 to 83.2%) that does not improve substantially with sound duration (a ~30% increase compared to 40 to 45% for the correlation-based classifiers). Thus, spectrum- and correlation-based neural codes can both contribute to texture recognition, and performance tends to improve with the sound duration.

Although texture recognition performance of the synthetic texture (+Corr condition) depends strongly on the sound duration and neural response statistics measured, texture discrimination was less dependent on both. In the texture discrimination task, the naïve Bayes classifier was trained using the responses for sound pairs of identical duration (*Methods*). For the example penetration site of Fig. 4*B*, most sounds are easily discriminated regardless of the sound duration used. One exception is for the water sound where, for this example, the temporal correlation-based classifier performance tends to decrease with duration, although it remains above chance (50%). Averaged across all sounds and all penetration sites (Fig. 4*D*), performance is high and above chance for all classifiers when the sound is 62.5 ms (spectrotemporal 70.0%, spectral 68.1%, temporal 60.1%, neural spectrum 88.0%) and shows slight increases with sound duration. The performance reaches and exceeds 90% with longer durations.

To further evaluate whether and the degree to which the sound spectrum itself may be driving correlated neural activity that might contribute toward the recognition and discrimination of textures, we repeated the experiments in $n = 11$ recording sites (from two animals) using the original sound textures and texture variants with an equalized 1/f power spectrum (4) (*SI Appendix* and Sounds S26–S30). This manipulation guarantees that average spectrum cues are the same across sounds while preserving many of the high-order sound cues in the original sounds. Despite the fact that these sounds have identical spectrum, the sounds are perceptually distinct and are uniquely perceived as the original sound. Here, the neural correlations of the equalized sounds are distinct from each other and remarkably similar to the original sounds (Fig. 5*A*, spectral; Fig. 5*B*, temporal). On the other hand, while the neural spectra of the original sounds are quite distinct, the neural spectra for the equalized sounds are much more similar to each other (Fig. 5*C*). Thus, although the neural spectrum is strongly affected by the sound spectrum, high-order structure in these equalized sounds appears to be largely preserved and encoded by the neural correlations within IC ensembles.

How does the removal of the sound spectrum affect classification performance? Fig. 5*D* shows the population average neural identification and discrimination for the spectrum equalized sounds. While the neural spectrum contributes substantially to recognition of both the synthetic and original sounds (Fig. 4*C*, far *Right*; *SI Appendix*, Fig. S3*A*, original [Ori], far *Right*) neural spectrum identification performance drops to near chance (from 90.1 to 26.6% at 1 s) for the spectrum equalized sounds (Fig. 5*D*, red). Recognition performance for the neural correlation also drops when compared to the original sounds (from 84.9 to 58.5% at 1 s) but still retains information for the spectrum equalized sounds that allows the classifier to perform well above chance (Fig. 5*D*, blue). Similar results are also shown for the discrimination task. The neural correlation classifier performance is only slightly reduced compared to the original sound (93.6% vs. 84.7% at 1 s), which is expected given that correlations are largely preserved for these equalized sounds. However, although the neural spectra of the equalized sounds are quite similar, there are still significant differences that enable the neural spectrum classifier to discriminate among the five sounds beyond chance (original condition = 96.0%; spectrum equalized = 79.3%). These findings suggest that high-order sound structure from the original sounds is retained in the neural correlations despite equalization and that this structure can contribute to neural recognition and discrimination independently of the cues in the sound spectrum.

Next, we explored how the addition of high-order sound statistics affects how well sound textures can be identified or discriminated from neural responses. The neural classifier performance at 1-s duration is shown in Fig. 6*A* as a function of the sound statistics that were included during the synthesis. For the recognition task, the performance of all classifiers increases substantially upon adding sound statistics (for spectral classifier and temporal classifier, see *SI Appendix*, Fig. S4). The spectrotemporal classifier shows the highest performance among the correlation-based classifiers, increasing from 40.0 ± 11.6% for the Spec condition to 83.0 ± 9.4% for Ori. This is followed by the spectral correlation-based classifier (39.2 ± 11.8% to 78.1 ± 10.2%) and the temporal correlation-based classifier (25.2 ± 6.0% to 63.6 ± 11.8%), which shows the lowest performance on average. In contrast to the correlation classifiers,

