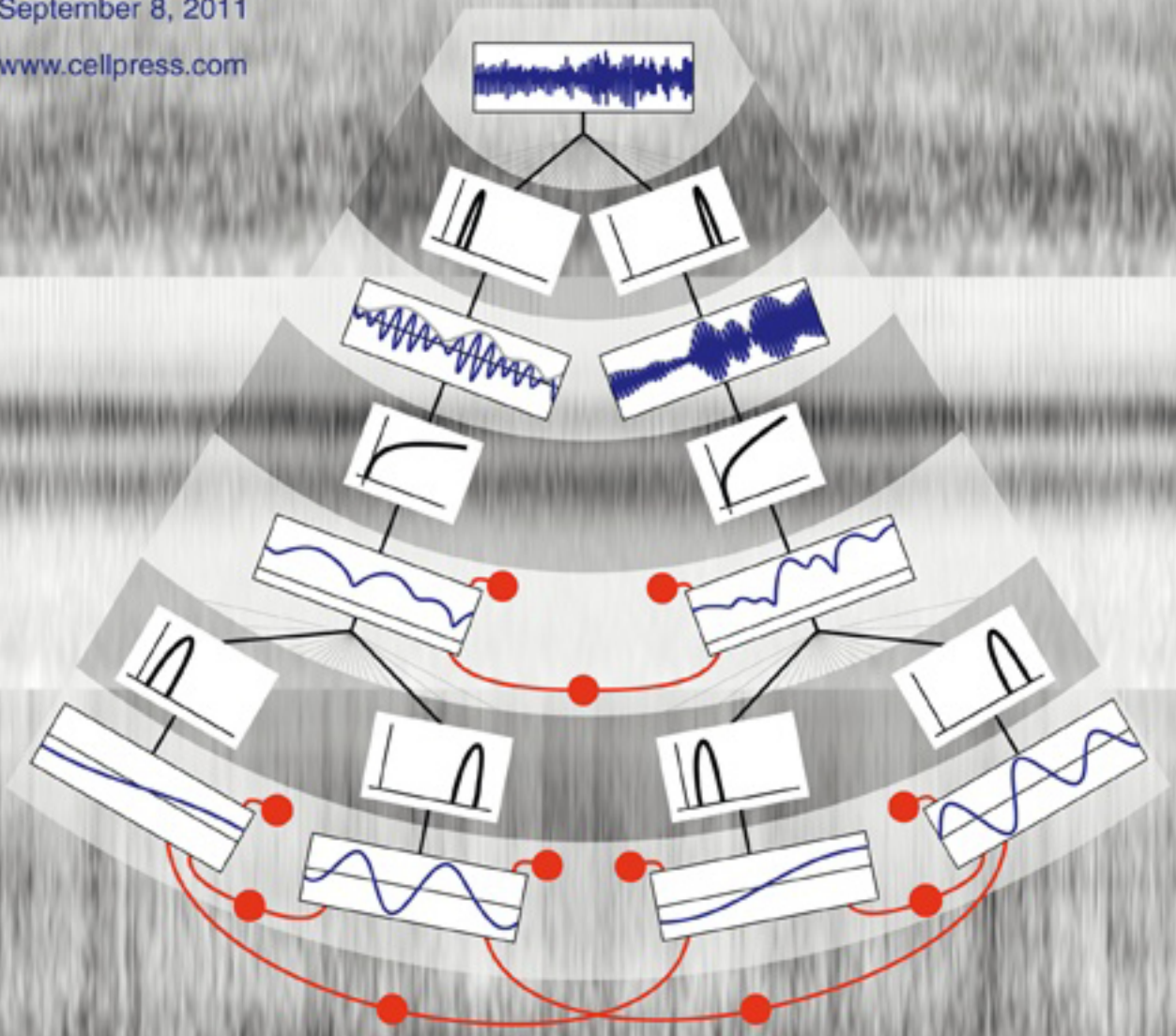
Neuron

Volume 71
Number 5
September 8, 2011

www.cellpress.com



Perceiving Sound Textures

Reviews:
Astrocytes in Neurovascular Coupling
Dendritic Spines and Distributed Circuits

Neuron
Article

Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis

Josh H. McDermott[1,2,*] and Eero P. Simoncelli[1,2,3]
[1]Howard Hughes Medical Institute
[2]Center for Neural Science
[3]Courant Institute of Mathematical Sciences
New York University, New York, NY 10003, USA
*Correspondence: jhm@cns.nyu.edu
DOI 10.1016/j.neuron.2011.06.032

SUMMARY

Rainstorms, insect swarms, and galloping horses produce "sound textures"—the collective result of many similar acoustic events. Sound textures are distinguished by temporal homogeneity, suggesting they could be recognized with time-averaged statistics. To test this hypothesis, we processed real-world textures with an auditory model containing filters tuned for sound frequencies and their modulations, and measured statistics of the resulting decomposition. We then assessed the realism and recognizability of novel sounds synthesized to have matching statistics. Statistics of individual frequency channels, capturing spectral power and sparsity, generally failed to produce compelling synthetic textures; however, combining them with correlations between channels produced identifiable and natural-sounding textures. Synthesis quality declined if statistics were computed from biologically implausible auditory models. The results suggest that sound texture perception is mediated by relatively simple statistics of early auditory representations, presumably computed by downstream neural populations. The synthesis methodology offers a powerful tool for their further investigation.

INTRODUCTION

Sensory receptors measure light, sound, skin pressure, and other forms of energy, from which organisms must recognize the events that occur in the world. Recognition is believed to occur via the transformation of sensory input into representations in which stimulus identity is explicit (for instance, via neurons responsive to one category but not others). In the auditory system, as in other modalities, much is known about how this process begins, from transduction through the initial stages of neural processing. Something is also known about the system's output, reflected in the ability of human listeners to recognize sounds. Less is known about what happens in the middle—the stages between peripheral processing and perceptual decisions. The difficulty of studying these mid-level processing stages partly reflects a lack of appropriate stimuli, as the tones and noises that are staples of classical hearing research do not capture the richness of natural sounds.

Here we study "sound texture," a category of sound that is well-suited for exploration of mid-level auditory perception. Sound textures are produced by a superposition of many similar acoustic events, such as arise from rain, fire, or a swamp full of insects, and are analogous to the visual textures that have been studied for decades (Julesz, 1962). Textures are a rich and varied set of sounds, and we show here that listeners can readily recognize them. However, unlike the sound of an individual event, such as a footstep, or of the complex temporal sequences of speech or music, a texture is defined by properties that remain constant over time. Textures thus possess a simplicity relative to other natural sounds that makes them a useful starting point for studying auditory representation and sound recognition.

We explored sound texture perception using a model of biological texture representation. The model begins with known processing stages from the auditory periphery and culminates with the measurement of simple statistics of these stages. We hypothesize that such statistics are measured by subsequent stages of neural processing, where they are used to distinguish and recognize textures. We tested the model by conducting psychophysical experiments with synthetic sounds engineered to match the statistics of real-world textures. The logic of the approach, borrowed from vision research, is that if texture perception is based on a set of statistics, two textures with the same values of those statistics should sound the same (Julesz, 1962; Portilla and Simoncelli, 2000). In particular, our synthetic textures should sound like another example of the corresponding real-world texture if the statistics used for synthesis are similar to those measured by the auditory system.

Although the statistics we investigated are relatively simple and were not hand-tuned to specific natural sounds, they produced compelling synthetic examples of many real-world textures. Listeners recognized the synthetic sounds nearly as well as their real-world counterparts. In contrast, sounds synthesized using representations distinct from those in biological auditory systems generally did not sound as compelling. Our results suggest that the recognition of sound textures is based on statistics of modest complexity computed from the responses of the peripheral auditory system. These statistics likely reflect sensitivities of downstream neural populations. Sound textures and their synthesis thus provide a substrate for studying mid-level audition.

