Auditory abstraction from spectro-temporal features to coding auditory entities

Gal Chechik* and Israel Nelken**

*The Gonda Multidisciplinary Brain Research Center, Bar Ilan University, Ramat Gan 52900, Israel; and **The Edmond and Lily Safra Center for Brain Sciences and the Interdisciplinary Center for Neural Computation and *Department of Neurobiology, Hebrew University, Jerusalem 91904, Israel

The auditory system extracts behaviorally relevant information from acoustic stimuli. The average activity in auditory cortex is known to be sensitive to spectro-temporal patterns in sounds. However, it is not known whether the auditory cortex also processes more abstract features of sounds, which may be more behaviorally relevant than spectro-temporal patterns. Using recordings from three stations of the auditory pathway, the inferior colliculus (IC), the ventral division of the medial geniculate body (MGB) of the thalamus, and the primary auditory cortex (A1) of the cat in response to natural sounds, we compared the amount of information that spikes contained about two aspects of the stimuli: spectro-temporal patterns, and abstract entities present in the same stimuli such as a bird chirp, its echoes, and the ambient noise. IC spikes conveyed on average approximately the same amount of information about spectro-temporal patterns as they conveyed about abstract auditory entities, but A1 and the MGB neurons conveyed on average three times more information about abstract auditory entities than about spectro-temporal patterns. Thus, the majority of neurons in auditory thalamus and cortex coded well the presence of abstract entities in the sounds without containing much information about their spectro-temporal structure, suggesting that they are sensitive to abstract features in these sounds.

Sensory systems have evolved to extract relevant information from the continuous flux of sensory stimuli. The cascade of processing stations along a sensory pathway transforms the stimuli so that the higher brain areas can make behaviorally relevant decisions. In the visual system, these transformations have been shown to involve increasingly complex representations: Center-surround thalamic representations feed the simple and complex V1 neurons, eventually feeding face sensitive neurons, which generalize over a large class of behaviorally relevant stimuli (1). In the auditory system, cochlear nerve fibers at the periphery are narrowly tuned in frequency, and neurons in primary cortex have been shown to be sensitive to specific spectro-temporal (ST) patterns (2–4). However, it is unclear whether ST sensitivity fully reflects the information processing performed by cortical neurons. An alternative account of spectro-temporal sensitivity in A1 suggests that cortical neurons “inherit” their sensitivity to ST patterns from their afferent inputs. Indeed, even neurons in the inferior colliculus, the obligatory midbrain auditory station, are sensitive to ST patterns (5, 6).

If this alternative hypothesis is true, then auditory cortical neurons may actually be extracting additional, and potentially more abstract, stimulus features. To characterize ST sensitivity, auditory neurons are often analyzed by computing the average stimulus that elicits spikes, yielding the so-called ST receptive field (STRF) (7). Unfortunately, the stimulus average may be a coarse descriptor of the complex features that excite neurons. For example, the average of all stimuli that excite a visual, face-sensitive, neuron would be an elongated lump, rather than a face. For this reason, even if the average stimulus of an auditory neuron contains distinctive ST features, the neuron may be even more informative about other aspects of the sound. High-order sensitivities could allow the neurons to identify a class of complex sounds even when the ST structure in that class varies.

In this paper, we set to quantify the strength of the relationships between neural response and ST structures in a complex stimulus set. Instead of looking at averages, we took a neural decoding approach. Specifically, to test whether neurons code information about high-order features that go beyond ST patterns, we studied two different coding problems based on the same neural responses to the same stimuli: (i) coding the ST structure of the stimulus; and (ii) coding the presence of abstract auditory entities in the stimulus (Fig. 1). The first coding problem aims to infer “physical” properties of stimuli from the neural responses. The second problem focuses on inferring a more abstract notion, which generalizes across physical realizations. This inherent difference between the two coding problems can reveal the relevant processing performed by neurons: Neurons that encode well the detailed ST structure of sounds may also encode abstract auditory entities, but it is also possible for neurons to code the presence of an abstract entity without explicitly coding the ST structure.