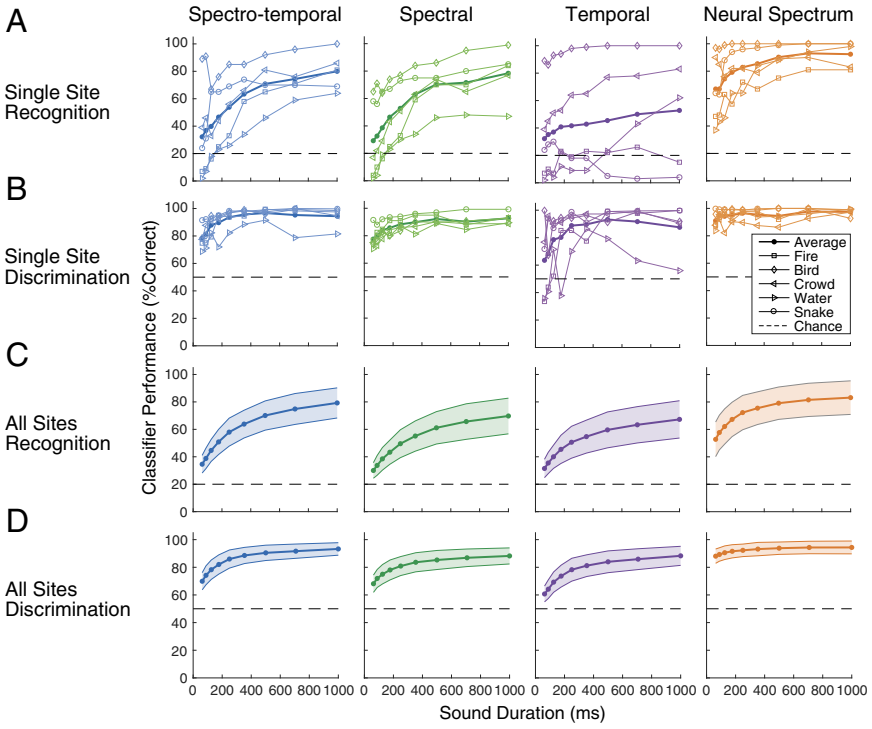

Fig. 4. Decoding neural correlations and neural spectrums for sound recognition and categorization. Classification results are shown as a function of sound duration. Stimuli were synthetic sounds with full statistics (+Corr condition; *Methods*). (*A*) Single-trial classification results for the recognition task (*Methods*) for the penetration site shown in Fig. 3. The performance of each individual sound (black curves) and the average results (red curve) are shown for different classifiers (spectrotemporal, spectral, temporal, and neural spectrum). (*B*) Classification results for the discrimination task for the same penetration site in *A*. (*C* and *D*) Average performance across $n = 29$ penetration sites. Shaded areas, SD.

the neural spectrum classifier exhibits very high performance that is much less dependent on the included statistics. The recognition accuracy is 65.1 ± 14.3% for the Spec condition and improves to 87.8 ± 8.4% for Ori condition. Note that the spectrotemporal and neural spectrum classifiers have roughly similar high performance (about 80%) for the synthetic sounds with correlation statistics and original sounds. However, in the discrimination task, the performance doesn't improve substantially with sound statistics for the spectrotemporal, spectral, and neural spectrum classifiers, and improves only slightly (~18%) for the temporal classifier. Together, these findings suggest that correlation- and spectrum-based cues can contribute differently to neural recognition and discrimination of texture sounds.

The evidence accumulation times for the neural correlation and spectrum decoders differed significantly as a function of task (recognition vs. discrimination), the statistics that were included in the synthetic texture stimuli, and the neural response statistics used for classification (neural correlation vs. spectrum) ($P < 0.001$, m-way ANOVA). Fig. 6*B* shows the time constant, defined as the time it takes to reach 90% of the corresponding maximum classifier performance, as a function of sound statistics. First, for both the neural spectrum and combined spectrotemporal correlation classifiers, time constants for the discrimination task are shorter than the recognition task and they exhibit different behaviors as a function of added sound statistics. In the recognition task, the time constant increases systematically with additional sound statistics (change between conditions Spec to Ori: 65.4% increase, from 355 ± 195 ms to 587 ± 135 ms for correlation; 53.2% increase, from 222 ± 143 ms to 340 ± 121 ms for neural spectrum). This systematic increase in the time constant is partly attributed to performance improvement with changing statistic, which is most prominent for the long sound durations (*SI Appendix*, Fig. S3). For short duration sounds, there is only a modest improvement in the classifier recognition performance, indicating that the classifier does not effectively make use of the added statistics for very short duration sounds. Regardless of task, it is worthwhile noting that the time constant of the neural spectrum classifier is ~200 to 300 ms faster than that of the neural spectrotemporal correlation classifiers (also true for spectral and temporal correlations; *SI Appendix*, Fig. S4). For the +Corr sound condition, for instance, the time constant is 581 ± 121 ms for the neural correlation and 350 ± 127 ms for the neural spectrum classifier. We also find a similar difference in the discrimination task, where the time constant is 235 ± 109 ms for the neural correlation classifier and 68 ± 14 ms for the neural spectrum classifier for the +Corr sound condition. However, for the discrimination task, the time constant is relatively stable and does not change substantially with sound statistics (change between conditions Spec to Ori: 28.8% decrease from 281 ± 101 ms to 200 ± 85 ms for neural correlation; 7.6% from 66 ± 7 ms to 71 ± 22 ms for neural spectrum). The dissociation between neural spectrum and correlation classifiers is not trivially due to the differences in the training procedure between the two task (e.g., different training sound durations; *Methods*), because the differences in time constant are observed within each task. Furthermore, matching the training sound duration for the recognition classifier does not appreciably affect the general trends (*SI Appendix*, Fig. S5), suggesting that the differences reflect task-specific information that is available in the neural spectrum and correlation signals. Thus overall, texture discrimination can be accomplished by the classifier more quickly than recognition, and spectrum-based cues accumulate evidence about the sound more quickly in both tasks.

Sound Texture Statistics Facilitate Recognition but Not Discrimination of Natural Sounds. A series of parallel studies using an identical sound paradigm were carried out to determine how human listeners discriminate and identify sound textures and to determine how different statistics contribute to both tasks. In the texture recognition task, subjects listened to one of five sounds and were asked to identify the sound they heard, whereas for texture discrimination, subjects listened to two sounds and were asked to report whether the sounds were the same or different (two-alternative forced choice [2AFC]; *Methods*). As for the neural classifier, distinct differences are observed in the texture recognition and discrimination tasks and distinct performance trends are observed across sound

