nature neuroscience

Summary statistics in auditory perception

Josh H McDermott[1], Michael Schemitsch[2] & Eero P Simoncelli[2]

Sensory signals are transduced at high resolution, but their structure must be stored in a more compact format. Here we provide evidence that the auditory system summarizes the temporal details of sounds using time-averaged statistics. We measured discrimination of 'sound textures' that were characterized by particular statistical properties, as normally result from the superposition of many acoustic features in auditory scenes. When listeners discriminated examples of different textures, performance improved with excerpt duration. In contrast, when listeners discriminated different examples of the same texture, performance declined with duration, a paradoxical result given that the information available for discrimination grows with duration. These results indicate that once these sounds are of moderate length, the brain's representation is limited to time-averaged statistics, which, for different examples of the same texture, converge to the same values with increasing duration. Such statistical representations produce good categorical discrimination, but limit the ability to discern temporal detail.

The sounds that enter our ears are transduced with a temporal resolution that is notably fine-grained. Listeners can detect gaps in continuous sounds that are as short as a few milliseconds[1] and can localize sounds using time differences between the two ears as brief as a few hundredths of a millisecond[2]. However, the information rate implied by this temporal resolution is prohibitive for storage—the representations retained for short- or long-term memory must somehow be compressed. Moreover, although fine-grained measurements of the sound waveform capture temporal detail, they do not make explicit the similarities and differences between categories that are important for behavior. The auditory system must derive more compact and abstract representations for memory and recognition.

Sound textures provide an attractive means to study this process of abstraction. Textures are formed from the superposition of many similar acoustic events, collectively giving rise to aggregate statistical properties[3–7]. These properties can in turn be captured by statistical measurements: time averages of short-term acoustic characteristics, which summarize the qualities of a sound[8,9]. Such time averages might be measured by the auditory system following peripheral filtering operations (**Fig. 1a**) and could provide a compact representation for encoding sound. Previously, we found that statistical measurements could be used to synthesize realistic textures: sounds generated to match the statistics of real-world texture recordings (rain, fire, wind, insect swarms, etc.) often sounded like new examples of the original recording[8,9] (**Fig. 1b**). The success of this synthesis method suggests that statistics can capture perceptually important information and could, in principle, underlie texture recognition. However, it remains to be seen whether statistical representations are actually used by the auditory system and under what circumstances.

We asked whether the human auditory system summarizes sounds with time-averaged statistics, whether these statistics are retained in lieu of the fine-grained temporal details from which they are computed and whether any such dependence on summary statistics is specific to textures. To explore these questions, we assessed the ability of listeners to discriminate texture examples. If textures are represented exclusively with time-averaged statistics, signals with distinct statistical properties should be discriminable, whereas those with similar statistics should be difficult to distinguish even when their temporal details are completely different.

We manipulated texture statistics by first generating multiple examples of textures with particular long-term statistics. We then extracted excerpts of different durations from these textures, exploiting the natural variability that arises in statistics that are computed from small samples. Specifically, the statistics computed from brief excerpts of a particular texture vary considerably from one excerpt to the next (as they are biased by the particular features that happen to occur in each excerpt), but converge to their 'true' values as the duration increases (**Fig. 1c**). This convergence must occur for any statistic that the brain might measure, assuming that the statistic is an average over time and that the excerpts are drawn from a stationary signal such as a texture.

We designed two tasks in which the discrimination of these texture excerpts was measured as a function of their duration. In the first task, listeners discriminated texture excerpts with different long-term statistics. If perception relies on these statistics, performance should improve with excerpt duration as the measured statistics converge to their long-term values. In the second task, listeners discriminated texture excerpts with the same long-term statistics. If perception is again determined by summary statistics, performance in this case should decline with duration despite the concomitant increase in stimulus information, because the statistics converge to similar values.

RESULTS

We first conducted experiments using synthetic textures whose statistics matched those of natural sound textures. The statistics were measured from a model of the auditory periphery (**Fig. 1a**) and

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Howard Hughes Medical Institute, Center for Neural Science, and Courant Institute of Mathematical Sciences, New York University, New York, New York, USA. Correspondence should be addressed to J.H.M. (jhm@mit.edu).

Received 24 September 2012; accepted 31 January 2013; published online 24 February 2013; doi:10.1038/nn.3347