Comparisons between cortical coding of different stimulus aspects are often hard to interpret, because the amount of information depends on the features whose coding is being tested. We address this issue by analyzing the neural activity of multiple auditory regions and using information in one region as a baseline. Specifically, we studied neural responses to the same set of stimuli in three subsequent stations in the auditory pathway: the inferior colliculus (IC), the ventral division of the medial geniculate body of the thalamus (MGB), and the primary auditory cortex (A1). This approach allowed us to learn how stimulus representation changes along the auditory processing hierarchy.

Results
All recordings were performed in halothane-anesthetized cats by using a single set of stimuli consisting of natural and modified bird vocalizations (8–10). To investigate changes in stimulus representations across processing stations, it is necessary to use stimuli that are rich enough to engage processing mechanisms at all levels of the auditory pathway. We therefore used stimuli based on natural sounds and their modifications. Specifically, the set contained three natural recordings of bird vocalizations that consisted mostly of an amplitude- and frequency-modulated pure tone, and five variants of these stimuli that were created by separating each stimulus into a small number of high-level “auditory abstractions”. These entities included the main chirp (defined by tracking the frequency and amplitude of the energy peak), the echo component of the main chirp, and the background ambient noise. We also included a variant that contained both the main and the echo. See Fig. 4 for examples of these variants for one bird chirp. A detailed description of the stimuli is provided in (8). We studied how neurons in the same three stations cooperate to code these sounds (11).

Author contributions: G.C. and I.N. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

The authors declare that they have no conflict of interest.

1To whom correspondence should be addressed. E-mail: gal.chechik@biu.ac.il.

NEUROSCIENCE

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1111242109/-/DCSupplemental.
Coding Spectro-Temporal Structures. Neurons in the auditory system are often characterized by their ST selectivity (2, 12–18), implying that they discriminate between stimuli based on their ST content. A standard tool for studying ST selectivity is spike triggered averaging of a time-frequency representation, with corrections for possible second-order correlations in the stimulus (7, 19). This family of methods is useful when the spikes are time locked to the sounds and events that are relevant to the neuron. Otherwise, averaging may actually remove the stimulus features that excite the neurons. When the stimulus set has a rich correlation structure, averaging the stimuli that precede spikes usually reveals structures that characterize the stimulus itself (regardless of the responses). At the same time, averaging ignores high-order nonlinear dependencies in the stimuli that the neural responses may be sensitive to (20). All these cautionary notes apply to the natural and modified stimuli used in this paper and, as a result, the mean stimulus preceding spikes is not sensitive to interneuron differences. Although a number of sensitive methods have been devised to overcome these problems (19, 21), using linear STRFs may have limited predictive accuracy even when correcting for the correlations in the stimuli (20, 22).

With the goal of comparing how cells code different aspects of the stimuli, we take here a different approach for quantifying stimulus–response relationships—the decoding approach. Decoding methods aim to complement the linear analysis by taking into account high-order dependencies. Instead of calculating the average of all the sounds that precede a spike, the object of study is the joint distribution of sound segments and the occurrence of a spike, considered as a binary variable. If the distribution of sound segments preceding a spike differs from the distribution of sound segments with no spiking response, the spike event can be used to infer the stimulus that just occurred. Shannon’s mutual information (MI; refs. 23 and 24) is a useful quantity that measures the strength of this statistical relation and provides an upper bound on the transmitted information of any decoder based on the spiking activity. High levels of information in this case would indicate that spikes are strongly time locked to specific ST structures in sounds, which is the underlying assumption of the reverse-correlation methods.

To study the amount of information neurons carry about the stimuli, we start by calculating the information that single spikes convey about the preceding sound segment. In this context, one can think of the experiment as a long series of concatenated short experiments, each presenting a short stimulus segment and recording the (binary) presence of a spike in the consecutive 1 ms, as is usually done in reverse-correlation experiments. Because each stimulus was presented a number of times, spike probability could be estimated directly from the joint distribution of spikes and presented sound segments. Each 1 ms of IC neuron activity conveyed 0.0659 bit of information on each segment (biased corrected; see Methods), approximately four times more information than conveyed by A1 neurons (mean of 0.0163 bit) and MGB neurons (mean of 0.0106 bit). Normalizing by spike rates, these values are 0.0183, 0.0111, and 0.00103 bit per spike in IC, MGB, and A1.