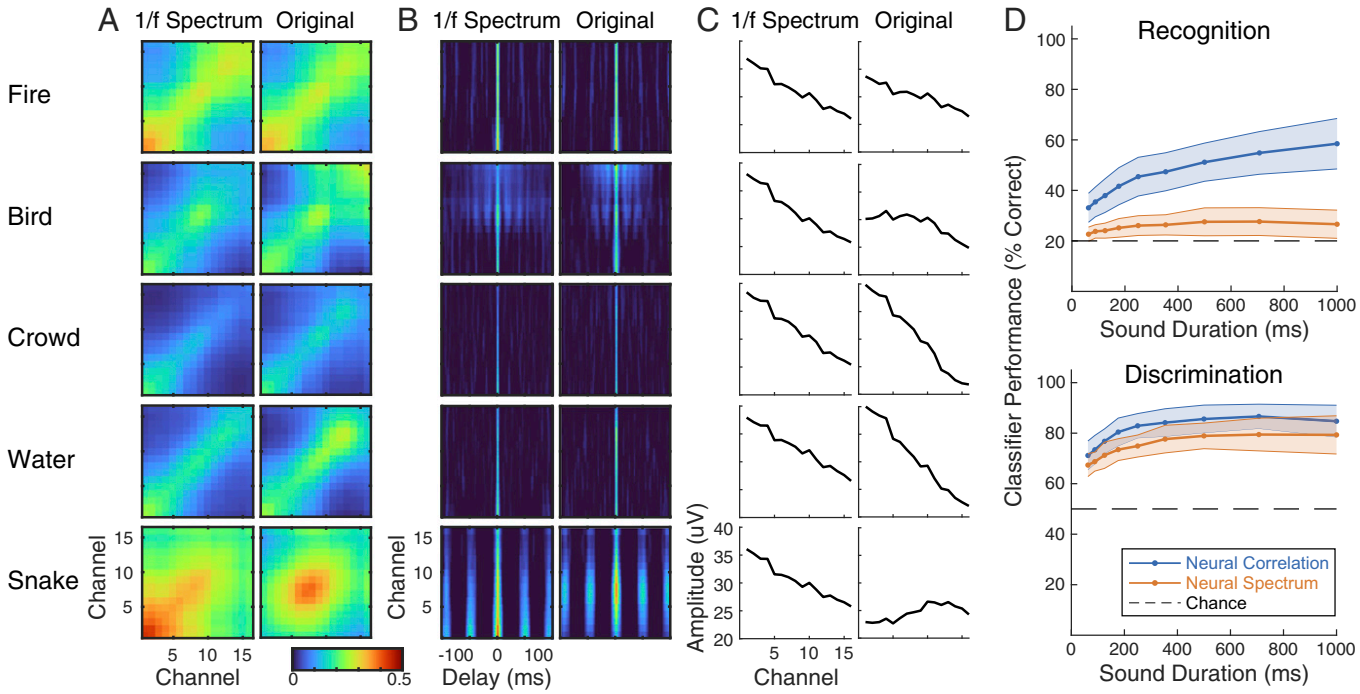

Fig. 5. Neural correlations and neural spectrums for sound variants with equalized power spectrum. (*A*–*C*) Response metrics of a single IC penetration site. Spectral (*A*) and temporal (*B*) neural correlations show similar structures between the 1/f equalized and original sounds. Neural spectrums (*C*) show nearly identical structures across sounds for the 1/f condition while they are different from those in the original sounds. (*D*) Average single-trial classification performance across $n = 11$ penetration sites ($n = 3$, 8 from two animals). In the recognition task, the neural classifier is trained using the responses to the original sounds, while the validation data are from the spectrum equalized sounds. In the discrimination task, both training and testing data are from the responses to the spectrum equalized sounds (*Methods*).

durations and added statistics (Fig. 7*A*; $P < 0.001$, m-way ANOVA). First, in the recognition task, performance tends to improve upon adding statistics to the synthetic variants (48% improvement from 50.0 ± 10.6% to 98.0 ± 2.7%, Spec to +Corr for the 1-s duration; $P <$ 0.05, paired *t* test) and a significant increase in recognition performance is also observed across sound durations (29% improvement from 71.0 ± 12.9% to 100.0 ± 0.0%, shortest to longest duration for the Ori condition; $P < 0.05$, paired *t* test). Thus, the including high-order statistics seem to have a pervasive role in the subject's ability to identify the textures used and evidence of these high-order statistics can accumulate over relatively long durations, consistent with prior studies reporting increased perceptual realism with added statistics and sound duration (1, 16). This behavior sharply contrasts human discrimination trends, where the performance is nearly maxed out and is much more homogenous. Performance improves only subtly upon adding statistics for the short duration sounds (93.0 ± 2.7% to 96.5 ± 3.4%, Spec to +Corr for the 62.5-ms duration; $P < 0.05$, paired *t* test), whereas it is nearly 100% for all statistics for the longest duration (98.5 ± 1.4% to 99.5 ± 1.0%, Spec to +Corr for the 1-s duration; N.S., paired *t* test). Thus overall, texture discrimination can be performed relatively quickly requiring few statistical cues and the spectrum cue on its own accounts for most of the discrimination performance. Texture recognition, by comparison, is greatly impacted by adding high-order statistics to the synthetic variants (beyond Spec) and such information can accumulate over time so that recognition performance is highest for the longest sounds.

The human perceptual trends resemble results from the neural classifier where discrimination is fast and recognition is slow and where multiple high-order statistics contribute most profoundly to recognition. For this reason, we compared neural classifier against the human listener trends using matched conditions. Fig. 7 *B* and *C* show the corresponding neural classification trends using the neural correlation (spectrotemporal; spectral-only and temporal-only correlation classifiers are shown in *SI Appendix*, Fig. S6) and neural spectrum classifiers. Intriguingly, several parallels between the neural classifier and human trends are observed. First, in the recognition task, performance varies significantly as a function of both duration and added statistics ($P < 0.001$, m-way ANOVA). The neural correlation classifier performance increases substantially with added statistics (39% improvement from 40.0 ± 11.6% to 79.3 ± 11.0%, Spec to +Corr for the 1-s duration; $P < 0.05$, paired *t* test) and duration (48% improvement from 35.4 ± 6.3% to 83.0 ± 9.4%, shortest to longest duration for the Ori condition; $P < 0.05$, paired *t* test) and the data for 1-s duration follows a similar trend to the human recognition data. The spectrum-based classifier performance likewise shows some improvements with statistics (from 65.1 ± 14.3% to 83.2 ± 12.2%, Spec to +Corr for the 1-s duration; $P < 0.05$, paired *t* test) and duration (from 55.0 ± 12.2% to 87.8 ± 8.4%, shortest to longest duration for the Ori condition; $P < 0.05$, paired *t* test), but these are more subtle and the performance trends are less similar than for recognition. In contrast, the general agreement between neural classifier and human performance swap for the discrimination task. Here, the correlation-based classifier shows graded improvements with both added statistics (from 86.5 ± 6.5% to 93.3 ± 4.5%, Spec to +Corr for the 1-s duration; $P < 0.05$, paired *t* test) and duration (from 72.6 ± 6.4% to 92.6 ± 4.8%, shortest to longest duration for the Ori condition; $P < 0.05$, paired *t* test), which are not observed for human results. By comparison, the spectrum-based classifier follows a nearly identical and much more similar trend to the human data where discrimination performance is much more homogenous across stimulus conditions and independent of the sound duration (from 88.5 ± 4.6% to 94.9 ± 4.0%, shortest to longest duration for the Ori condition; $P < 0.05$, paired *t* test) and statistics included (from 94.4 ± 4.6% to 94.4 ± 4.6%, Spec to +Corr for the 1-s duration; N.S., paired *t* test). Thus, overall, human perception and neural decoding in the auditory midbrain