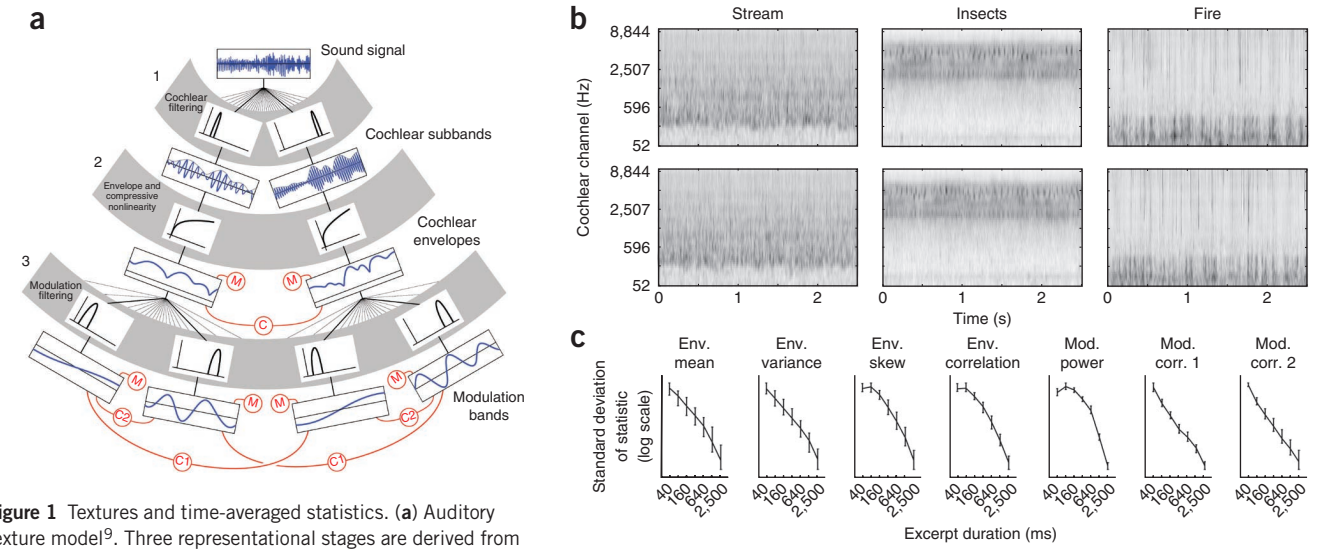

Figure 1 Textures and time-averaged statistics. (**a**) Auditory texture model[9]. Three representational stages are derived from the sound waveform, mimicking peripheral auditory processing from the cochlea through roughly the thalamus: (1) frequency subbands are generated as responses of cochlear bandpass filters, (2) their envelopes are extracted and passed through a compressive nonlinearity, and (3) modulation bands are generated by filtering the compressed cochlear envelopes. Summary statistics are measured from the envelopes and modulation bands and are proposed to underlie the perception of texture. Synthetic textures are generated by adapting samples of Gaussian noise to cause them to have the same statistics as a real-world sound recording. Model diagram is modified from ref. 9. (**b**) Spectrograms of two 2.5-s exemplars of each of three different textures. The top and bottom exemplars of each pair have the same long-term (5 s) statistics, but differ in temporal detail. (**c**) Variability of texture statistics. Graphs show the median s.d. of each of seven classes of statistic (computed for the textures used in our experiments by measuring the statistics in multiple excerpts of each texture). The measured statistics are variable for short excerpts, but converge to fixed values as the excerpt length is increased. Error bars show lower and upper quartiles; s.d. is plotted on a logarithmic scale. The slight nonmonotonicity evident in some of the plots is a result of the onset and offset of the excerpt, which can influence the statistics when the excerpts are short.

synthetic textures were generated by adjusting a 5-s sample of random noise until it attained the same values of these statistics[9].

These sounds offer two notable benefits as experimental stimuli. First, synthetic textures generated in this way share many statistical and perceptual properties of natural sounds, but allow a degree of control that is difficult to attain with real-world recordings. In particular, synthetic textures lack the occasional extraneous sounds that are often present in recordings of naturally occurring sound textures. Second, the synthesis procedure can generate multiple distinct exemplars with nearly identical summary statistics, simply by initializing the procedure with different samples of noise. Visual inspection reveals that the exemplars in each pair (**Fig. 1b**) differ in their acoustic details, but have the same global texture characteristics. Our synthesis method thus provides a way to generate examples of stationary signals that nonetheless have meaningful, naturalistic structure. To confirm that our results would also hold for real-world sounds and to explore the relation of statistical representations to texture, we also experimented with textures generated by combining multiple recordings of individual real-world sound sources (experiments 4 and 5).

Experiment 1: texture discrimination

We began by measuring the ability of listeners to discriminate textures with different long-term statistics. On each trial, listeners heard three sounds of equal duration: two different exemplars of one texture (excerpts of synthetic signals with the same long-term statistics) and one exemplar of another (**Fig. 2a**). The excerpt with distinct long-term statistics could occur first or last in the sequence, and listeners were asked to identify whether the first or last sound was generated by a source that was distinct from that of the other two.

Performance in this task was relatively poor for brief texture excerpts, but improved gradually with duration ($F_{5,55}$ = 50.96, $P < 0.001$; **Fig. 2b**). Improvement in performance with duration is arguably expected, as the information available for discrimination grows with duration, but is nevertheless consistent with the hypothesis that perception is based on time-averaged statistics. Over short durations, these statistics are variable, such that they are different for all three sounds in a trial, rendering discrimination difficult. However, as the excerpts lengthen, the statistics converge to values that are similar for the two excerpts of the same texture and different for the excerpt of the alternative texture, enabling the listener to identify which of the three sounds is distinct.

Experiment 2: exemplar discrimination

We next measured the ability of listeners to discriminate different texture exemplars with the same long-term statistics. We again presented three sounds on each trial. Two of them were physically identical (the same excerpt of a particular texture repeated), and the third was a different excerpt of the same texture (which differed in temporal detail from the other two sounds) (**Fig. 2a**). Because all three stimuli were samples of the same texture, they had the same long-term statistics.

In this experiment, performance was high for brief excerpts and paradoxically declined with duration ($F_{5,55}$ = 40.66, $P < 0.001$; **Fig. 2b**). Although listeners received more information in the longer duration texture excerpts, they could not use it to discriminate sounds; indeed, the added information actually impaired performance. This result is consistent with the hypothesis that discrimination is based on time-averaged statistics, which, for the three sounds presented in a trial, converged to the same values as the excerpt duration increased (**Fig. 1c**).

The high performance at short durations indicates that the two texture excerpts in a trial differed in their details to a degree that was readily detectable if the excerpts were short. Moreover, the fact that performance continuously increased with duration in experiment 1 suggests that details over the full stimulus time course contribute to the discrimination of sounds with distinct statistics and that listeners

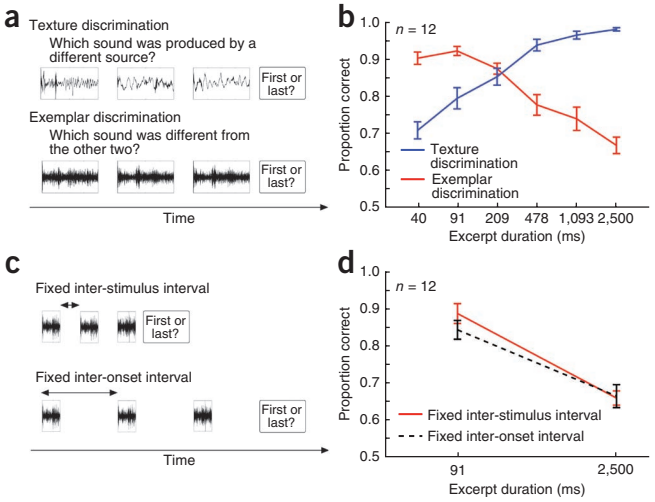

Figure 2 Texture and exemplar discrimination results. (**a**) Schematic of trial structure for experiments 1 and 2. Three sounds were played in succession, separated by a fixed interval. In the texture-discrimination task, two of the sounds were distinct excerpts of the same texture and the third (presented first or last) was an excerpt of a different texture. In the exemplar-discrimination task, two of the sounds were physically identical excerpts of a texture and the third was a different excerpt from the same texture. (**b**) Results of experiments 1 and 2. (**c**) Schematic trial structure for experiment 3. One set of conditions fixed the ISI, as in experiments 1 and 2. Another set of conditions fixed the inter-onset interval, such that shorter excerpts were separated by longer intervals than longer excerpts. Short excerpts are depicted for comparison to the long excerpts shown in **a** for fixed ISI, and are not drawn to scale. The actual ISI for the short duration conditions was 400 ms in the fixed-ISI configuration and 2,809 ms in the fixed inter-onset interval configuration. (**d**) Results of experiment 3. Error bars represent s.e.