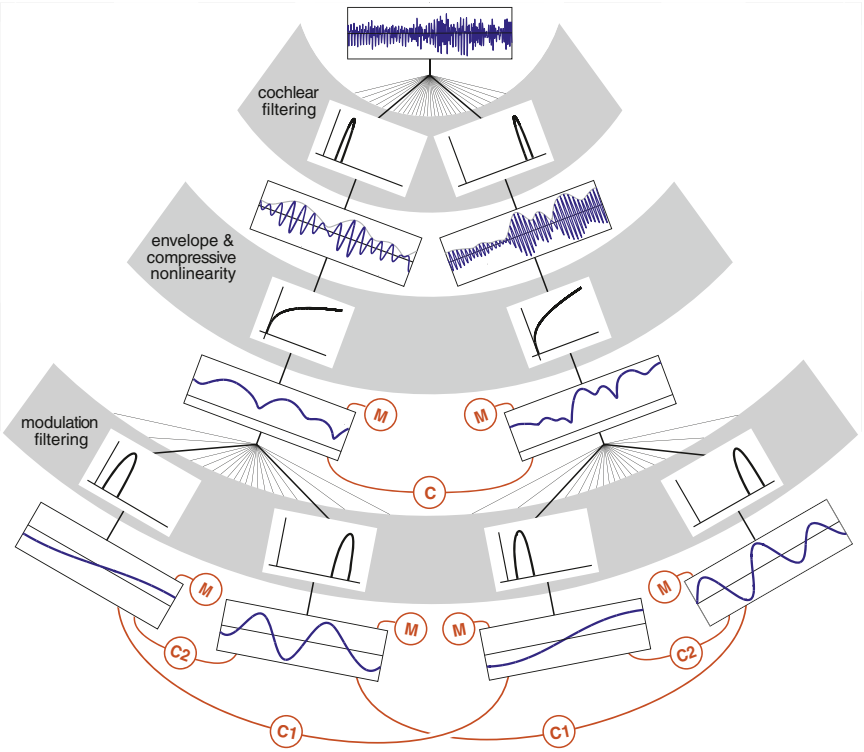


Figure 1. Model Architecture

A sound waveform (top row) is filtered by a "cochlear" filterbank (gray stripe contains two example filters at different frequencies, on a log-frequency axis). Cochlear filter responses (i.e., subbands) are bandlimited versions of the original signal (third row), the envelopes of which (in gray) are passed through a compressive nonlinearity (gray stripe, fourth row), yielding compressed envelopes (fifth row), from which marginal statistics and cross band correlations are measured. Envelopes are filtered with a modulation filter bank (gray stripe, sixth row, containing two example filters for each of the two example cochlear channels, on a log-frequency axis), the responses of which (seventh row) are used to compute modulation marginals and correlations. Red icons denote statistical measurements: marginal moments of a single signal or correlations between two signals.

RESULTS

Our investigations of sound texture were constrained by three sources of information: auditory physiology, natural sound statistics, and perceptual experiments. We used the known structure of the early auditory system to construct the initial stages of our model and to constrain the choices of statistics. We then established the plausibility of different types of statistics by verifying that they vary across natural sounds and could thus be useful for their recognition. Finally, we tested the perceptual importance of different texture statistics with experiments using synthetic sounds.

Texture Model

Our model is based on a cascade of two filter banks (Figure 1) designed to replicate the tuning properties of neurons in early stages of the auditory system, from the cochlea through the thalamus. An incoming sound is first processed with a bank of 30 bandpass cochlear filters that decompose the sound waveform into acoustic frequency bands, mimicking the frequency selectivity of the cochlea. All subsequent processing is performed on the amplitude envelopes of these frequency bands. Amplitude envelopes can be extracted from cochlear responses with a low-pass filter and are believed to underlie many aspects of peripheral auditory responses (Joris et al., 2004). When the envelopes are plotted in grayscale and arranged vertically, they form a spectrogram, a two-dimensional (time versus frequency) image commonly used for visual depiction of sound (e.g., Figure 2A). Perceptually, envelopes carry much of the important information in natural sounds (Gygi et al., 2004; Shannon et al., 1995; Smith et al., 2002), and can be used to reconstruct signals that are perceptually indistinguishable from the original in which the envelopes were measured. Cochlear transduction of sound is also distinguished by amplitude compression (Ruggero, 1992)—the response to high intensity sounds is proportionally smaller than that to low intensity sounds, due to nonlinear, level-dependent amplification. To simulate this phenomenon, we apply a compressive nonlinearity to the envelopes.

Each compressed envelope is further decomposed using a bank of 20 bandpass modulation filters. Modulation filters are conceptually similar to cochlear filters, except that they operate on (compressed) envelopes rather than the sound pressure waveform, and are tuned to frequencies an order of magnitude lower, as envelopes fluctuate at relatively slow rates. A modulation filter bank is consistent with previous auditory models (Bacon and Grantham, 1989; Dau et al., 1997) as well as reports of modulation tuning in midbrain and thalamic neurons (Baumann et al., 2011; Joris et al., 2004; Miller et al., 2002; Rodríguez et al., 2010). Both the cochlear and modulation filters in our model had bandwidths that increased with their center frequency (such that they were approximately constant on a logarithmic scale), as is observed in biological auditory systems.

From cochlear envelopes and their modulation bands, we derive a representation of texture by computing statistics (red symbols in Figure 1). The statistics are time-averages of nonlinear functions of either the envelopes or the modulation bands. Such statistics are in principle suited to summarizing stationary signals like textures, whose properties are constant over some moderate timescale. A priori, however, it is not obvious whether simple, biologically plausible statistics would have much explanatory power as descriptors of natural sounds or of their perception. Previous attempts to model sound texture have come from the machine audio and sound rendering communities (Athineos and Ellis, 2003; Dubnov et al., 2002; Saint-Arnaud and Popat, 1995; Verron et al., 2009; Zhu and Wyse, 2004) and

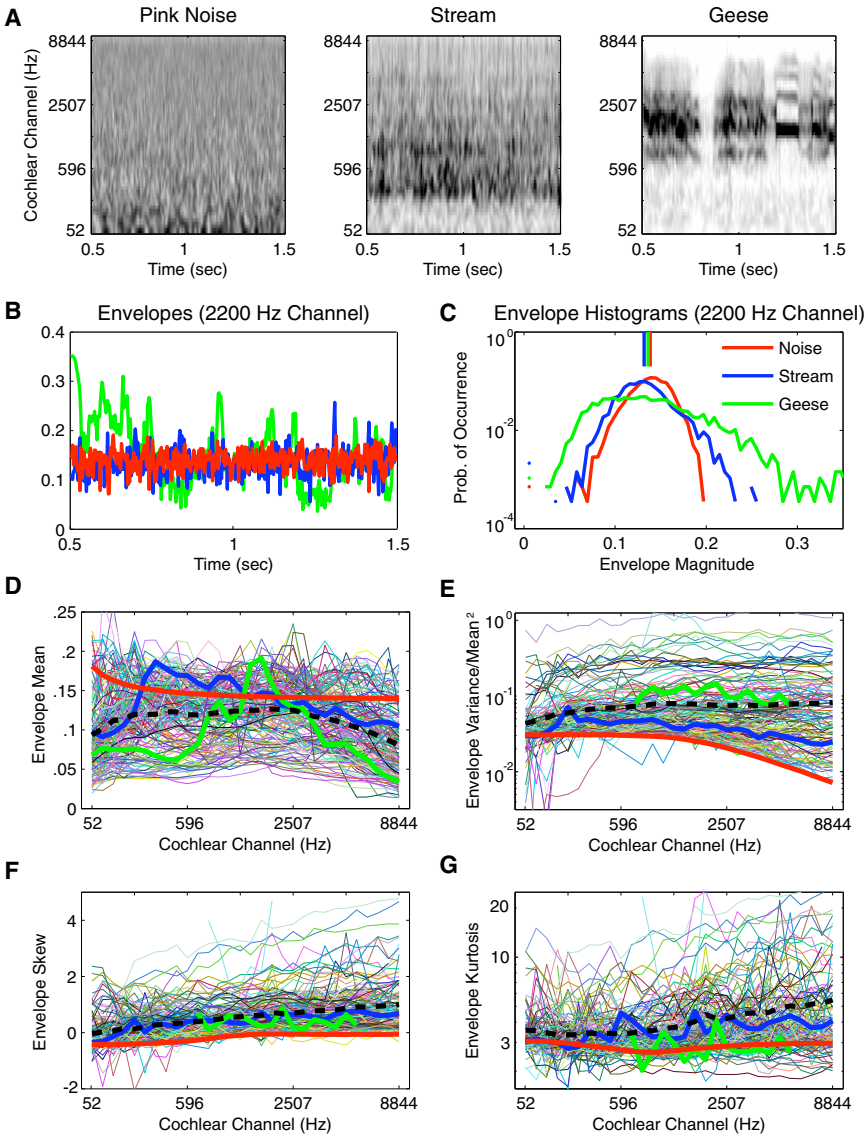


Figure 2. Cochlear Marginal Statistics

(A) Spectrograms of three sound excerpts, generated by plotting the envelopes of a cochlear filter decomposition. Gray-level indicates the (compressed) envelope amplitude (same scale for all three sounds).

(B) Envelopes of one cochlear channel for the three sounds from (A).

(C) Histograms (gathered over time) of the envelopes in (B). Vertical line segments indicate the mean value of the envelope for each sound.