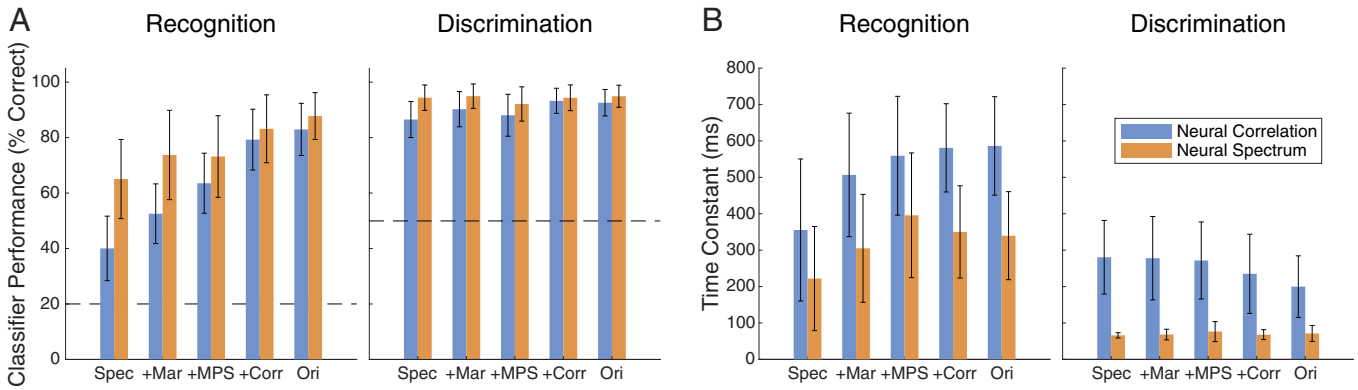

Fig. 6. Neural classifier performance and time constant as a function of sound statistics. (*A*) Classification results for the spectrotemporal and neural spectrum classifiers in the recognition and discrimination tasks. The average performance is shown for 1-s sound duration (averaged across sounds and penetration sites). Recognition performance improves substantially upon adding sound statistics to the synthetic variants. Performance for the discrimination task is much more stable and does not improve substantially with added statistics. (*B*) Time constant (time required to reach 90% of the maximum measured performance) as a function of sound statistics in the recognition and discrimination tasks. Time constants for the recognition task are substantially longer than the discrimination task. Error bars, SD.

follow similar trends where discrimination can be accomplished quickly with only spectrum-based cues. By comparison, recognition by both human listeners and the neural decoder is much more dependent on high-order statistics and requires longer evidence accumulation times to achieve high performance.

Discussion

The findings here demonstrate that the spectrum and high-order summary statistics of natural sounds are reflected in the response statistics of auditory midbrain ensembles. Results from the neural decoders are consistent with and predict perceptual patterns for recognition and discrimination in human listeners, suggesting that such neural activity can mediate auditory behaviors. Together the neural and behavioral results support the idea that the statistical structure in sounds is reflected in the neural spectrum and neural correlation statistics of IC neural ensembles and that such neural response statistics serve distinct roles for the recognition and discrimination of natural sounds.

Neural Representation of Natural Sound Textures. How does the brain represent natural sound textures? The synthetic textures used in this study are constructed from summary statistics of natural textures as measured with a relatively simple generative model of the auditory pathway, consisting of a stage of peripheral frequency-tuned filters that is followed by a stage of modulation filters. Given the sequential transformation from frequency decomposition in the cochlea to modulation decomposition along the ascending auditory pathway, it is likely that multiple auditory structures contribute to the extraction and representation of summary statistics. Here we have measured three neural response statistics from auditory midbrain ensembles (neural spectrum and spectral and temporal correlations) and shown that these can convey critical sound-related information.

The auditory midbrain is uniquely situated for representing natural sound summary statistics, given its central position in the ascending auditory pathway where multiple brainstem targets with distinct selectivities converge and where responses from single neurons have been shown to be modulated by various natural sound statistics (4–6, 10). The inferior colliculus is the first stage in the auditory pathway with a preponderance of modulation-tuned neurons (17) and where neurons are uniquely sensitive to correlation statistics of sounds (4, 7). Although correlated firing between neurons has been reported to be an inefficient mode of information transfer (redundant activity) and correlated firing has been proposed to diminish between the inferior colliculus and auditory cortex (18), correlated firing in IC is in fact substantially faster and is associated with fast temporal modulation features that cortical neurons do not synchronize to (15). As we demonstrate here, such correlated firing can be quite informative and may in fact provide critical information about the identity of natural sounds. Furthermore, the IC has a unique anatomical organization and circuitry, with neurons being topographically organized for frequency, frequency resolution, and modulation preferences along three anatomically orthogonal dimensions (17, 19), which as a population, could represent the modulation power spectrum of sounds (MPS statistics). Collectively, we have demonstrated that the neural ensembles in IC are sensitive to the high-order statistics of natural texture sounds and that response statistics from these frequency-organized neural ensembles may contribute toward recognition and discrimination of natural sound textures.