can flawlessly discriminate such sounds even when their duration is long. Taken together, these findings indicate that temporal detail is used to construct a statistical summary of a texture, but is otherwise inaccessible to the listener once the texture is of moderate length.

Experiment 3: control for time delay

An alternative explanation for the poor performance at long durations in experiment 2 is that temporal detail simply cannot be retained over such durations. In experiments 1 and 2, the interstimulus interval (ISI) was fixed across conditions. As a result, the elapsed time between the temporal details that listeners were asked to compare was greater for the long durations than for the short (**Fig. 2a,c**). To test whether elapsed time alone could explain the poor performance at long durations, we conducted a second version of the task in which the inter-onset interval was fixed across conditions such that a long silent period separated the short duration excerpts (**Fig. 2c**). For comparison, the same listeners also completed the original exemplar discrimination task of experiment 2.

We found that discrimination was largely unaffected by time delay *per se* (**Fig. 2d**). Performance with short excerpts was similar for the two task variants (t_{11} = 1.75, P = 0.11) and, in both cases, listeners were markedly better at discriminating short excerpts than long excerpts ($F_{1,11}$ = 34.29, P < 0.001; no interaction between duration and delay, $F_{1,11}$ = 2.96, P = 0.11). Elapsed time alone therefore does not have a large effect on memory for temporal detail, at least not for the sounds and timescales that we explored (also see ref. 10). Instead, the ability to distinguish sounds of moderate or longer duration appears to be constrained by their statistics.

Experiments 4 and 5: single sources versus textures

How general is the perceptual dependence on statistics? In real-world conditions, textures arise from the superposition of multiple sources (raindrops, animal calls, handclaps, etc.). It is natural to ask whether the effects seen in experiments 2 and 3 can be obtained for textures generated by superimposing sources and whether similar effects would occur for individual sound sources in addition to textures. To address these questions, we measured discrimination of superpositions of different numbers of individual real-world sound sources.

We first generated 'cocktail party' textures by superimposing recordings of different speakers (**Fig. 3a**). To avoid lexical or linguistic confounds, we used German speech (and selected subjects who did not understand German). The task was the same as in experiment 2: listeners were presented with two excerpts taken from the same signal (different clips of the same group of people talking), one of which was presented twice, and had to judge which of the three sounds was different from the other two. The signals from which the excerpts were drawn consisted of 1, 7, 29 or 115 different speakers.

In this experiment, short duration exemplars were highly discriminable for all signal types (**Fig. 3b**). In contrast, performance for long durations varied considerably across conditions, producing an interaction between duration and the number of sources ($F_{3,33}$ = 16.30, P < 0.001). For a single speaker, performance improved with duration (t_{11} = 2.79, P = 0.018), as one would expect given that the longer excerpts provide more information to support discrimination. For mixtures of speakers, however, performance declined with duration, with a larger decrement for mixtures of more speakers.

We conducted an analogous experiment with random sequences of drum hits that varied from sparse to dense (**Fig. 3c**; also see refs. 11,12). Performance was again good at short durations in all conditions, but exhibited an interaction between the effect of duration and density ($F_{2,22}$ = 16.04, P < 0.001; **Fig. 3d**); performance improved with duration in the sparse condition (t_{11} = 2.29, P = 0.04) and decreased with duration in the denser conditions.

In both experiments, one might expect that adding together large numbers of sources would produce a completely uniform sound, rendering discrimination difficult for uninteresting reasons. However, the high levels of performance at short durations suggest that this is not the case; the signals contained readily detectable variation irrespective of their density. For longer excerpts, this variation was inaccessible when the signals were composed of a large number of sound sources, but it remained available for individual sources.

One might imagine that these results could be consistent with discrimination based on summary statistics, as the statistics of single sources, or sparse mixtures thereof, might plausibly remain highly variable even for longer excerpts. To explore this issue, we measured the variability of our model's texture statistics for the speech and drum stimuli used in our experiments. The texture statistics were indeed more variable for single sources than for dense textures, which, at first glance, seems to mirror the behavioral results (**Fig. 3e**). However, note that the variability of the statistics of single sources nonetheless decreased with duration, and, in particular, was lower at long durations than that of the densest stimuli at short durations. In contrast, discrimination was better for single sources at long durations than for any of the short-duration conditions. This discrepancy between the statistic variability and discrimination suggests that something other than, or in addition to, time-averaged statistics underlies the perception of single sound sources (see Discussion).