(D–G) Envelope marginal moments for each cochlear channel of each of 168 natural sound textures. Moments of sounds in (A–C) are plotted with thick lines; dashed black line plots the mean value of each moment across all sounds.

have involved representations unrelated to those in biological auditory systems.

Texture Statistics

Of all the statistics the brain could compute, which might be used by the auditory system? Natural sounds can provide clues: in order for a statistic to be useful for recognition, it must produce different values for different sounds. We considered a set of generic statistics and verified that they varied substantially across a set of 168 natural sound textures.

We examined two general classes of statistic: marginal moments and pairwise correlations. Both types of statistic involve averages of simple nonlinear operations (e.g., squaring, products) that could plausibly be measured using neural circuitry at a later stage of neural processing. Moments and correlations derive additional plausibility from their importance in the representation of visual texture (Heeger and Bergen, 1995; Portilla and Simoncelli, 2000), which provided inspiration for our work. Both types of statistic were computed on cochlear envelopes as well as their modulation bands (Figure 1). Because modulation filters are applied to the output of a particular cochlear channel, they are tuned in both acoustic frequency and modulation frequency. We thus distinguished two types of modulation correlations: those between bands tuned to the same modulation frequency but different acoustic frequencies (C1), and those between bands tuned to the same acoustic frequency but different modulation frequencies (C2).

To provide some intuition for the variation in statistics that occurs across sounds, consider the cochlear marginal moments: statistics that describe the distribution of the envelope amplitude for a single cochlear channel. Figure 2A shows the envelopes, displayed as spectrograms, for excerpts of three example sounds (pink [1/f] noise, a stream, and geese calls), and Figure 2B plots the envelopes of one particular channel for each sound. It is visually apparent that the envelopes of the three sounds are distributed differently—those of the geese contain more high-amplitude and low-amplitude values than those of the stream or noise. Figure 2C shows the envelope distributions for one cochlear channel. Although the mean envelope values are nearly equal in this example (because they have roughly the same average acoustic power in that channel), the envelope distributions differ in width, asymmetry about the mean, and the presence of a long positive tail. These properties can be captured by the marginal moments (mean, variance, skew, and kurtosis, respectively). Figures 2D–2G show these moments for our full set of sound textures. Marginal moments have previously been proposed to play a role in envelope discrimination (Lorenzi et al., 1999; Strickland and Viemeister, 1996), and often reflect

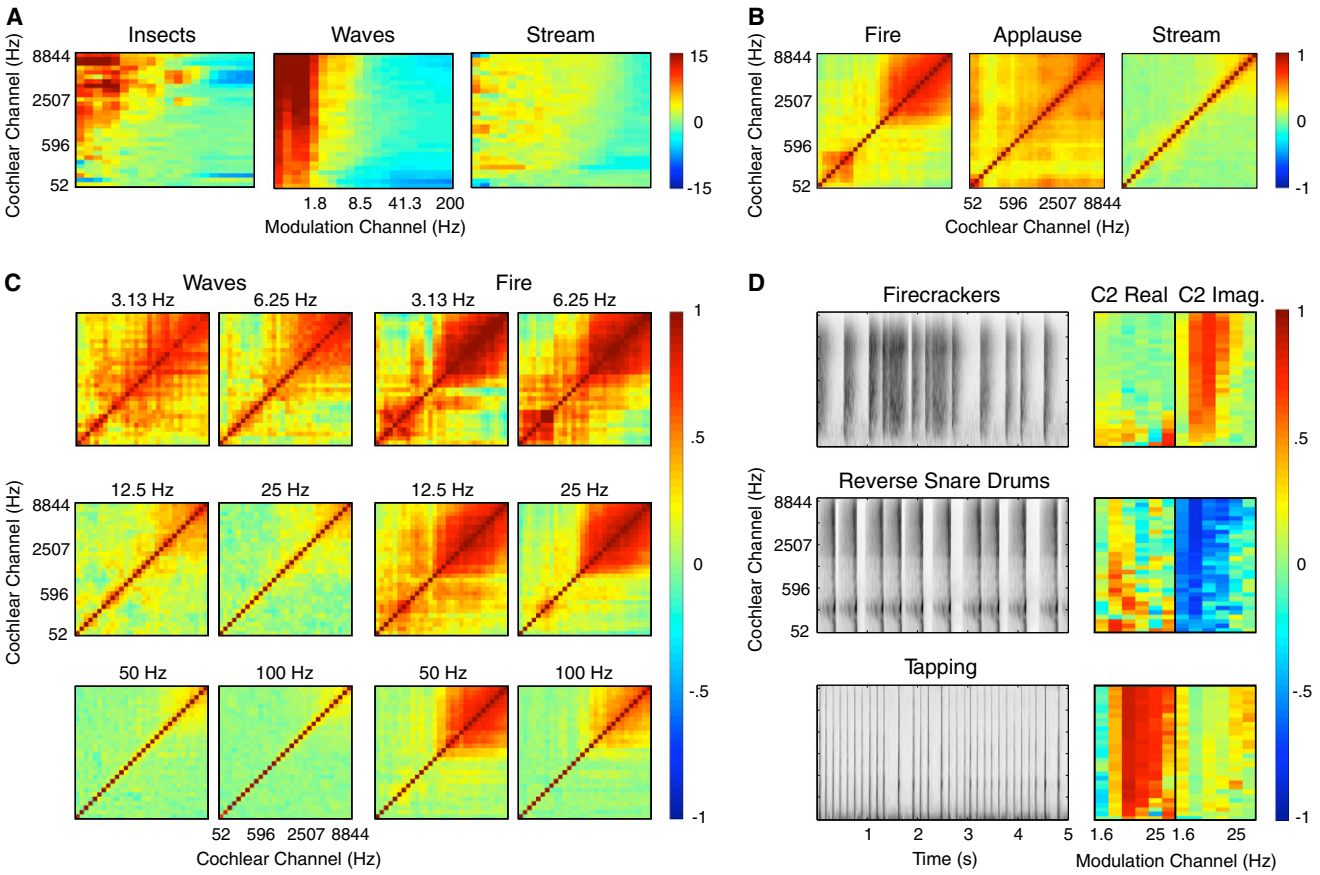


Figure 3. Modulation Power and Correlation Statistics

(A) Modulation power in each band (normalized by the variance of the corresponding cochlear envelope) for insects, waves, and stream sounds of Figure 4B. For ease of display and interpretation, this statistic is expressed in dB relative to the same statistic for pink noise.

(B) Cross-band envelope correlations for fire, applause, and stream sounds of Figure 4B. Each matrix cell displays the correlation coefficient between a pair of cochlear envelopes.

(C) C1 correlations for waves and fire sounds of Figure 4B. Each matrix contains correlations between modulation bands tuned to the same modulation frequency but to different acoustic frequencies, yielding matrices of the same format as (B), but with a different matrix for each modulation frequency, indicated at the top of each matrix.

(D) Spectrograms and C2 correlations for three sounds. Note asymmetric envelope shapes in first and second rows, and that abrupt onsets (top), offsets (middle), and impulses (bottom) produce distinct correlation patterns. In right panels, modulation channel labels indicate the center of low-frequency band contributing to the correlation. See also Figure S6.

the property of sparsity, which tends to characterize natural sounds and images (Field, 1987; Attias and Schreiner, 1998). Intuitively, sparsity reflects the discrete events that generate natural signals; these events are infrequent, but produce a burst of energy when they occur, yielding high-variance amplitude distributions. Sparsity has been linked to sensory coding (Field, 1987; Olshausen and Field, 1996; Smith and Lewicki, 2006), but its role in the perception of real-world sounds has been unclear.

Each of the remaining statistics we explored (Figure 1) captures distinct aspects of acoustic structure and also exhibits large variation across sounds (Figure 3). The moments of the modulation bands, particularly the variance, indicate the rates at which cochlear envelopes fluctuate, allowing distinction between rapidly modulated sounds (e.g., insect vocalizations) and slowly modulated sounds (e.g., ocean waves). The correlation statistics, in contrast, each reflect distinct aspects of coordination between envelopes of different channels, or between their modulation bands. The cochlear correlations (C) distinguish textures with broadband events that activate many channels simultaneously (e.g., applause), from those that produce nearly independent channel responses (many water sounds; see Experiment 1: Texture Identification). The cross-channel modulation correlations (C1) are conceptually similar except that they are computed on a particular modulation band of each cochlear channel. In some sounds (e.g., wind, or waves) the C1 correlations are large only for low modulation-frequency bands, whereas in others (e.g., fire) they are present across all bands. The within-channel modulation correlations (C2) allow discrimination between sounds with sharp onsets or offsets (or both), by capturing the relative phase relationships between modulation bands within a cochlear channel. See Experimental Procedures for detailed descriptions.