Although the mechanisms underlying the transformation from high-order sound statistics to neural ensemble statistics are not yet clear, specific response statistics may partly correspond to distinct statistics within the underlying texture synthesis model. For example, the temporal correlations of the neural ensemble are theoretically related to the modulation power spectrum summary statistic (via Fourier transform, Fig. 1*C*), while the spectral correlation reflects synchronous activity across frequency channels, analogous to the correlation summary statistic. It is important to note, however, that there are substantial differences between the auditory pathway and the simplified texture synthesis model that preclude a one-to-one mapping between sound and neural response statistics. For instance, the correlation sound statistics can affect the strength and pattern of neural correlations in IC (Fig. 3); the relationship between the sound correlation statistics and neural correlations is weak ($r^2 = 0.09$, figure S5 in ref. 4). Such disparity is likely due to the fact that IC neurons are selective to multiple acoustic features, including spectral and temporal modulations, and the fact that nonlinearities can strongly impact correlated activity in IC (15). By comparison, the model filters used to compute the correlation statistics are relatively simple and are strictly selective for the frequency content of the sound. Despite such differences, the exact form of the measured statistics may not be critical. More biologically realistic synthesis models with different summary statistics could certainly be selected and might more accurately predict behavior or the neural activity. More important to our conclusions is the general observation that statistics of the sound modulate neural

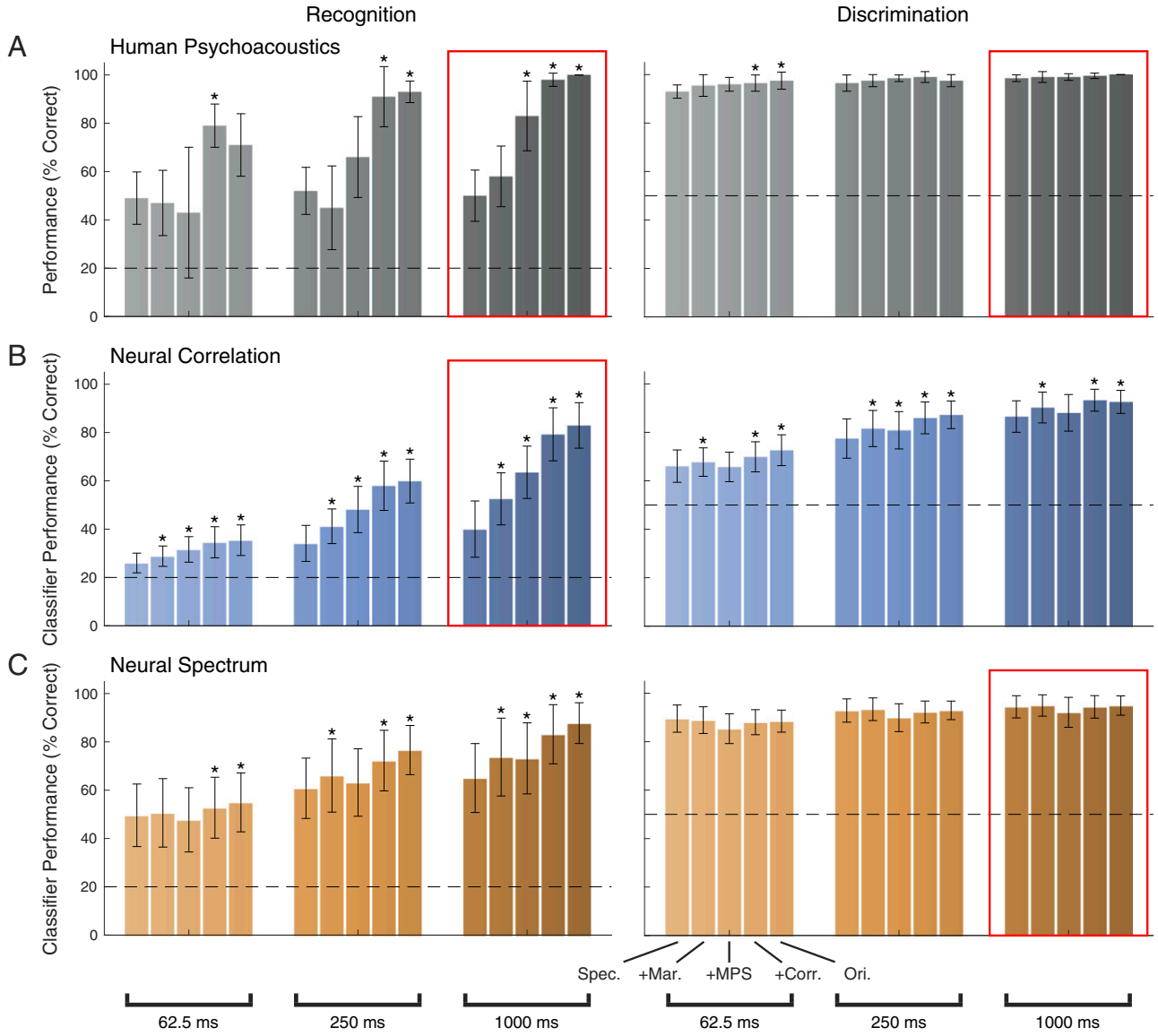

Fig. 7. Comparing the task performance between the neural classifiers and human psychoacoustics. (*A*) Human test scores in the sound recognition and discrimination tasks (*Methods*). Results are shown for 62.5-, 250-, and 1,000-ms sound durations and for different summary statistic conditions (Spec, +Mar, +MPS, +Corr, and Ori). (*B*) Spectrotemporal neural classifier performance as a function of sound statistics in the recognition and discrimination tasks for the same conditions. (*C*) Neural spectrum classifier performance. Asterisks indicate a statistically significance increase when compared to the Spec condition ($P < 0.05$, paired *t* test, corrected for multiple comparisons).

response statistics and that statistics beyond the spectra appear to influence neural coding and perception.