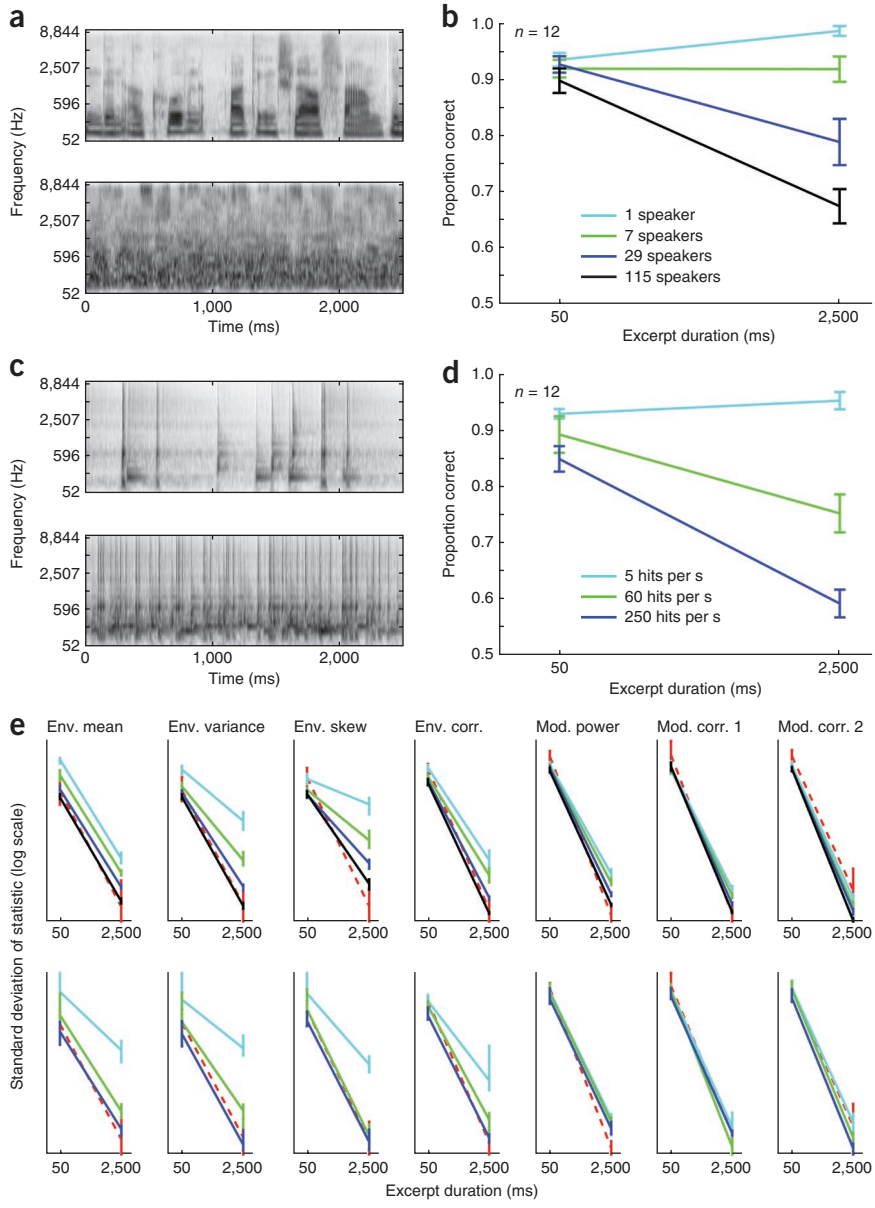

Figure 3 Exemplar discrimination with mixtures of sources. (**a**) Spectrograms of a single speaker (top) and a mixture of 29 speakers (bottom). (**b**) Results of experiment 4. (**c**) Spectrograms of a sparse sequence of drum hits (5 hits per s, top) and a denser sequence (60 hits per s, bottom). (**d**) Results of experiment 5. (**e**) Statistic variability for mixtures of sources: speech (top) and drum (bottom) stimuli. The red dashed line represents the statistic variability for the synthetic textures used in experiments 1–3. Other lines are colored as in **b** and **d**. Data are presented as in **Figure 1c**. Error bars represent s.e.

DISCUSSION

Our results suggest that the brain summarizes collections of acoustic features with their time-averaged statistics. As sounds become more texture-like and of longer duration, listeners lose access to detailed temporal structure and become dependent on these summary statistics. As a result, excerpts of dense mixtures of sources (that is, textures) are difficult to discriminate once they are of moderate length, even though the individual sources, as well as short excerpts of the same textures, are readily discriminable. The real-world implication of our results is that for most naturally occurring sound textures (rain, fire, wind, etc.), which typically have durations well in excess of our experimental stimuli, the brain's representation is likely to consist primarily of time-averaged statistics.

Our results provide an explanation for previous findings that discrimination of white noise samples is best for relatively short duration samples (less than 100 ms)[11,13–16]. In most cases, however, performance on psychophysical tasks improves with stimulus duration, be it absolute detection[17], frequency discrimination[18], modulation detection[19,20] or localization[21,22]. Discrimination of textures with distinct statistical properties also improved with duration (experiment 1), as did discrimination of the temporal detail of single sources (experiments 4 and 5). The decline in performance with duration (experiment 2) seems to be limited to cases in which discrimination depends on the retention of fine-grained temporal detail in dense signals.

Two modes of auditory representation

Why does the effect of duration interact with that of density? We believe our results collectively suggest a framework for the organization of auditory processing, built on two modes of representation. In brief, we suppose the following: (i) Incoming sound signals are encoded as sequences of features (perhaps including phoneme-like elements in speech, notes in music, or onsets in environmental sounds). Their sequential order makes temporal structure explicit. (ii) Concurrent with this sequential encoding, temporal detail is summarized with statistics, computed as running averages over time. (iii) The sequential feature encodings used by the auditory system are typically sparse for individual natural sound sources[23–25], perhaps because the encoding process is designed to recover sparse generative processes in the world when they exist. As a consequence, individual sound sources, or superpositions of small numbers of sources, are encoded with small numbers of features. Typical textures, however, require larger numbers of features per unit time. (iv) Memory capacity places limits on the number of features that can be retained (for instance, for comparison with a subsequent stimulus), rather than on the duration of the sound to be retained. It is therefore possible to remember a longer segment of an individual source than of a texture. (v) The signal that enters the ears is continuously and obligatorily encoded, and once the feature memory capacity is reached, its contents are overwritten by subsequent input. At that point, previous temporal detail is available only via its contribution to summary statistics.

Obligatory encoding of incoming sound information into memory could explain why the discrimination of texture excerpts worsens with increasing duration, rather than saturating (as might be expected if a memory buffer reached capacity): listeners cannot simply remember the initial part of a long texture excerpt, as its representation is replaced by temporal details from the later portions of the excerpt. Nor can they remember the end of a long texture excerpt, as its representation is replaced by the details from initial portions of the

subsequent excerpt in a trial. In support of this obligatory encoding hypothesis, we found that exemplar discrimination remained poor when a brief initial or final segment of a long texture excerpt was separated from the rest of the excerpt with a silent gap (**Supplementary Fig. 1**). Subsequent detail appears to overwrite previous detail even when listeners are explicitly cued to remember a particular subsegment of a sound.