Sound Synthesis

Our goal in synthesizing sounds was not to render maximally realistic sounds per se, as in most sound synthesis applications (Dubnov et al., 2002; Verron et al., 2009), but rather to test hypotheses about how the brain represents sound texture, using realism as an indication of the hypothesis validity. Others have also noted the utility of synthesis for exploring biological auditory representations (Mesgarani et al., 2009; Slaney, 1995); our work is distinct for its use of statistical representations. Inspired by methods for visual texture synthesis (Heeger and Bergen, 1995; Portilla and Simoncelli, 2000), our method produced novel signals that matched some of the statistics of a real-world sound. If the statistics used to synthesize the sound are similar to those used by the brain for texture recognition, the synthetic signal should sound like another example of the original sound.

To synthesize a texture, we first obtained desired values of the statistics by measuring the model responses (Figure 1) for a real-world sound. We then used an iterative procedure to modify a random noise signal (using variants of gradient descent) to force it to have these desired statistic values (Figure 4A). By starting from noise, we hoped to generate a signal that was as random as possible, constrained only by the desired statistics.

Figure 4B displays spectrograms of several naturally occurring sound textures along with synthetic examples generated from their statistics (see Figure S1 available online for additional examples). It is visually apparent that the synthetic sounds share many structural properties of the originals, but also that the process has not simply regenerated the original sound—here and in every other example we examined, the synthetic signals were physically distinct from the originals (see also Experiment 1: Texture Identification [Experiment 1b, condition 7]). Moreover, running the synthesis procedure multiple times produced exemplars with the same statistics but whose spectrograms were easily discriminated visually (Figure S2). The statistics we studied thus define a large set of sound signals (including the original in which the statistics are measured), from which one member is drawn each time the synthesis process is run.

To assess whether the synthetic results sound like the natural textures whose statistics they matched, we conducted several experiments. The results can also be appreciated by listening to example synthetic sounds, available online (http://www.cns.nyu.edu/~lcv/sound_texture.html).

Experiment 1: Texture Identification

We first tested whether synthetic sounds could be identified as exemplars of the natural sound texture from which their statistics were obtained. Listeners were presented with example sounds, and chose an identifying name from a set of five. In Experiment 1a, sounds were synthesized using different subsets of statistics. Identification was poor when only the cochlear channel power was imposed (producing a sound with roughly the same power spectrum as the original), but improved as additional statistics were included as synthesis constraints (Figure 5A; $F[2.25, 20.25] = 124.68$, $p < 0.0001$; see figure for paired comparisons between conditions). Identifiability of textures synthesized using the full model approached that obtained for the original sound recordings.

Inspection of listeners' responses revealed several results of interest (Figures 5B and 5C). In condition 1, when only the cochlear channel power was imposed, the sounds most often correctly identified were those that are noise-like (wind, static, etc.); such sounds were also the most common incorrect answers. This is as expected, because the synthesis process was initialized with noise and in this condition simply altered its spectrum. A more interesting pattern emerged for condition 2, in which the cochlear marginal moments were imposed. In this condition, but not others, the sounds most often identified correctly, and chosen incorrectly, were water sounds. This is readily apparent from listening to the synthetic examples—water often sounds realistic when synthesized from its cochlear marginals, and most other sounds synthesized this way sound water-like.

Because the cochlear marginal statistics only constrain the distribution of amplitudes within individual frequency channels, this result suggests that the salient properties of water sounds are conveyed by sparsely distributed, independent, bandpass acoustic events. In Experiment 1b, we further explored this result: in conditions 1 and 2 we again imposed marginal statistics, but used filters that were either narrower or broader than the filters found in biological auditory systems. Synthesis with these alternative filters produced overall levels of performance similar to the auditory filter bank (condition 3; Figure 5D), but in both cases, water sounds were no longer the most popular choices (Figures 5E and 5F; the four water categories were all identified less well, and chosen incorrectly less often, in conditions 1 and 2 compared to condition 3; $p < 0.01$, sign test). It thus seems that the bandwidths of biological auditory filters are comparable to those of the acoustic events produced by water (see also Figure S3), and that water sounds often have remarkably simple structure in peripheral auditory representations.

Although cochlear marginal statistics are adequate to convey the sound of water, in general they are insufficient for recognition (Figure 5A). One might expect that with a large enough set of filters, marginal statistics alone would produce better synthesis, because each filter provides an additional set of constraints on the sound signal. However, our experiments indicate otherwise. When we synthesized sounds using a filter bank with the bandwidths of our canonical model, but with four times as many filters (such that adjacent filters overlapped more than in the original filter bank), identification was not significantly improved [Figure 5D; condition 4 versus 3, $t(9) = 1.27$, $p = 0.24$]. Similarly, one might suppose that constraining the full marginal distribution (as opposed to just matching the four moments in our model) might capture more structure, but we found that this also failed to produce improvements in identification [Figure 5D; condition 5 versus 3, $t(9) = 1.84$, $p = 0.1$; Figure S4]. These results suggest that cochlear marginal statistics alone, irrespective of how exhaustively they are measured, cannot account for our perception of texture.

Because the texture model is independent of the signal length, we could measure statistics from signals much shorter or longer than those being synthesized. In both cases the results generally sounded as compelling as if the synthetic and original signals were the same length. To verify this empirically, in condition 7

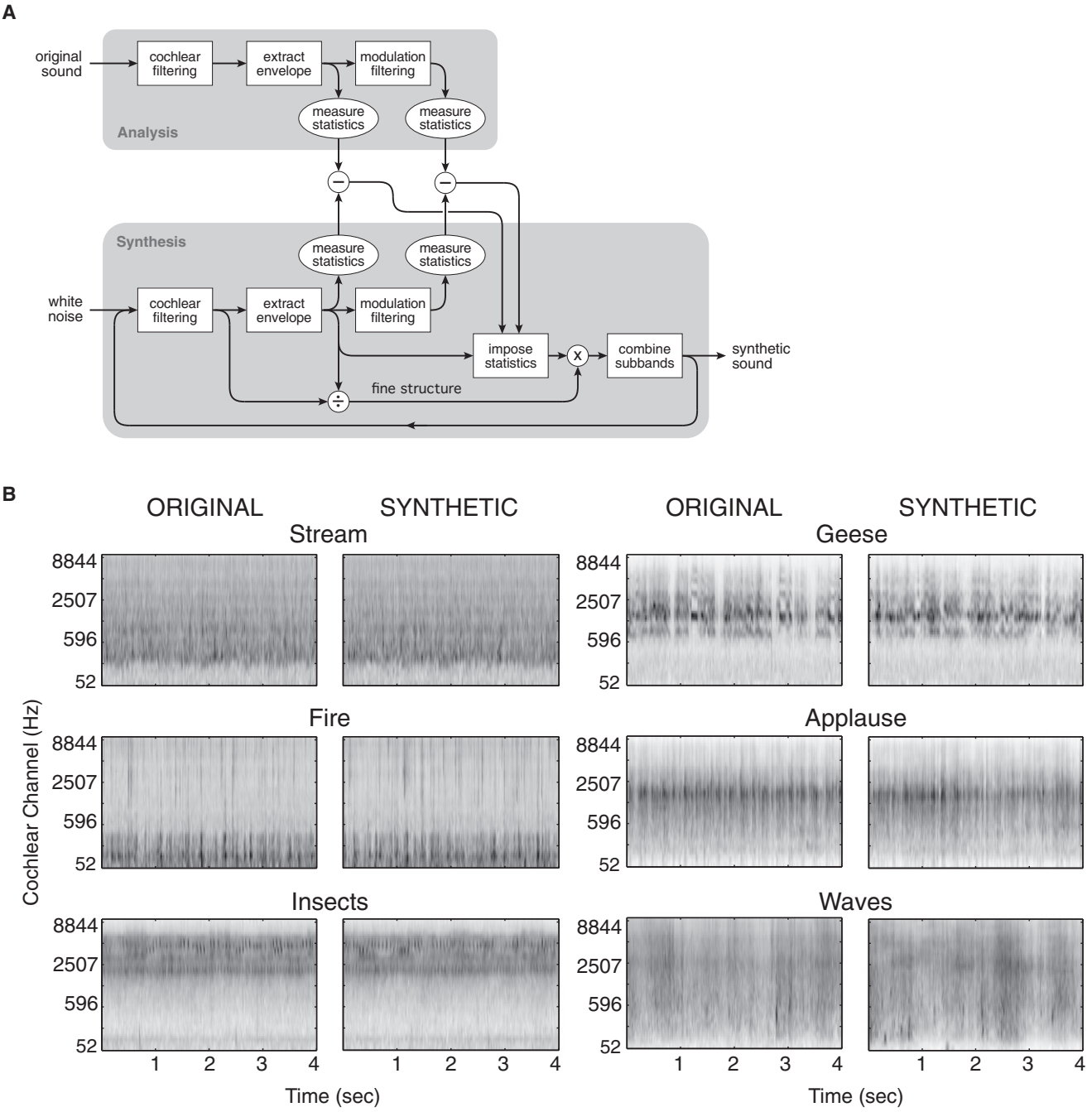


Figure 4. Synthesis Algorithm and Example Results

(A) Schematic of synthesis procedure. Statistics are measured after a sound recording is passed through the auditory model of Figure 1. Synthetic signal is initialized as noise, and the original sound's statistics are imposed on its cochlear envelopes. The modified envelopes are multiplied by their associated fine structure, and then recombined into a sound signal. The procedure is iterated until the synthesized signal has the desired statistics.