We then proceeded to measure the information that single spikes convey about the ST structures in the short sound segments that precede them. To quantify this information, we adopt a quantization approach that is widely used in machine learning and signal processing for modeling complex continuous signals such as speech or images (e.g., refs. 25–27). We used vector quantization to learn a “codebook” of typical ST patterns in the stimulus set. This codebook consisted of a set of clusters of ST segments, each cluster treated as one “code word.” At the second phase, we represented each signal as a series of discrete symbols, corresponding to code words (clusters of ST segments) from that codebook. Finally, we computed the joint distribution and the mutual information between spikes (or spike patterns) and clusters of ST segments in their discrete representation. Intuitively, the joint distribution of spikes and ST clusters can inform us about neural processing. A neuron whose activity is determined by the ST content alone (e.g., a neuron whose responses are governed by an STRF) is expected to respond similarly to all segments in the same cluster and to respond differently to segments from different cluster. As a result, the MI that it conveys about the ST cluster should be high. This estimate will generally be smaller than the information about segments, because it is based on a reduced representation of the acoustic input. Although approaches exist for obtaining a reduced but informative representation of the stimulus (e.g., ref. 28), the current approach highlights the relationships between the ST patterns and the occurrence of single spikes or spike patterns and, hence, puts the MI computation in the context of spike-triggered-averaging literature discussed above.

In practice, the ST segments were calculated by first cutting the stimuli into 50-ms overlapping segments (Fig. 2A) and then grouping the segments into 32 clusters by the similarity of their ST patterns (the number of clusters was selected based on the entropy of the distribution of segments into clusters, see Methods). For visualization purposes, we also computed the mean of each cluster of ST segments (see examples in Fig. 2B)—these are the typical ST patterns that characterize each cluster. The sequence of segments in the original stimulus was then represented as a sequence of code words corresponding to the cluster identifiers (Fig. 2C, Left). Finally, we collected the joint occurrences of a spike and a cluster at every point in time with a resolution of 1 ms (Fig. 2C, Right) and computed the MI of the resulting joint distribution (Methods). Computing the MI by averaging over all neurons from the same cluster would generally be smaller than the information related to ST structure, because it is based on a reduced representation of the acoustic input. Normalizing by spike rates, the MI was 0.0178, 0.006, and 0.0065 bit per spike in the IC, MGB, and A1. To verify that these values represent a good estimate of the available information, the information in spikes of A1 neurons was also estimated by using STRFs fitted to the responses of the individual neurons (see Methods for details). The two estimates were similar, with a slight but consistent advantage for the clustering method used here (Fig. S1), showing that it was fine enough to extract at least as much information related to ST structure as the more standard STRF approach.

Going beyond coding using single spikes, neurons could transmit more information by using binary spike patterns that extend to multiple milliseconds. To test this possibility, we computed the mutual information between stimulus patterns and the binary spike patterns that followed each sound pattern. This computational procedure is similar to the way spike-triggered averaging is extended to event-triggered averaging, where the events are specific spike patterns. For example, using binary patterns extending for 4 bins (1 ms each; Fig. 2C), the mean mutual information with the stimulus was 0.171 bit per segment (per sample) in the IC, but was more than six times lower in...
while spikes are C3o f6 PNAS Early Edition | 3 of 6

Coding Auditory Entities. The above results suggest that IC spikes provide an order of magnitude more information about the ST structure of the stimuli compared with A1 and MGB spikes. This major difference could be explained by various coding schemes. First, it is possible that A1 and MGB spikes are more weakly determined by the stimulus, for example, due to top-down modulation that makes their responses context sensitive. Alternatively, it is also possible that A1 neurons respond to features of the stimuli that are weakly correlated with the raw ST structures. In such a case, if spikes provide independent information about the clusters of ST segments, discriminating among 32 equiprobable clusters (5 bits of information) would require collecting at least 550 ms of A1 spiking activity, but only 106 ms of IC spiking activity.