Similar auditory interference effects are well documented in other contexts (for example, see refs. 26,27). However, our findings indicate that although this interference can prevent direct conscious access to temporal detail, it does not eliminate its contribution to a statistical summary. Taken together, our results are suggestive of two modes of auditory representation: one in which temporal detail is made explicit, but whose memory capacity is limited, and one in which temporal detail is continuously incorporated into summary statistics, but not otherwise retained.

In both cases, the sound waveform is converted into a more abstract and compact representation suitable for recognition and retention. Although the temporal details of textures are left inaccessible by this arrangement once they are of moderate duration, for most real-world tasks that listeners must perform on textures (for example, distinguishing wind from fire or judging the intensity of rain), the summary statistics are more informative than the precise arrangement of features, and this strategy may thus be both economical and adaptive. Moreover, the two types of representation need not be mutually exclusive. Even in signals that produce sparse temporal encodings, such as speech or music, summary statistics may have a complementary role in representing properties that are time invariant, such as voice qualities and speaker identity[28] or aspects of musical genre and style[29]. Dense textures, such as those that we explored in this study, may be unique only in allowing us to isolate the role of time-averaged statistics by removing direct access to temporal detail.

It remains unclear whether summary statistics are also involved in representing brief sounds. The variability in the statistics of short texture excerpts (**Fig. 1c**) is consistent with the performance levels that we observed at short durations (high for exemplar discrimination, low for texture discrimination) under the hypothesis that discrimination is based on the excerpt statistics. However, at short durations, the summary statistic hypothesis becomes less distinct from alternative possibilities, in part because there is less to summarize. The set of sounds with the same statistics shrinks as the duration decreases, and the distinction between a statistical representation and one that explicitly encodes temporal detail therefore becomes blurred, as the statistics inevitably capture aspects of temporal detail that are lost through the averaging process at longer durations.

Related findings

One exception to our findings may occur when a particular texture excerpt is presented repeatedly (that is, more than the two presentations that occurred in the trials of our experiments). Individual features of repeated segments of noise, for instance, can be detected and learned[30–34]. In these cases, repetition seems to allow detail to be retained that would otherwise be discarded. Repetition may provide a cue that a sound is not actually a texture and could promote the learning of features for its temporal description, allowing a representation that can be retained in memory.

Statistical representations may also be evident in auditory perceptual completion. Tones that are modulated in amplitude or frequency are heard as continuous when interrupted with a brief burst of intense noise[35,36]. However, continuity is perceived even if the modulations on either side of the noise are 180 degrees out of phase, such that a continuous modulation is physically inconsistent with the stimulus. Apparently, the auditory system represents the modulation in a way that discards the phase while retaining the rate, perhaps using summary statistics similar to those we have implicated in texture representation. We found informally that textures interrupted by intense noise bursts are also heard as continuous, despite the absence of individual features that might be tracked or completed across the noise, suggesting that a statistical representation is used to fill in the masked portion of the texture.

Although the mechanisms seem likely to be different, our finding that details are discarded in lieu of a more abstract representation is reminiscent of the well-known phenomenon of categorical perception, in which stimulus variation that does not straddle a learned category boundary is reduced in salience[37]. The need to encode speech categorically can also apparently influence access to fine-grained time differences between the two ears[38]. All these instances reflect the transformation of the sound waveform to a more abstract and compact representation.

Our results have interesting parallels in the visual system, which has been hypothesized to represent summary measures of sets of image features[39–41]. The averaging that has been characterized in vision occurs over spatial regions of the visual field rather than time and is most evident in the periphery, where the dependence on ensemble statistics has been related to the phenomenon of crowding[42–45]. Statistical measures also seem to dominate visual perception in the absence of attention[46]. In contrast, our auditory effects involved pooling information over time, and the dependence on statistical measures seems to be related to memory capacity limitations. It is possible that analogous effects would occur for visual stimuli that vary over time.

Potential mechanisms

Our conclusions are based, in part, on experiments with textures synthesized from statistics measured in real-world texture recordings. Because the textures are defined by multiple statistics, it is difficult to know which statistics might have been used by listeners as they performed our tasks. We examined whether trial-to-trial variation in the difference between the values of a particular statistic for the two excerpts could predict performance in experiment 2. We found that all of the statistics used in our texture model exhibited correlations with performance when analyzed in this way: the difference in their values for the two excerpts was, on average, larger on trials that produced correct responses. However, the differences in statistics between two excerpts are themselves typically correlated across statistics, making it difficult to establish a causal role for any particular statistic. Moreover, the convergence phenomenon on which our experiments relied (**Fig. 1c**) is not specific to the statistics in our model, and must hold for any time-averaged statistic. We therefore do not claim that the brain computes statistics that are identical to those in our model, only that the brain's perceptual representations of textures consist of temporal summaries, of which our model statistics provide examples.

Our experimental results also do not constrain the details of the temporal averaging process by which statistics are derived. In particular, averages are presumably computed over an integration window of some duration[47,48] (possibly different for different statistics). It is notable that, for even the longest excerpts that we used, exemplar discrimination was not fully at chance. One explanation is that listeners do not exclusively average over the full signal length, but instead retain some degree of access to information averaged over shorter timescales, at which the statistics vary enough to support above-chance discrimination. A sequence of such statistics could form a coarse representation of temporal structure, raising the possibility that summary statistics and the feature encodings discussed above could lie on the ends of a continuum of representations of varying temporal specificity.