Relationship to Visual Textures. Texture synthesis methods were originally developed for vision (20), and both visual and auditory texture synthesis models involve similar receptive field hierarchies and similar statistics (marginals, power spectrum, correlations) (1, 20). As for audition, natural scene statistics from such models can drive the firing of visual cortex neurons in ways that likely contribute to visual perception (21, 22). Furthermore, similar to what we report here for IC, correlated firing in visual cortex can be driven directly by high-order visual scene structure (23).

The recent use of texture synthesis to study vision and audition (16, 20) is in part motivated by the fact that realistic texture stimuli can be generated with models that account for underlying transformations of both modalities and which suggest common sensory processing principles. However, there are some noticeable difference between the two modalities. Perhaps the most noticeable distinction is the fact that auditory textures overwhelmingly rely on temporal sound structure, whereas visual textures do not. Visual textures can be generated with a purely static model (no time dependency) which contains spatially segregated receptive fields and summary statistics that account for the high-order spatial statistics between image pixels. By comparison, natural sound texture synthesis involves receptive fields that operate simultaneously in time and frequency and where all of the measured summary statistics are averaged over time. While most of the relevant information in dynamic natural scenes is relatively slow (<10 Hz) (24), natural sounds contain perceptually salient temporal cues which cover several orders of magnitude (~0.5 Hz to 1.0 kHz) (25). These cues span distinct perceptual ranges of rhythm, roughness, and pitch, and auditory neurons selectively synchronize to these ranges across the auditory hierarchy (25, 26). Thus, in contrast to vision, where textures can be identified with purely statistic images that contain spatial

statistics only (27), natural sound textures require temporal structure to be meaningful (1).

As shown here, many of the informative high-order neural response statistics are inherently temporal. For instance, the spectral and temporal correlation statistics are both derived from temporal response that are synchronized to the sound envelopes up to several hundredths of hertz and thus both statistics rely heavily on temporal synchronized neural activity. Furthermore, a static sound (constant spectrum, as for Spec) that lacks temporal structure, does not form a realistic auditory impression (1) and, as shown here, cannot be easily identified as arising from a unique auditory source by humans or the neural decoders. Thus, a major advantage of auditory texture stimuli is that they have modifiable high-order structure which is critical perceptually. As shown, neural responses are sensitive to this structure and, given the highly nonlinear transformations in audition, these stimuli may lead to a more complete understanding of perceptual and neural coding strategies for such high-order sound structure when compared to more traditional auditory stimuli, such as tones or amplitude-modulated sounds.

Comparing Perception and Neural Decoding. Neural response statistics in IC ensembles reflect and are sensitive to the summary sound statistics which enabled both discrimination and recognition of the natural sound textures studied here. As shown, the pattern and strength of neural correlations is strongly influenced by the summary statistics included in the synthetic variants. Upon adding summary statistics to the synthetic texture, neural correlations converge upon and more closely resemble that of the original texture sound resulting in patterned activity that is more easily recognized by the neural classifier. Similar behavior was observed for the human listeners whereby the recognition performance improves upon adding summary statistics to the synthetic variants. By comparison, the neural spectrum does not vary substantially when adding summary statistics to the synthetic variants, despite large perceptual differences in these sounds. Such insensitivity to the added statistics mirrors the neural discrimination trends and human performance, both of which are largely insensitive to the included summary statistics.

While texture recognition appears to be linked to high-order sound and response structure, texture discrimination is much more stable and appears to depend much less on which summary statistics are included. One notable difference between neural classification and human psychoacoustical data is that the neural classifier was trained using the five original texture sounds for the recognition task, whereas humans were presented these sounds blindly without feedback. Thus, human listeners had no knowledge of the particular statistics for the five texture sounds and instead they relied on prior learned knowledge for these sound categories. Despite this, both humans and the neural classifier appear to be able to utilize summary statistics for recognition since, in both instances, adding statistics to the synthetic variants improves performance. In contrast, the neural discrimination classifier used the neural response statistics directly from the two sounds being discriminated which more closely resembles the perceptual comparison between the two sounds carried out by human listeners. Sounds could be easily discriminated, both by human observers and the neural classifier, even if high-order summary statistics are not included in the synthetic variants. This suggests that the sound spectrum, while not the sole cue used for discrimination, may be sufficient for discrimination on its own. Spectrum equalized sounds can be discriminated and recognized from neural ensemble responses (Fig. 5*D*). Moreover, these sounds are perceptually quite different and can be readily identified or distinguished from each other (*SI Appendix* and Sounds S26–S30). Thus, if spectrum cues are not available it is possible to take advantage of high-order structure in the neural activity (neural correlations) for both recognition and discrimination.

An intriguing aspect of both the neural and human findings is that the evidence integration time scales can depend critically on both the chosen statistic and the perceptual task. Texture discrimination can be accomplished relatively easily with spectrum-based cues only and, with both the human listeners and neural decoders, and the integration of spectrum cues is exceptionally fast. Although this does not exclude the possibility of high-order cues being used, it suggests that spectrum cues are sufficient for high accuracy natural texture discrimination. This is consistent with prior findings suggesting that texture excerpts (different exemplars of the same texture) can be discriminated readily with short duration sounds where spectrum cues differ and that they become increasingly difficult to discriminate for longer durations, where presumably the average spectrum and high-order statistics are similar (16). By comparison, our data suggest that recognition of natural sound textures appears to depend much more heavily on the availability of high-order summary statistics and evidence about the summary statistics needs to be accumulated over relatively long durations to be informative, both neurally and behaviorally. Such findings are consistent with prior perceptual studies where it has been shown that high-order summary statistics are necessary for creating realistic impressions of sounds (1) and that texture discrimination (from different categories) improves with increasing sound duration (16). Our results build on these findings by suggesting that low- and high-order summary statistics appear to have distinct evidence integration times and these appear to contribute differently to recognition and discrimination of sound textures.