MGB (mean of 0.021 bit) and A1 (mean of 0.030) (Table 1, second row). This effect was also preserved for patterns of other sizes, including 2, 8, and even 16 bins. Interestingly, we observed little redundancy across time in these spike trains. For instance, 0.171 bit for a pattern of 4 ms corresponds to 0.043 bit per ms, which is very close to the 0.047 bit obtained with 1-ms bins.

These results suggest that IC spikes can be used to infer the cluster of the presented ST patterns significantly more accurately than A1 and MGB spikes. For example, if spikes provide independent information about the clusters of ST segments, discriminating among 32 equiprobable clusters (5 bits of information) would requires collecting at least 550 ms of A1 spiking activity, but only 106 ms of IC spiking activity.

**Table 1. Summary of information levels across regions and tasks**

<table>
<thead>
<tr>
<th>Mutual information in bits</th>
<th>IC</th>
<th>MGB</th>
<th>A1</th>
<th>IC/MGB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean MI about ungrouped sound segments, using single spikes (binary patterns of 4-ms bins)</td>
<td>0.0659 (0.244)</td>
<td>0.0106 (0.055)</td>
<td>0.0163 (0.073)</td>
<td>6.21 (4.43)</td>
</tr>
<tr>
<td>Mean MI about sound segments grouped by ST similarity, using single spikes (binary patterns of 4-ms bins)</td>
<td>0.0466 (0.171)</td>
<td>0.0050 (0.021)</td>
<td>0.0081 (0.030)</td>
<td>9.4 (8.2)</td>
</tr>
<tr>
<td>Mean MI about auditory entities, using spike patterns</td>
<td>0.090</td>
<td>0.026</td>
<td>0.027</td>
<td>3.4</td>
</tr>
</tbody>
</table>

All values are measured in bits, quantifying the MI in the full joint distribution of stimuli and response. Values in parentheses were computed by using binary patterns of length 4.

Chechik and Nelken
Coding of ST patterns vs. coding auditory entities. (Fig. 4B; $P < 4 \times 10^{-4}$, two-sample $t$ test).

Fig. 4C provides another view of the difference in relative MI levels in the two coding tasks in IC and A1. For each A1 neuron, we computed the MI it conveyed in the two coding problems, normalized by the mean MI conveyed by IC neurons in the same coding problem. In A1, 32 of 45 neurons conveyed higher MI about abstract auditory entities, and the mean MI conveyed about abstract entities was more than three times as large as the MI conveyed about ST clusters (A1/IC MI ratio = 0.29 on abstract auditory entity and 0.09 on ST clusters). We also repeated this analysis on a set of 32 stimuli based on 4 bird chirps and 8 combinations of abstract entities for each chirp, leading to very similar results (18 of 19 A1 neurons with higher A1/IC MI ratio about abstract entities than about ST clusters, Fig. S3).

We further selected three neurons whose relative MI for abstract entities in Fig. 4 was high and studied their responses to the stimuli in our set. Fig. 5 shows the responses of these three neurons to five variants of one bird chirp, where the first three variants shared the same main component and, hence, also shared the most prominent ST patterns. Interestingly, the responses of these neurons clearly discriminate between the third (Main) and the second (Main + Echo) variants, although they share ST patterns. For neuron 15, the timing of the response was different: for neuron 1, the echo elicited strong responses 50–55 ms after stimulus onset ($t$ test for spike counts in stimuli 2 and 3, $df = 19$, $P = 0.004$); neuron 4 had different total spike counts ($df = 19$, $P = 0.02$). Furthermore, the spiking patterns were often only loosely locked to the stimulus, further reducing the information about ST structures.