We have argued that information from the different sound events within a texture is averaged over time to form a summary and that the individual events become inaccessible once the sound is of moderate length. However, it is clear that if an event is sufficiently incongruous with a texture, it may be heard and remembered as a distinct event. This phenomenon was apparent in pilot versions of experiment 2 that we conducted with recordings of real-world textures. We observed informally that different excerpts of a recording were often discriminable by virtue of particular distinctive acoustic events (for example, a bird call that was faintly audible in one excerpt of a recording of a stream). The audibility of such discrepant events suggests that the averaging process that generates texture representations occurs somewhat selectively[49], perhaps in conjunction with a clustering or grouping process, as has been proposed to underlie auditory stream segregation[50]. Coordination of integration with segregation could allow for efficient estimates of texture properties along with temporally precise descriptions of foreground sounds in auditory scenes.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

The authors thank B. Anderson, S. Keshvari and J. Traer for comments on earlier versions of the manuscript. Research was funded by the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

J.H.M., M.S. and E.P.S. designed the experiments. M.S. conducted the experiments. J.H.M. analyzed the data. J.H.M. and E.P.S. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Plomp, R. Rate of decay of auditory sensation. *J. Acoust. Soc. Am.* **36**, 277–282 (1964).
2. Dye, R.H. & Hafter, E.R. The effect of intensity on the detection of interaural differences of time in high-frequency trains of clicks. *J. Acoust. Soc. Am.* **75**, 1593–1598 (1984).
3. Saint-Arnaud, N. & Popat, K. Analysis and synthesis of sound texture. *Proc. AJCAI Workshop Comput. Auditory Scene Anal.* 293–308 (1995).
4. Dubnov, S., Bar-Joseph, Z., El-Yaniv, R., Lischinski, D. & Werman, M. Synthesizing sound textures through wavelet tree learning. *IEEE Comput. Graph. Appl.* **22**, 38–48 (2002).
5. Athineos, M & Ellis, D. Sound texture modeling with linear prediction in both time and frequency domains. *IEEE Workshop Appl. Signal Processing Audio Acoustics* 648–651 (2003).
6. Lu, L., Wenyin, L. & Zhang, H. Audio textures: theory and applications. *IEEE Trans. Speech Audio Process.* **12**, 156–167 (2004).
7. Schwarz, D. State of the art in sound texture synthesis. *14th Int. Conf. Digital Audio Effects* 221–231 (2011).
8. McDermott, J.H., Oxenham, A.J. & Simoncelli, E.P. Sound texture synthesis via filter statistics. *IEEE Workshop Appl. Signal Processing Audio Acoustics* 297–300 (2009).
9. McDermott, J.H. & Simoncelli, E.P. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron* **71**, 926–940 (2011).
10. Demany, L., Trost, W., Serman, M. & Semal, C. Auditory change detection: simple sounds are not memorized better than complex sounds. *Psychol. Sci.* **19**, 85–91 (2008).
11. Goossens, T., van de Par, S. & Kohlrausch, A. On the ability to discriminate Gaussian-noise tokens or random tone-burst complexes. *J. Acoust. Soc. Am.* **124**, 2251–2262 (2008).
12. Geffen, M.N., Gervain, J., Werker, J.F. & Magnasco, M.O. Auditory perception of self-similarity in water sounds. *Front. Integr. Neurosci.* **5**, 15 (2011).
13. Hanna, T.E. Discrimination of reproducible noise as a function of bandwidth and duration. *Percept. Psychophys.* **36**, 409–416 (1984).
14. Coble, S.F. & Robinson, D.E. Discriminability of bursts of reproducible noise. *J. Acoust. Soc. Am.* **92**, 2630–2635 (1992).
15. Heller, L.M. & Trahiotis, C. The discrimination of samples of noise in monotic, diotic, and dichotic conditions. *J. Acoust. Soc. Am.* **97**, 3775–3781 (1995).
16. Goossens, T., van de Par, S. & Kohlrausch, A. Gaussian-noise discrimination and its relation to auditory object formation. *J. Acoust. Soc. Am.* **125**, 3882–3893 (2009).
17. Gerken, G.M., Bhat, V.K.H. & Hutchinson-Clutter, M.H. Auditory temporal integration and the power-function model. *J. Acoust. Soc. Am.* **88**, 767–778 (1990).
18. Moore, B.C.J. Frequency difference limens for short-duration tones. *J. Acoust. Soc. Am.* **54**, 610–619 (1973).
19. Viemeister, N.F. Temporal modulation transfer functions based upon modulation thresholds. *J. Acoust. Soc. Am.* **66**, 1364–1380 (1979).
20. Sheft, S. & Yost, W.A. Temporal integration in amplitude modulation detection. *J. Acoust. Soc. Am.* **88**, 796–805 (1990).
21. Houtgast, T. & Plomp, R. Lateralization threshold of a signal in noise. *J. Acoust. Soc. Am.* **44**, 807–812 (1968).
22. Hafter, E.R., Dye, R.H. & Gilkey, R.H. Lateralization of tonal signals which have neither onsets nor offsets. *J. Acoust. Soc. Am.* **65**, 471–477 (1979).
23. Klein, D.J., Konig, P. & Kording, K.P. Sparse spectrotemporal coding of sounds. *EURASIP J. Appl. Signal Process.* **7**, 659–667 (2003).
24. Smith, E.C. & Lewicki, M.S. Efficient auditory encoding. *Nature* **439**, 978–982 (2006).
25. Hromadka, T., DeWeese, M.R. & Zador, A.M. Sparse representations of sounds in the unanesthetized auditory cortex. *PLoS Biol.* **6**, 124–137 (2008).
26. Deutsch, D. Tones and numbers: Specificity of interference in short-term memory. *Science* **168**, 1604–1605 (1970).
27. Starr, G.E. & Pitt, M.A. Interference effects in short-term memory for timbre. *J. Acoust. Soc. Am.* **102**, 486–494 (1997).
28. Latinus, M. & Belin, P. Human voice perception. *Curr. Biol.* **21**, R143–R145 (2011).
29. Tzanetakis, G. & Cook, P. Musical genre classification of audio signals. *IEE Trans. Speech Audio Processing* **10**, 293–302 (2002).
30. Guttman, N. & Julesz, B. Lower limits of auditory periodicity analysis. *J. Acoust. Soc. Am.* **35**, 610 (1963).
31. Warren, R.M., Bashford, J.A. Jr., Cooley, J.M. & Brubaker, B.S. Detection of acoustic repetition for very long stochastic patterns. *Percept. Psychophys.* **63**, 175–182 (2001).
32. Kaernbach, C. The memory of noise. *Exp. Psychol.* **51**, 240–248 (2004).
33. Agus, T.R., Thorpe, S.J. & Pressnitzer, D. Rapid formation of auditory memories: insights from noise. *Neuron* **66**, 610–618 (2010).
34. McDermott, J.H., Wrobleski, D. & Oxenham, A.J. Recovering sound sources from embedded repetition. *Proc. Natl. Acad. Sci. USA* **108**, 1188–1193 (2011).
35. Carlyon, R.P., Micheyl, C., Deeks, J.M. & Moore, B.C.J. Auditory processing of real and illusory changes in frequency modulation (FM) phase. *J. Acoust. Soc. Am.* **116**, 3629–3639 (2004).
36. Lyzenga, J., Carlyon, R.P. & Moore, B.C.J. Dynamic aspects of the continuity illusion: perception of level and of the depth, rate and phase of modulation. *Hear. Res.* **210**, 30–41 (2005).
37. Cutting, J.E. & Rosner, B. Categories and boundaries in speech and music. *Percept. Psychophys.* **16**, 564–571 (1974).
38. Nahum, M., Nelken, I. & Ahissar, M. Low-level information and high-level perception: The case of speech in noise. *PLoS Biol.* **6**, e126 (2008).
39. Ariely, D. Seeing sets: Representation by statistical properties. *Psychol. Sci.* **12**, 157–162 (2001).
40. Chong, S.C. & Treisman, A. Representation of statistical properties. *Vision Res.* **43**, 393–404 (2003).
41. Haberman, J. & Whitney, D. Seeing the mean: ensemble coding for sets of faces. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 718–734 (2009).
42. Parkes, L., Lund, J., Angelucci, A., Solomon, J.A. & Morgan, M. Compulsory averaging of crowded orientation signals in human vision. *Nat. Neurosci.* **4**, 739–744 (2001).
43. Greenwood, J.A., Bex, P.J. & Dakin, S.C. Positional averaging explains crowding with letter-like stimuli. *Proc. Natl. Acad. Sci. USA* **106**, 13130–13135 (2009).
44. Balas, B., Nakano, L. & Rosenholtz, R. A summary-statistic representation in peripheral vision explains visual crowding. *J. Vis.* **9**, 1–18 (2009).
45. Freeman, J. & Simoncelli, E.P. Metamers of the ventral stream. *Nat. Neurosci.* **14**, 1195–1201 (2011).
46. Alvarez, G.A. & Oliva, A. Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proc. Natl. Acad. Sci. USA* **106**, 7345–7350 (2009).
47. Yabe, H. *et al.* Temporal window of integration of auditory information in the human brain. *Psychophysiology* **35**, 615–619 (1998).
48. Poeppel, D. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Commun.* **41**, 245–255 (2003).
49. Viemeister, N.F. & Wakefield, G.H. Temporal integration and multiple looks. *J. Acoust. Soc. Am.* **90**, 858–865 (1991).
50. Elhilali, M. & Shamma, S.A. A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. *J. Acoust. Soc. Am.* **124**, 3751–3771 (2008).