(B) Spectrograms of original and synthetic versions of several sounds (same amplitude scale for all sounds). See also Figure S1 and Figure S2.

we used excerpts of 15 s signals synthesized from 7 s originals. Identification performance was unaffected [Figure 5D; condition 7 versus 6; $t(9) = 0.5$, $p = 0.63$], indicating that these longer signals captured the texture qualities as well as signals more comparable to the original signals in length.

Experiment 2: Necessity of Each Class of Statistic

We found that each class of statistic was perceptually necessary, in that its omission from the model audibly impaired the quality of some synthetic sounds. To demonstrate this empirically, in Experiment 2a we presented listeners with excerpts

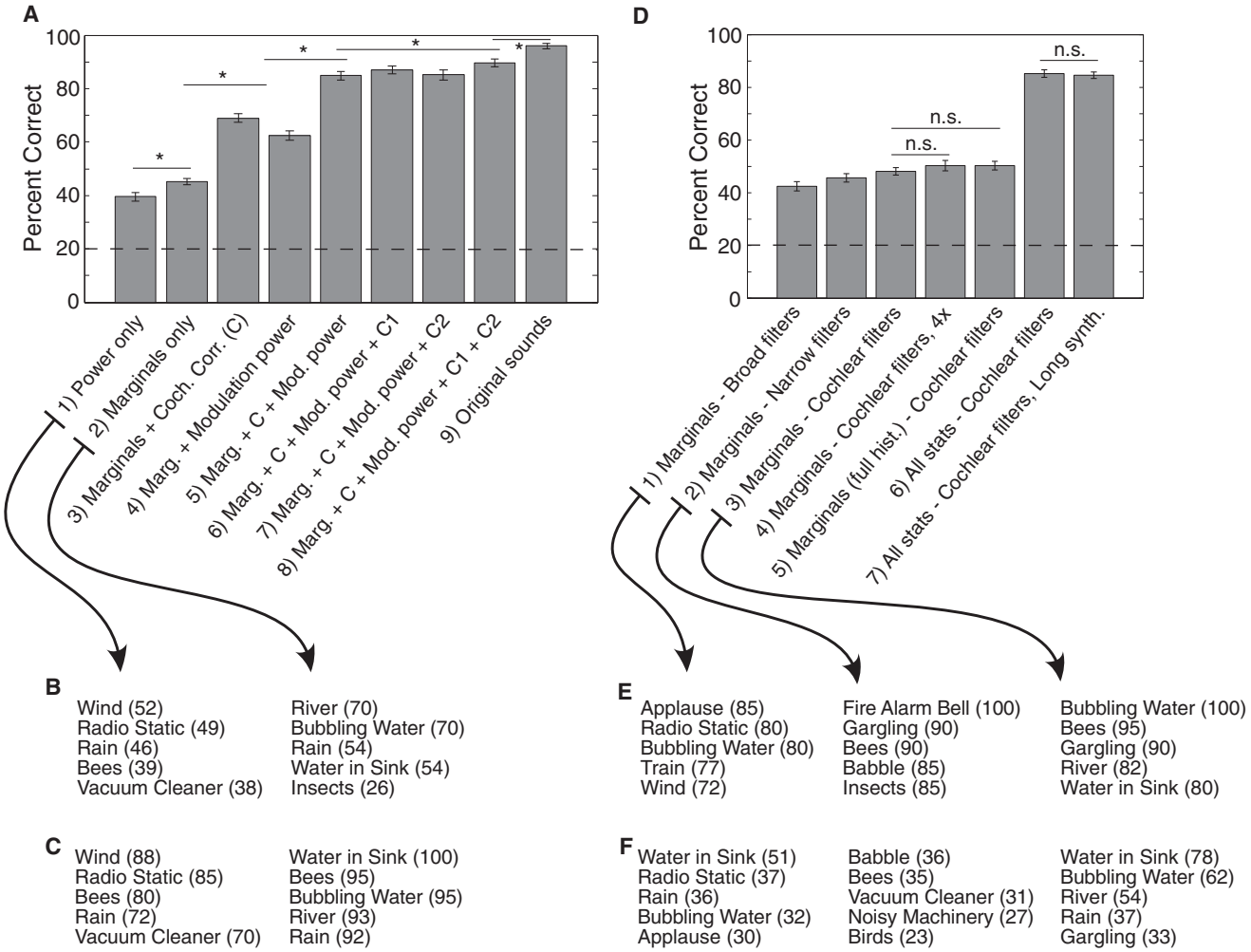


Figure 5. Experiment 1: Texture Identification

(A) Identification improves as more statistics are included in the synthesis. Asterisks denote significant differences between conditions, $p < 0.01$ (paired t tests, corrected for multiple comparisons). Here and elsewhere, error bars denote standard errors and dashed lines denote the chance level of performance.

(B) The five categories correctly identified most often for conditions 1 and 2, with mean percent correct in parentheses.

(C) The five categories chosen incorrectly most often for conditions 1 and 2, with mean percent trials chosen (of those where they were a choice) in parentheses.

(D) Identification with alternative marginal statistics, and long synthetic signals. Horizontal lines indicate nonsignificant differences ($p > 0.05$).

(E and F) The five (E) most correctly identified and (F) most often incorrectly chosen categories for conditions 1–3. See also Figure S3 and Figure S4.

of original texture recordings followed by two synthetic versions—one synthesized using the full set of model statistics, and the other synthesized with one class omitted—and asked them to judge which synthetic version sounded more like the original. Figure 6A plots the percentage of trials on which the full set of statistics was preferred. In every condition, this percentage was greater than that expected by chance (t tests, $p < 0.01$ in all cases, Bonferroni corrected), though the preference was stronger for some statistic classes than others [$F(4,36) = 15.39$, $p < 0.0001$].

The effect of omitting a statistic class was not noticeable for every texture. A potential explanation is that the statistics of many textures are close to those of noise for some subset of statistics, such that omitting that subset does not cause the statistics of the synthetic result to deviate much from the correct values (because the synthesis is initialized with noise). To test this idea, we computed the difference between each sound's statistics and those of pink (1/f) noise, for each of the five statistic classes. When we reanalyzed the data including only the 30% of sounds whose statistics were most different from those of noise, the proportion of trials on which the full set of statistics was preferred was significantly higher in each case (t tests, $p < 0.05$). Including a particular statistic in the synthesis process thus tends to improve realism when the value of that statistic deviates from that of noise. Because of this, not all statistics are necessary for the synthesis of every texture (although all statistics presumably contribute to the perception of every texture—if the values were actively perturbed from their correct values, whether noise-like or not, we found that listeners generally noticed).

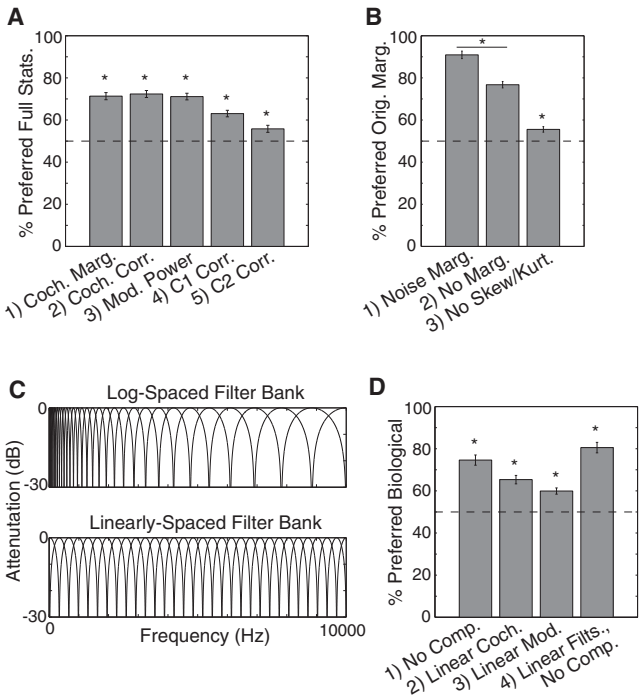


Figure 6. Experiments 2 and 3: Omitting and Manipulating Statistics

(A) Experiment 2a: synthesis with the full set of statistics is preferred over synthesis omitting any single class. Condition labels indicate statistic class omitted. In condition 1, the envelope mean was imposed, to ensure that the spectrum was approximately correct. Asterisks denote significant differences from chance, $p < 0.01$.