Altogether, the findings here suggest that information from neural response statistics contribute differentially to recognition and discrimination of sound textures. Low-order summary statistics (e.g., spectrum) and the corresponding neural response statistics (e.g., neural spectrum), accumulate information quickly, allowing for fast and accurate sound discrimination. High-order sound statistics, by comparison, are reflected in coordinated activity across IC neural ensembles (neural correlations). Such activity requires longer evidence accumulation times to be useful yet it contributes substantially more toward the recognition of natural sounds.

Methods

Natural Texture Sounds and Audio Delivery. Five natural sound textures were used in this study: crackling fire, bird chorus, outdoor crowd, running water, and rattling snake sounds. These sounds were selected since they each have distinct spectral and temporal properties. These real-world textures were initially analyzed in a generative auditory model that contained hierarchical filters representing the signal processing of the cochlea and midlevel (e.g., auditory midbrain) auditory system, and the statistics of the resulting decomposition were measured and used to generate synthetic variants with reduced statistics (1). For each sound, synthetic variants were generated that included only the sound power spectrum (Spec), or which sequentially incorporated the channel marginal statistics (+Mar), modulation power spectrum (+MPS), and the correlations between frequency channels (+Corr) (Fig. 1). To control for the spectral cues in each sound, we also generated sound variants with a matched 1/f power spectrum (pink noise) (4). All sounds were delivered at 65 dB SPL (sound pressure level) in a block randomized fashion through a calibrated closed speaker audio system. Details of the texture synthesis procedures and sound delivery are outlined in *SI Appendix, Expanded Methods*.

Animal Procedures. Four female Dutch Belted rabbits (age of 0.5 to 2 y) with a weight of 1.5 to 2.5 kg were used. We measured the auditory response properties of neuron ensembles in the auditory midbrain (IC) of unanesthetized animals during passive listening to experimental sounds. All experimental procedures were approved by the University of Connecticut Animal Care and Use Committee and in accordance with NIH and the American Veterinary Medical Association guidelines.

Briefly, aMUA (4, 28) was recorded from tonotopically organized regions of IC using linear arranged multichannel silicon probes (NeuroNexus; Fig. 2 *A* and *B*). For the sound texture paradigm, we recorded from a total of 38 penetration sites in four animals and 29 sites were used in this study due to the data quality based on response stability, recording duration (a minimum of

15 trials per sound variant), and the covered frequency range (at least two octaves). For the spectrum equalized paradigm, we recorded from an additional 11 penetration sites from two animals. Details of surgical, neural recording procedures and aMUA analysis are outlined in *SI Appendix, Expanded Methods*.

Population Response Metrics: Neural Spectrum and Neural Correlations. We separately assessed the role of spectrum- and correlation-based codes and their contribution toward neural recognition and discrimination of natural sound textures. Spectrum-based codes account for the tonotopic decomposition of sounds along the auditory pathway. They can be viewed as a conventional place-rate code where the strength of activity is driven bottom up by the power in the sound at each frequency channel. Here, the neural spectrum (Fig. 2*E*) for a given sound at a specific recording location was estimated by averaging the responses from each channel across trials and time. Since neurons in IC can also potentially encode stimulus information through coordinated firing, we estimated the stimulus-driven neural correlations across the 16 tonotopic recording channels. Here we refer to the correlations between different frequency organized channels at zero lag as the spectral correlations (Fig. 2*C*). The spectral correlations account for the similarity of the neural activity across tonotopic locations, yet they do not account for the time scales of the coordinated neural activity. We thus also measured the temporal correlations (Fig. 2*D*) which account for the timing and pattern of the neural activity for each of the recording channels.

We quantified how the neural population responses change upon adding stimulus statistics to the synthetic sound variants and asked how incorporating higher-order structure in the texture sounds alters neural responses. We considered each of three-response metrics (neural spectrum, spectral correlation, or temporal correlation) and quantified how these differed from the original sound responses. First, we computed a similarity index (SI, equivalent to a correlation coefficient) for each of the response metrics. The SI quantifies the structural similarity in the population response metrics between the synthetic condition being tested and the original sound. To characterize how the strength of each of the response metric changes upon adding statistics to the synthetic textures, we also computed a strength ration (SR), defined as the ratio of the response metric power between the synthetic condition being tested and the original sound.

The analysis procedure used are identical to those described in a recent publication (4) and are outlined in additional detail in *SI Appendix, Expanded Methods*.

Single-Trial Neural Classifiers. To evaluate the extent to which the neural spectrum and correlation statistics contribute to recognition and discrimination of texture sounds, we developed a single-trial neural classifier which we applied separately in texture recognition and texture discrimination paradigms. The classifier has been described in detail previously (4) and is outlined in detail in *SI Appendix, Expanded Methods*. Briefly, we used a cross-validated naïve Bayes classifier for each task. In order to account for task-specific differences (recognition vs. discrimination) the Bayesian model priors were intentionally designed and trained to account for the information that is available to subjects when performing each task. In the case of texture recognition, the classifier was required to identify a sound from the provided neural response (e.g., neural spectrum or correlation) in a five-alternative forced choice task and the maximum a posteriori (MAP) rule was used to identify the sounds from the neural activity. For the sound discrimination paradigm, the goal was to determine whether the neural responses to the two sounds provided (at a specified sound duration) could be differentiated from one another. In this case, in addition to having identical durations, both sounds tested were selected from variants for the same statistic condition used (synthetic or original; e.g., fire vs. bird sounds both containing +MPS condition) and the model priors were selected to match these conditions. We then used the classifier to determine whether the two sounds were different from each other using the MAP rule.