ONLINE METHODS

Synthetic textures. Sound texture stimuli were synthesized using a previously published method[9]. Statistics were first measured in 7-s recordings of real-world sound textures processed in an auditory model (**Fig. 1a**). Sounds were then synthesized to have the same values of these statistics. The statistics that were used for synthesis are depicted in **Figure 1a** and are time averages of simple functions of either the cochlear envelopes or the modulation bands. They include the mean, variance, skew and kurtosis of each cochlear envelope, the correlation coefficient between pairs of cochlear envelopes, the normalized power in each modulation band, and two types of correlations between modulation bands. Statistics were measured with a temporal weighting function that faded to zero over the first and last second of the original recordings to avoid boundary artifacts. The statistics were imposed on 5-s samples of Gaussian noise, with a flat weighting function (circular boundary conditions). All synthesis parameters were as described previously[9].

Four 5-s synthetic exemplars were synthesized from the statistics of each real-world recording. For the analysis shown in **Figure 1c**, we measured the statistics in excerpts taken from the first and last half (2.5 s) of each of the four exemplars of each of the 61 textures used in the experiments (**Supplementary Table 1**), yielding eight sets of statistics per texture per excerpt length (it would have been largely equivalent to synthesize eight 2.5-s exemplars, but these are somewhat more vulnerable to artifacts from the circular boundary conditions of the synthesis process). The s.d. of these eight values for each texture and excerpt length was computed; the median and quartiles of the s.d. across textures are plotted in **Figure 1c** for the textures used in the experiments over a range of excerpt lengths. Statistics involving modulation bands centered at rates whose period was more than the excerpt length were not included in this analysis, as such modulations are too slow to be meaningfully measured. However, this decision was not crucial to the qualitative form of **Figure 1c**.

The measurements of variability shown in **Figure 1c** are computed for the statistics in our texture model, the same set used for synthesis. The textures were generated such that these statistics were matched for a 5-s segment (within the tolerance used to test for convergence of the synthesis procedure), which ensures that the s.d. of those particular statistics would approach zero by 5 s. However, convergence with increasing excerpt length is not unique to these statistics; any statistic that is an average over time will converge to fixed values as the excerpt length increases, assuming that the excerpts are drawn from a stationary signal for which the statistic in question is well-defined. Convergence is also not unique to the synthetic textures that we used. For instance, the same phenomenon is evident in the textures that we generated for experiments 4 and 5 by superimposing real-world recordings of individual sound sources (**Fig. 3e**).

Experiments 1 and 2. To prevent performance in the texture discrimination task of experiment 1 from reaching ceiling levels, we chose pairs of textures that sounded relatively similar. For each texture in the set used previously[9], we selected a companion texture that was judged subjectively by the first two authors to be most similar in its texture qualities (out of the other textures in the set). The quality of the match was then rated on a scale of 1–5, and pairs with ratings below 3 were eliminated. We make no claim that these pairings were optimal or objective, but they served to keep performance levels in an informative range (that is, below ceiling).

From these pairs, we selected textures whose statistics had relatively low variability when measured in 2.5-s segments. To get a single measure of variability (pooled across statistics), we divided each statistic's s.d. by the mean of the statistic for that texture (because different statistics have different units), and averaged these normalized s.d. measurements across statistics to get a single measure of the statistic variability of each texture. We then ranked the texture pairs according to the statistic variability of the first texture in the pair, and chose the 50 pairs with the lowest variation for use in experiments. Both textures in a pair were used in experiment 1; the first texture in a pair was used in experiment 2 (**Supplementary Table 1**).