(B) Experiment 2b: sounds with the correct cochlear marginal statistics were preferred over those with (1), the cochlear marginal moments of noise; (2), all cochlear marginals omitted (as in condition 1 of [A]); or (3), the skew and kurtosis omitted. Asterisks denote significant differences from chance or between conditions, $p < 0.01$.

(C) Frequency responses of logarithmically and linearly spaced cochlear filter banks.

(D) Experiment 3: sounds synthesized with a biologically plausible auditory model were preferred over those synthesized with models deviating from biology (by omitting compression, or by using linearly spaced cochlear or modulation filter banks). Asterisks denote significant differences from chance, $p < 0.01$.

We expected that the C2 correlation, which measures phase relations between modulation bands, would help capture the temporal asymmetry of abrupt onsets or offsets. To test this idea, we separately analyzed sounds that visually or audibly possessed such asymmetries (explosions, drum beats, etc.). For this subset of sounds, and for other randomly selected subsets, we computed the average proportion of trials in which synthesis with the full set of statistics was preferred over that with the C2 correlation omitted. The preference for the full set of statistics was larger in the asymmetric sounds than in 99.96% of other subsets, confirming that the C2 correlations were particularly important for capturing asymmetric structure.

It is also notable that omitting the cochlear marginal moments produced a noticeable degradation in realism for a large fraction of sounds, indicating that the sparsity captured by these statistics is perceptually important. As a further test, we explicitly forced sounds to be nonsparse and examined the effect on perception. We synthesized sounds using a hybrid set of statistics in which the envelope variance, skew, and kurtosis were taken from pink noise, with all other statistics given the correct values for a particular real-world sound. Because noise is nonsparse (the marginals of noise lie at the lower extreme of the values for natural sounds; Figure 2), this manipulation forced the resulting sounds to lack sparsity but to maintain the other statistical properties of the original sound. We found that the preference for signals with the correct marginals was enhanced in this condition [1 versus 2, $t(9) = 8.1$, $p < 0.0001$; Figure 6B], consistent with the idea that sparsity is perceptually important for most natural sound textures. This result is also an indication that the different classes of statistic are not completely independent: constraining the other statistics had some effect on the cochlear marginals, bringing them closer to the values of the original sound even if they themselves were not explicitly constrained. We also found that listeners preferred sounds synthesized with all four marginal moments to those with the skew and kurtosis omitted ($t(8) = 4.1$, $p = 0.003$). Although the variance alone contributes substantially to sparsity, the higher-order moments also play some role.

Experiment 3: Statistics of Nonbiological Sound Representations

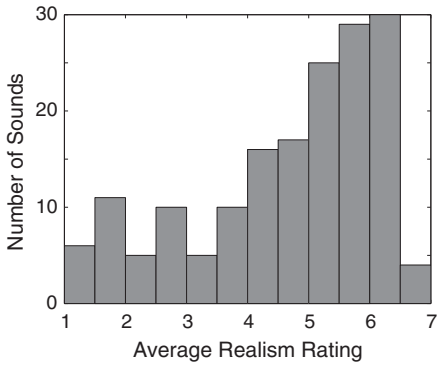
How important are the biologically inspired features of our model? One might expect that any large and varied set of statistics would produce signals that resemble the originals. As a test, we altered our model in three respects: (1) removing cochlear compression, (2), altering the bandwidths of the "cochlear" filters, and (3) altering the bandwidths of the modulation filters (rows four, two, and six of Figure 1). In the latter two cases, linearly spaced filter banks were substituted for the log-spaced filter banks found in biological auditory systems (Figure 6C). We also included a condition with all three alterations. Each altered model was used both to measure the statistics in the original sound signal, and to impose them on synthetic sounds. In all cases, the number of filters was preserved, and thus all models had the same number of statistics.

We again performed an experiment in which listeners judged which of two synthetic sounds (one generated from our biologically inspired model, the other from one of the nonbiological models) more closely resembled the original from which their statistics were measured. In each condition, listeners preferred synthetic sounds produced by the biologically inspired model (Figure 6D; sign tests, $p < 0.01$ in all conditions), supporting the notion that the auditory system represents textures using statistics similar to those in this model.

Experiment 4: Realism Ratings

To illustrate the overall effectiveness of the synthesis, we measured the realism of synthetic versions of every sound in our set. Listeners were presented with an original recording followed by a synthetic signal matching its statistics. They rated the extent to which the synthetic signal was a realistic example of the original sound, on a scale of 1–7. Most sounds yielded average ratings above 4 (Figures 7A and 7B; Table S1). The sounds with low ratings, however, are of particular interest, as they are

A



B

Synthetic sounds with high realism ratings

6.57	Insects in swamp
6.57	Heavy rain on hard surface
6.53	Frogs
6.47	Applause - big room
6.43	Radio static
6.43	Stream
6.40	Air conditioner
6.37	Frying eggs
6.33	Wind blowing
6.20	Sparrows - large excited group
6.17	Jackhammer
6.13	Water trickling into pool
6.00	Fire - forest inferno
5.90	Bee swarm
5.90	Rustling paper
5.87	Train speeding down tracks
5.87	Rattlesnake rattle
5.70	Cocktail party babble
5.67	Shaking coins
5.67	Helicopter

C

Synthetic sounds with lowest realism ratings

1.93	Railroad crossing
1.90	Tapping rhythm - quarter note pairs
1.77	Wind chimes
1.77	Corkscrew scraping desk edge
1.70	Reverse snare drum beats
1.70	Tapping rhythm - quarter note triplets
1.67	Snare drum beats
1.63	Walking on gravel
1.60	Snare drum rimshot sequence
1.60	Music - drum break
1.50	Music - mambo
1.50	Bongo drum loop
1.47	Firecracker explosions
1.40	Person speaking French
1.37	Church bells
1.20	Person speaking English

Figure 7. Experiment 4: Realism Ratings

(A) Histogram of average realism ratings for each sound in our set.

(B) List of 20 sound textures with high average ratings. Multiple examples of similar sounds are omitted for brevity.

(C) List of all sounds with average realism ratings <2, along with their average rating. See Table S1 for complete list. See also Figure S5.

statistically matched to the original recordings and yet do not sound like them. Figure 7C lists the sounds with average ratings below 2. They fall into three general classes—those involving pitch (railroad crossing, wind chimes, music, speech, bells), rhythm (tapping, music, drumming), and reverberation (drum beats, firecrackers); see also Figure S5. This suggests that the perception of these sound attributes involves measurements substantially different from those in our model.

DISCUSSION

We have studied "sound textures," a class of sounds produced by multiple superimposed acoustic events, as are common to many natural environments. Sound textures are distinguished by temporal homogeneity, and we propose that they are represented in the auditory system with time-averaged statistics. We embody this hypothesis in a model based on statistics (moments and correlations) of a sound decomposition like that found in the subcortical auditory system. To test the role of these statistics in texture recognition, we conducted experiments with synthetic sounds matching the statistics of various real-world textures. We found that (1) such synthetic sounds could be accurately recognized, and at levels far better than if only the spectrum or sparsity was matched, (2) eliminating subsets of the statistics in the model reduced the realism of the synthetic results, (3) modifying the model to less faithfully mimic the mammalian auditory system also reduced the realism of the synthetic sounds, and (4) the synthetic results were often realistic, but failed markedly for a few particular sound classes.

Our results suggest that when listeners recognize the sound of rain, fire, insects, and other such sounds, they are recognizing statistics of modest complexity computed from the output of the peripheral auditory system. These statistics are likely measured at downstream stages of neural processing, and thus provide clues to the nature of mid-level auditory computations.