Human Psychoacoustics. We carried out complementary experiments in human participants to determine how different sound statistics contribute to recognition and discrimination of texture sounds. All procedures were approved by the Institutional Review Board at the University of Connecticut. Participants were provided verbal and written descriptions of the experiment procedures and study rationale and they consented to participate in the experiments.

Two male and three female participants ages 20 to 35 were recruited for the study. Briefly, participants were asked to recognize or discriminate texture sounds of varying durations (62.5, 250, and 1,000 ms) and statistics (Spec, +Mar, +MPS, +Corr, and Ori). In the recognition task, participants listen to one of the five sounds for a given condition (same as physiology) and are required to identify the sound they hear. In the discrimination task, subjects listen to two sounds, and are required to respond as to whether the sounds are the same or different (2AFC). A detailed account of the psychoacoustic procedures are provided in *SI Appendix, Expanded Methods*.

Data Availability. All study data are included in the article and supporting information.

ACKNOWLEDGMENTS. We thank the late Dr. S. Kuwada for generous support and guidance on experimental procedures in the unanesthetized rabbits. This work was supported by the National Institute on Deafness and Other Communication Disorders of the NIH under award R01DC015138 and by a grant from the University of Connecticut. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NSF. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

1. J. H. McDermott, E. P. Simoncelli, Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron* **71**, 926–940 (2011).
2. M. N. Geffen, J. Gervain, J. F. Werker, M. O. Magnasco, Auditory perception of self-similarity in water sounds. *Front. Integr. Nuerosci.* **5**, 15 (2011).
3. R. McWalter, T. Dau, Cascaded amplitude modulations in sound texture perception. *Front. Neurosci.* **11**, 485 (2017).
4. M. Sadeghi, X. Zhai, I. H. Stevenson, M. A. Escabí, A neural ensemble correlation code for sound category identification. *PLoS Biol.* **17**, e3000449 (2019).
5. F. A. Rodríguez, C. Chen, H. L. Read, M. A. Escabí, Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. *J. Neurosci.* **30**, 15969–15980 (2010).
6. M. A. Escabí, L. M. Miller, H. L. Read, C. E. Schreiner, Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *J. Neurosci.* **23**, 11489–11504 (2003).
7. M. A. Escabi, C. E. Schreiner, Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *J. Neurosci.* **22**, 4114–4131 (2002).
8. D. L. Barbour, X. Wang, Contrast tuning in auditory cortex. *Science* **299**, 1073–1075 (2003).
9. N. C. Rabinowitz, B. D. Willmore, J. W. Schnupp, A. J. King, Contrast gain control in auditory cortex. *Neuron* **70**, 1178–1191 (2011).
10. H. Attias, C. Schreiner, Coding of naturalistic stimuli by auditory midbrain neurons. *Adv. Neural Inf. Process. Syst.* **10**, 103–109 (1998).
11. I. Nelken, Y. Rotman, O. Bar Yosef, Responses of auditory-cortex neurons to structural features of natural sounds. *Nature* **397**, 154–157 (1999).
12. S. M. Woolley, T. E. Fremouw, A. Hsu, F. E. Theunissen, Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat. Neurosci.* **8**, 1371–1379 (2005).
13. S. Shamma, D. Klein, The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *J. Acoust. Soc. Am.* **107**, 2631–2644 (2000).
14. L. A. Jeffress, A place theory of sound localization. *J. Comp. Physiol. Psychol.* **41**, 35–39 (1948).
15. C. Chen, H. L. Read, M. A. Escabí, Precise feature based time scales and frequency decorrelation lead to a sparse auditory code. *J. Neurosci.* **32**, 8454–8468 (2012).
16. J. H. McDermott, M. Schemitsch, E. P. Simoncelli, Summary statistics in auditory perception. *Nat. Neurosci.* **16**, 493–498 (2013).
17. C. E. Schreiner, G. Langner, Periodicity coding in the inferior colliculus of the cat. II. Topographical organization. *J. Neurophysiol.* **60**, 1823–1840 (1988).
18. G. Chechik *et al*., Reduction of information redundancy in the ascending auditory pathway. *Neuron* **51**, 359–368 (2006).
19. M. Braun, Auditory midbrain laminar structure appears adapted to f0 extraction: Further evidence and implications of the double critical bandwidth. *Hear. Res.* **129**, 71–82 (1999).
20. J. Portilla, E. P. Simoncelli, A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* **40**, 49–70 (2000).
21. G. Okazawa, S. Tajima, H. Komatsu, Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E351–E360 (2015).
22. C. M. Ziemba, J. Freeman, J. A. Movshon, E. P. Simoncelli, Selectivity and tolerance for visual texture in macaque V2. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E3140–E3149 (2016).
23. M. Bányai *et al*., Stimulus complexity shapes response correlations in primary visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2723–2732 (2019).
24. D. W. Dong, J. J. Atick, Statistics of natural time-varying images. *Network* **6**, 345–358 (1995).
25. P. X. Joris, C. E. Schreiner, A. Rees, Neural processing of amplitude-modulated sounds. *Physiol. Rev.* **84**, 541–577 (2004).
26. B. C. J. Moore, *An Introduction to the Psychology of Hearing* (Academic Press, 1997).
27. B. J. Balas, Texture synthesis and perception: Using computational models to study texture representations in the human visual system. *Vision Res.* **46**, 299–309 (2006).
28. J. W. Schnupp, J. A. Garcia-Lazaro, N. A. Lesica, Periodotopy in the gerbil inferior colliculus: Local clustering rather than a gradient map. *Front. Neural Circuits* **9**, 37 (2015).