The sound stimuli presented on a trial of experiment 1 consisted of one excerpt from each of two exemplars of the first texture in a pair, and one excerpt from one exemplar of the second texture in that pair. The excerpts on a trial were either 40, 91, 209, 478, 1,093 or 2,500 ms in duration (chosen to be equally spaced on a logarithmic scale). The beginning of each excerpt was randomly chosen from within the 5-s exemplar, with the same beginning point for each of the three sounds. A 10-ms half-Hanning window was applied to the beginning and end of each sound. Each excerpt was presented at 70 dB SPL. Excerpts were separated by an ISI of 400 ms, and the window in which subjects entered their response (by clicking a button with a mouse) appeared 40 ms after the end of the last sound.

The order in which the two experiments were completed was counterbalanced across subjects. Subjects completed two blocks of trials across 2 d, each of which contained one trial per condition per sound pair for a total of 300 trials, randomly intermixed. In all experiments, subjects were recruited from the New York University community via poster advertisements (and as such were inexperienced psychophysical observers, naive as to the hypotheses), and had self-reported normal hearing. Subjects gave written consent (overseen by the New York University Institutional Review Board) and were paid an hourly wage in compensation for their efforts.

Nineteen subjects participated in experiments 1 and 2. Subjects were included in the analysis if their performance exceeded 85% correct in at least one condition of both experiments 1 and 2. This inclusion criterion was neutral with respect to the hypotheses and was designed to exclude subjects who were unable or unmotivated to perform the task. Seven subjects were excluded by this criterion, yielding 12 subjects whose data were analyzed (all female, averaging 21.1 years of age, s.d. of 3.0 years).

Here and in all other experiments, subjects were first given a small number (less than 20) of practice trials with feedback. Feedback consisted of a message stating whether their response was correct or incorrect. Stimulus excerpts on practice trials were distinct from those in the main experiment. In the main experiment, feedback was not provided.

Experiment 3. Subjects completed two versions of the exemplar discrimination task: part A, which was identical to experiment 2 (in which the ISI was fixed across conditions), and part B, in which the inter-onset interval was fixed. We tested only two excerpt durations in part B (91 ms and 2,500 ms), as the purpose was to investigate the effect of time delay on performance with short excerpts. In part B, the inter-onset interval was set to 2,900 ms (the inter-onset interval for the longest duration condition in part A). The short excerpts were thus separated by extended silences. The time interval between the onset of the third sound in a trial and the onset of the response window was also fixed across condition, to 2,540 ms (again, to match part A). All other parameters in part A and part B were identical to those of experiment 2.

Eighteen new subjects participated in experiment 3. The order in which parts A and B were completed was counterbalanced across subjects. Subjects completed one block of 300 trials of part A and one block of 200 trials (two trials per texture per condition) of part B. Six subjects were again excluded from analysis because they did not exceed 85% correct in at least one condition, yielding 12 subjects whose data were analyzed (nine female, averaging 22.2 years of age, s.d. of 5.3 years).

Experiment 4. Speech stimuli were generated from freely available online recordings of German speakers reading books. We obtained recordings of 115 unique speakers (62 male), from which pauses and breaths were excised. Mixtures of these speakers were created by adding together random selections of speakers at equal amplitudes. The excerpts used on a trial were randomly selected from two unique 5-s segments of the resulting signals as they were in experiment 2. The analysis of **Figure 3e** was conducted on 30 speech textures produced in this way for each density (using four 5-s excerpts, divided into two 2.5-s segments, with the same procedure used in the analysis shown in **Fig. 1c**). The 115-speaker textures had the same set of speakers in each instantiation; each of the 30 textures for the other densities had a distinct set of speakers. To bring performance below ceiling, we embedded the speech mixtures in pink noise (5 dB below the level of the speech). Without noise, a similar interaction was observed between duration and mixture density, but ceiling levels of performance obscured differences between short and long durations for the single- and seven-speaker conditions in most subjects (**Supplementary Fig. 2**). The combined stimulus was presented at 70 dB SPL.

Seventeen subjects completed 30 trials in each of eight conditions (short and long durations × four mixture densities). Five subjects were excluded from analysis because they were at ceiling in more than one condition, yielding 12 subjects whose data were analyzed (nine female, averaging 21.3 years of age, s.d. of 4.4 years).

Experiment 5. Drum stimuli were generated from MIDI renditions of 25 different drum or percussion sounds. Single drum hits were synthesized as WAV files (1.8 s in duration), and multiple copies of the resulting waveforms were superimposed with different random time offsets to generate texture stimuli. Four exemplars of 25 different textures were generated at each density, each containing four randomly selected drum types, used with equal probability. Textures were generated with 5, 60 and 250 hits per second. We generated 8-s signals by sampling 8*d (where d is the density in hits per s) onset times from a uniform distribution between 0 and 8 and laying down a drum hit at each onset time. The level of each drum hit was randomly selected from a uniform distribution spanning 10 dB. The first and last 1.5 s were then truncated, leaving a 5-s texture without boundary artifacts. The analysis of **Figure 3e** was conducted on these 5-s exemplars with the same procedure used for **Figure 1c**. The short and long excerpts presented in the experimental trials were drawn from the texture examples as they were in experiment 2, and were presented at 70 dB SPL. The task was otherwise like that of experiment 2. 19 subjects completed 25 trials (one per texture) in each of six conditions (short and long duration × three mixture densities). Seven subjects were excluded from the analysis because their performance did not exceed 0.85 in at least one condition or because they were at ceiling in more than one condition. This yielded 12 subjects whose data were analyzed (ten female, averaging 24.4 years old, s.d. of 6.7 years).

Statistical tests. ANOVAs and *t* tests (two tailed) were used to test for statistical significance. These tests were performed on percent correct scores passed through the inverse of the normal cumulative distribution function (which served to make their distribution more closely approximate a normal distribution). Prior to this, scores of 0 and 1 were changed to 0.001 and 0.999, respectively, to avoid infinite values. Sphericity was verified with Maunchly's test.

doi:10.1038/nn.3347