Neural Implementation

Because texture statistics are time averages, their computation can be thought of as involving two steps: a nonlinear function applied to the relevant auditory response(s), followed by an average over time. A moment, for instance, could be computed by a neuron that averages its input (e.g., a cochlear envelope) after raising it to a power (two for the variance, three for the skew, etc.). We found that envelope moments were crucial for producing naturalistic synthetic sounds. Envelope moments convey sparsity, a quality long known to differentiate natural signals from noise (Field, 1987) and one that is central to many recent signal-processing algorithms (Asari et al., 2006; Bell and Sejnowski, 1996). Our results thus suggest that sparsity is represented in the auditory system and used to distinguish sounds. Although definitive characterization of the neural locus awaits, neural responses in the midbrain often adapt to particular amplitude distributions (Dean et al., 2005; Kvale and Schreiner, 2004), raising the possibility that envelope moments may be computed subcortically. The modulation power (also a marginal moment) at particular rates also seems to be reflected in the tuning of many thalamic and midbrain neurons (Joris et al., 2004).

The other statistics in our model are correlations. A correlation is the average of a normalized product (e.g., of two cochlear envelopes), and could be computed as such. However, a correlation can also be viewed as the proportion of variance in one variable that is shared by another, which is partly reflected in the variance of the sum of the variables. This formulation provides an alternative implementation (see Experimental Procedures), and illustrates that correlations in one stage of representation (e.g., bandpass cochlear channels) can be reflected in the marginal statistics of the next (e.g., cortical neurons that sum input from multiple channels), assuming appropriate convergence. All of the texture statistics we have considered could thus reduce to marginal statistics at different stages of the auditory system.

Neuronal tuning to texture statistics could be probed using synthetic stimuli whose statistics are parametrically varied.

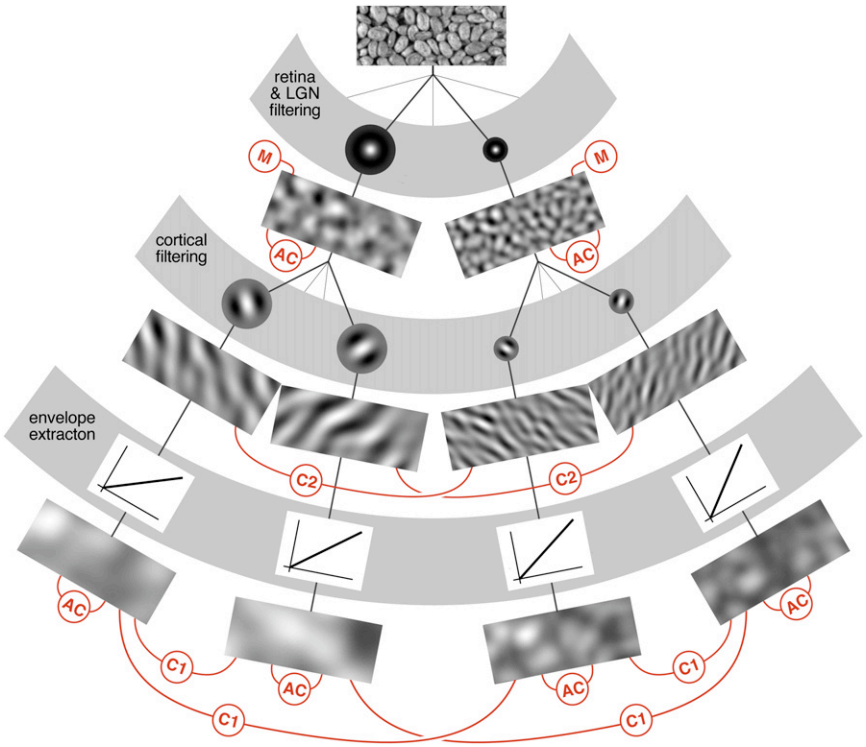


Figure 8. Analogous Model of Visual Texture Representation

Model is depicted in a format like that of the auditory texture model in Figure 1. An image of beans (top row) is filtered into spatial frequency bands by center-surround filters (second row), as happens in the retina/LGN. The spatial frequency bands (third row) are filtered again by orientation selective filters (fourth row) analogous to V1 simple cells, yielding scale and orientation filtered bands (fifth row). The envelopes of these bands are extracted (sixth row) to produce analogs of V1 complex cell responses (seventh row). The linear function at the envelope extraction stage indicates the absence of the compressive nonlinearity present in the auditory model. As in Figure 1, red icons denote statistical measurements: marginal moments of a single signal (M) or correlations between two signals (AC, C1, or C2 for autocorrelation, cross-band correlation, or phase-adjusted correlation). C1 and C2 here and in Figure 1 denote conceptually similar statistics. The autocorrelation (AC) is identical to C1 except that it is computed within a channel. This model is a variant of Portilla and Simoncelli (2000).

Stationary artificial sounds have a long history of use in psychoacoustics and neurophysiology, with recent efforts to incorporate naturalistic statistical structure (Attias and Schreiner, 1998; Garcia-Lazaro et al., 2006; McDermott et al., 2011; Overath et al., 2008; Rieke et al., 1995; Singh and Theunissen, 2003). Stimuli synthesized from our model capture naturally occurring sound structure while being precisely characterized within an auditory model. They offer a middle ground between natural sounds and the tones and noises of classical hearing research.

Relation to Visual Texture

Visual textures, unlike their auditory counterparts, have been studied intensively for decades (Julesz, 1962), and our work was inspired by efforts to understand visual texture using synthesis (Heeger and Bergen, 1995; Portilla and Simoncelli, 2000; Zhu et al., 1997). How similar are visual and auditory texture representations? For ease of comparison, Figure 8 shows a model diagram of the most closely related visual texture model (Portilla and Simoncelli, 2000), analogous in format to our auditory model (Figure 1) but with input signals and representational stages that vary spatially rather than temporally. The vision model has two stages of linear filtering (corresponding to LGN cells and V1 simple cells) followed by envelope extraction (corresponding to V1 complex cells), whereas the auditory model has the envelope operation sandwiched between linear filtering operations (corresponding to the cochlea and midbrain/thalamus), reflecting structural differences in the two systems. There are also notable differences in the stages at which statistics are computed in the two models: several types of visual texture statistics are computed directly on the initial linear filtering stages, whereas the auditory statistics all follow the envelope operation, reflecting the primary locus of structure in images versus sounds. However, the statistical computations themselves—marginal moments and correlations—are conceptually similar in the two models. In both systems, relatively simple statistics capture texture structure, suggesting that texture perception, like filling in (McDermott and Oxenham, 2008; Warren et al., 1972), and saliency (Cusack and Carlyon, 2003; Kayser et al., 2005), may involve analogous computations across modalities.

It will be interesting to explore whether the similarities between modalities extend to inattention, to which visual texture is believed to be robust (Julesz, 1962). Under conditions of focused listening, we are often aware of individual events composing a sound texture, presumably in addition to registering time-averaged statistics that characterize the texture qualities. A classic example is the "cocktail party problem," in which we attend to a single person talking in a room dense with conversations (Bee and Micheyl, 2008; McDermott, 2009). Without attention, individual voices or other sound sources are likely inaccessible, but we may retain access to texture statistics that characterize the combined effect of multiple sources, as is apparently the case in vision (Alvarez and Oliva, 2009). This possibility could be tested in divided attention experiments with synthetic textures.

Texture Extensions

We explored the biological representation of sound texture using a set of generic statistics and a relatively simple auditory model, both of which could be augmented in interesting ways. The three sources of information that contributed to the present work—auditory neuroscience, natural sound analysis, and perceptual experiments—all provide directions for such extensions.

The auditory model of Figure 1, from which statistics are computed, captures neuronal tuning characteristics of subcortical

structures. Incorporating cortical tuning properties would likely extend the range of textures we can account for. For instance, cortical receptive fields have spectral tuning that is more complex and varied than that found subcortically (Barbour and Wang, 2003; Depireux et al., 2001), and statistics of filters modeled on their properties could capture higher-order structure that our current model does not. As discussed earlier, the correlations computed on subcortical representations could then potentially be replaced by marginal statistics of filters at a later stage.

It may also be possible to derive additional or alternative texture statistics from an analysis of natural sounds, similar in spirit to previous derivations of cochlear and V1 filters from natural sounds and images (Olshausen and Field, 1996; Smith and Lewicki, 2006), and consistent with other examples of congruence between properties of perceptual systems and natural environments (Attias and Schreiner, 1998; Garcia-Lazaro et al., 2006; Lesica and Grothe, 2008; Nelken et al., 1999; Rieke et al., 1995; Rodríguez et al., 2010; Schwartz and Simoncelli, 2001; Woolley et al., 2005). We envision searching for statistics that vary maximally across sounds and would thus be optimal for recognition.