Discussion

By comparing information levels across processing stations, we identified a significant change in stimulus representation along the ascending auditory system. Whereas information about abstract auditory entities was reduced to one-third in A1 and MGB relative to IC, the information that single spikes encoded about the ST cluster of the immediately preceding sound segment was reduced six- to ninefold. Thus, the same stimuli are encoded by using a different strategy in IC on the one hand and in MGB and A1 on the other hand.

Our analysis is based on an experimental design that used the same set of responses for the same set of stimuli, but measured two different coding quantities. We use the fact that natural stimuli are rich enough to contain various ST structures and in higher order features. Analyzing the same responses in different coding tasks has the advantage that it reduces the variability compared with conducting multiple experiments.
There are two main features of our results. First, the significant reduction in information about ST clusters carried by A1 neurons relative to subcortical stations. Second, the substantially better preservation of the information about abstract entities in A1 despite the reduction in information about ST patterns. The first of these two is a well-known phenomenon. For example, although a typical neuron in IC easily follows temporal modulation in the range of 10–100 Hz, A1 neurons do not follow well modulation rates much above 10 Hz, even in awake animals (29, 30). Similarly, frequency tuning curves in auditory cortex tend to be wider than in IC (31) and in MGB (15), and neurons in MGB follow more faithfully the amplitude modulations of natural sounds than neurons in A1 (32). Our finding that spikes of A1 neurons carry six times less information about ST features than spikes in IC corresponds directly to this reduction in coding abilities of A1 neurons.

The reduced ability of A1 spikes to encode ST patterns has been accounted at least in part to the tendency of neurons in auditory cortex of anesthetized animals to have highly phasic response patterns. However, the responses of the neurons studied here, although recorded in anesthetized cats, are often tonic (8) and have rich temporal structure (Fig. 5 and ref. 33) that varies significantly across different stimuli. We showed that this variation could be used to decode the presence of abstract entities in the stimulus much better than to decode the ST pattern preceding the spike. In consequence, although A1 neurons were still less informative than IC neurons regarding the presence of abstract entities, this reduction was substantially smaller, by a factor of 2.5, and was partially accounted by the lower firing rates of A1 neurons. As a result, the information about abstract entities per spike is only twice smaller in A1 than in IC.

The findings in this paper extend a number of previous reports. First, STRFs computed for cortical neurons differ significantly if estimated by using natural stimuli or synthetic stimuli (19, 34, 35). This finding is consistent with the view that cortical neurons respond to more complex aspects of the stimuli. Indeed, Ahrens et al. (36) demonstrated how to turn STRFs into nonlinear predictors of neural responses in a principled way, improving their ability to describe the responses of cortical neurons. Their method, however, could not be applied for the data analyzed here, because the stimulus set used here contains high correlations and is relatively small. However, the set of stimuli we used emphasized the exquisite sensitivity of cortical neurons to rather small changes in their inputs (see also ref. 11) and, therefore, allowed us to directly address the issue of encoding abstract entities. Our results are also in line with the results of Averbeck and Romanski (37, 38). They studied the responses of neurons in ventrolateral prefrontal cortex of macaques to vocalizations. The observed responses were accounted less well by ST patterns than by a time-resolved estimate of the probabilities of the sound to belong to specific call class. They concluded that the inputs to these neurons are already encoding abstract features of the sounds. The responses studied here could supply the more abstract representation hypothesized by Averbeck and Romanski.

What are the possible mechanisms underlying this transformation in the representation of sounds? Shamma and coworkers (35, 39) suggested that some of these effects could be explained by depression of the thalamo-cortical synapse. Our results suggest that this process may actually occur earlier, because MGB neurons are also less informative about ST receptive fields but keep substantial information about auditory entities. Thus, the relevant depressing synapse may actually be at the collicular inputs to the thalamus. Additional thalamic mechanisms may be in play, like the strong interactions between the MGB and the thalamic reticular nucleus (40).

In the visual system, it is widely accepted that a processing layer combines inputs from earlier layers and extracts more complex features, which may be also invariant to some visual features such as the precise stimulus position. In the auditory system, creating an abstract representation that goes beyond ST structure, such as the one described here, may be essential for recognizing auditory events and process complex sounds like speech.

**Methods**

**Electrophysiological Recordings.** Extracellular recordings were made in A1 of nine halothane-anesthetized cats, in MGB of two halothane-anesthetized cats, and IC of nine isoflurane-anesthetized and two halothane-anesthetized cats. All experiments used standard protocols authorized by the Committee for Animal Care and Ethics of the Hebrew University (A1, MGB, and IC recordings) and the Johns Hopkins University (IC recordings). For additional details, see SI Methods and ref. 8.

**Stimulus Set.** The sound stimuli are based on natural bird chirp stimuli from field recordings (Cornell Laboratory of Ornithology), have lengths 80–120 ms and are described in details in ref. 8. The natural stimuli were separated to the main chirp component, the echo, and the background noise as described in ref. 9.

**Information About Auditory Entities.** The MI about auditory entities was estimated by using several methods and several statistics of the responses. The MI between spike counts and stimuli was estimated by using the histograms of the counts distribution per each stimulus. The bins of the histogram were chosen to maximize the bias corrected MI conveyed by each unit using the method of ref. 41, and as in ref. 33. The bias was corrected by using the first-order Taylor expansion by following refs. 41 and 42), and its maximal magnitude did not exceed 20% of the corrected MI. Latency MI was similarly computed by using a histogram estimation of latency distribution per stimulus. In addition, MI was estimated by using the distribution of binary words (43, 44). To this end, each spike train was discretized in several temporal resolutions (1, 2, 4, 8, 16, and 32 ms, yielding 3–120 bins per stimulus), and the resolution and temporal window that yielded maximal (bias corrected) MI were selected.

**Information About ST Segments.** The main idea behind the analysis is that grouping stimulus segments based on ST similarity can be used to uncover the stimulus aspects that a neuron is sensitive to. A neuron that is sensitive to an ST pattern is likely to respond in a similar way to all segments in such clusters, but differently to segments from another cluster of segments that do not contain the characteristic. As a result, it discriminates well between clusters and conveys high MI with respect to the clustered segments. At the same time, a neuron that responds to other aspects of the stimuli (not determined solely by the ST patterns) may yield similar responses to stimulus segments in different classes, leading to poor discrimination and low MI.

Specifically, to estimate MI in single spikes, the acoustic signal was cut into 50-ms segments (with 49-ms overlap), yielding a total of 1,215 segments from all 15 stimuli. The spectrogram of each segment was computed by using a Hanning window of length 5.6 ms and 2.8 ms overlap. Because neuronal responses often combine energy levels of different ST “bins” in a sublinear
way, each point in the time-frequency representation of each segment was raised to a power α chosen to maximize the information conveyed by the A1 neurons, yielding α = 0.5. Spectrograms were clustered into 32 clusters by using k-means (45), an iterative procedure that aggregates segments into homogeneous groups, such that segments are as similar as possible to the mean of their cluster. Even when using a larger number of clusters, the entropy of the distribution of spectrograms across clusters was ~5.1 bits, suggesting that 32 clusters capture the variability in the data. See additional details in SI Methods.

Information about ST segments was measured based on the distribution of segments samples. A sample is collected at every millisecond, hence the MI is reported in bits per ms and corresponds to the information conveyed by 1 ms of the response. In contrast, the information about entities is measured by repeated samples of responses to a stimulus presentation (a “trial”) and is reported as bits per trial.

To validate the clustering approach, we also estimated the MI between spikes and preceding acoustic segments by using an alternative method, based on modeling the STRF of each neuron. The information values were essentially the same as those that would result from state-of-the-art STRF estimation methods (Fig. 5).

ACKNOWLEDGMENTS. We thank E.D. Young for collecting a portion of the IC data. G.C. was supported by Israeli Science Foundation Grant 08/1001 and Marie Curie Grant PIRG06-GA-2009-256566. I.N. was supported by the Israeli Science Foundation, the US-Israeli Binational Science Foundation, and by the Gatsby Charitable Foundation.