The sound classes for which the model failed to produce convincing synthetic examples (revealed by Experiment 4) also provide directions for exploration. Notable failures include textures involving pitched sounds, reverberation, and rhythmic structure (Figure 7, Table S1, and Figure S5). It was not obvious a priori that these sounds would produce synthesis failures—they each contain spectral and temporal structures that are stationary (given a moderately long time window), and we anticipated that they might be adequately constrained by the model statistics. However, our results show that this is not the case, suggesting that the brain is measuring something that the model is not.

Rhythmic structure might be captured with another stage of envelope extraction and filtering, applied to the modulation bands. Such filters would measure "second-order" modulation of modulation (Lorenzi et al., 2001), as is common in rhythmic sounds. Alternatively, rhythm could involve a measure specifically of periodic modulation patterns. Pitch and reverberation may also implicate dedicated mechanisms. Pitch is largely conveyed by harmonically related frequencies, which are not made explicit by the pair-wise correlations across frequency found in our current model (see also Figure S5). Accounting for pitch is thus likely to require a measure of local harmonic structure (de Cheveigne, 2004). Reverberation is also well understood from a physical generative standpoint, as linear filtering of a sound source by the environment (Gardner, 1998), and is used to judge source distance (Zahorik, 2002) and environment properties. However, a listener has access only to the result of environmental filtering, not to the source or the filter, implying that reverberation must be reflected in something measured from the sound signal (i.e., a statistic). Our synthesis method provides an unexplored avenue for testing theories of the perception of these sound properties.

One other class of failures involved mixtures of two sounds that overlap in peripheral channels but are acoustically distinct, such as broadband clicks and slow bandpass modulations. These failures likely result because the model statistics are averages over time, and combine measurements that should be segregated. This suggests a more sophisticated form of estimating statistics, in which averaging is performed after (or in alternation with) some sort of clustering operation, a key ingredient in recent models of stream segregation (Elhilali and Shamma, 2008).

Using Texture to Understand Recognition

Recognition is challenging because the sensory input arising from different exemplars of a particular category in the world often varies substantially. Perceptual systems must process their input to obtain representations that are invariant to the variation within categories, while maintaining selectivity between categories (DiCarlo and Cox, 2007). Our texture model incorporates an explicit form of invariance by representing all possible exemplars of a given texture (Figure S2) with a single set of statistic values. Moreover, different textures produce different statistics, providing an implicit form of selectivity. However, our model captures texture properties with a large number of simple statistics that are partially redundant. Humans, in contrast, categorize sounds into semantic classes, and seem to have conscious access to a fairly small set of perceptual dimensions. It should be possible to learn such lower-dimensional representations of categories from our sound statistics, combining the full set of statistics into a small number of "metastatistics" that relate to perceptual dimensions. We have found, for instance, that most of the variance in statistics over our collection of sounds can be captured with a moderate number of their principal components, indicating that dimensionality reduction is feasible.

The temporal averaging through which our texture statistics achieve invariance is appropriate for stationary sounds, and it is worth considering how this might be relaxed to represent sounds that are less homogeneous. A simple possibility involves replacing the global time-averages with averages taken over a succession of short time windows. The resulting local statistical measures would preserve some of the invariance of the global statistics, but would follow a trajectory over time, allowing representation of the temporal evolution of a signal. By computing measurements averaged within windows of many durations, the auditory system could derive representations with varying degrees of selectivity and invariance, enabling the recognition of sounds spanning a continuum from homogeneous textures to singular events.

EXPERIMENTAL PROCEDURES

Auditory Model

Our synthesis algorithm utilized a classic "subband" decomposition in which a bank of cochlear filters were applied to a sound signal, splitting it into frequency channels. To simplify implementation, we used zero-phase filters, with Fourier amplitude shaped as the positive portion of a cosine function. We used a bank of 30 such filters, with center frequencies equally spaced on an equivalent rectangular bandwidth $(ERB)_N$ scale (Glasberg and Moore, 1990), spanning 52–8844 Hz. Their (3 dB) bandwidths were comparable to those of the human ear (~5% larger than ERBs measured at 55 dB sound pressure level (SPL); we presented sounds at 70 dB SPL, at which human auditory filters are somewhat wider). The filters did not replicate all aspects of biological

auditory filters, but perfectly tiled the frequency spectrum—the summed squared frequency response of the filter bank was constant across frequency (to achieve this, the filter bank also included lowpass and highpass filters at the endpoints of the spectrum). The filter bank thus had the advantage of being invertible: each subband could be filtered again with the corresponding filter, and the results summed to reconstruct the original signal (as is standard in analysis-synthesis subband decompositions [Crochiere et al., 1976]).

The envelope of each subband was computed as the magnitude of its analytic signal, and the subband was divided by the envelope to yield the fine structure. The fine structure was ignored for the purposes of analysis (measuring statistics). Subband envelopes were raised to a power of 0.3 to simulate basilar membrane compression. For computational efficiency, statistics were measured and imposed on envelopes downsampled (following lowpass filtering) to a rate of 400 Hz. Although the envelopes of the high-frequency subbands contained modulations at frequencies above 200 Hz (because cochlear filters are broad at high frequencies), these were generally low in amplitude. In pilot experiments we found that using a higher envelope sampling rate did not produce noticeably better synthetic results, suggesting the high frequency modulations are not of great perceptual significance in this context.

The filters used to measure modulation power also had half-cosine frequency responses, with center frequencies equally spaced on a log scale (20 filters spanning 0.5–200 Hz), and a quality factor of 2 (for 3 dB bandwidths), consistent with those in previous models of human modulation filtering (Dau et al., 1997), and broadly consistent with animal neurophysiology data (Miller et al., 2002; Rodríguez et al., 2010). Although auditory neurons often exhibit a degree of tuning to spectral modulation as well (Depireux et al., 2001; Rodríguez et al., 2010; Schönwiesner and Zatorre, 2009), this is typically less pronounced than their temporal modulation tuning, particularly early in the auditory system (Miller et al., 2002), and we elected not to include it in our model. Because 200 Hz was the Nyquist frequency, the highest frequency filter consisted only of the lower half of the half-cosine frequency response.

We used a smaller set of modulation filters to compute the C1 and C2 correlations, in part because it was desirable to avoid large numbers of unnecessary statistics, and in part because the C2 correlations necessitated octave-spaced filters (see below). These filters also had frequency responses that were half-cosines on a log-scale, but were more broadly tuned ($Q = \sqrt{2}$), with center frequencies in octave steps from 1.5625 to 100 Hz, yielding seven filters.

Boundary Handling

All filtering was performed in the discrete frequency domain, and thus assumed circular boundary conditions. To avoid boundary artifacts, the statistics measured in original recordings were computed as weighted time-averages. The weighting window fell from one to zero (half cycle of a raised cosine) over the 1 s intervals at the beginning and end of the signal (typically a 7 s segment), minimizing artifactual interactions. For the synthesis process, statistics were imposed with a uniform window, so that they would influence the entire signal. As a result, continuity was imposed between the beginning and end of the signal. This was not obvious from listening to the signal once, but it enabled synthesized signals to be played in a continuous loop without discontinuities.

Statistics

We denote the k^{th} cochlear subband envelope by $s_k(t)$, and the windowing function by $w(t)$, with the constraint that $\sum_t w(t) = 1$. The n^{th} modulation band of cochlear envelope s_k is denoted by $b_{k,n}(t)$, computed via convolution with filter f_n.

Cochlear Marginal Statistics

Our texture representation includes the first four normalized moments of the envelope:

$$M1_k = \mu_k = \sum_t w(t) s_k(t),$$

$$M2_k = \frac{\sigma_k^2}{\mu_k^2} = \frac{\sum_t w(t)(s_k(t) - \mu_k)^2}{\mu_k^2},$$

$$M3_k = \frac{\sum_t w(t)(s_k(t) - \mu_k)^3}{\sigma_k^3},$$

and

$$M4_k = \frac{\sum_t w(t)(s_k(t) - \mu_k)^4}{\sigma_k^4} \quad k \in [1 \ldots 32] \text{ in each case.}$$

The variance was normalized by the squared mean, so as to make it dimensionless like the skew and kurtosis.

The envelope variance, skew, and kurtosis reflect subband sparsity. Sparsity is often associated with the kurtosis of a subband (Field, 1987), and preliminary versions of our model were also based on this measurement (McDermott et al., 2009). However, the envelope's importance in hearing made its moments a more sensible choice, and we found them to capture similar sparsity behavior.

Figures 2D–2G show the marginal moments for each cochlear envelope of each sound in our ensemble. All four statistics vary considerably across natural sound textures. Their values for noise are also informative. The envelope means, which provide a coarse measure of the power spectrum, do not have exceptional values for noise, lying in the middle of the set of natural sounds. However, the remaining envelope moments for noise all lie near the lower bound of the values obtained for natural textures, indicating that natural sounds tend to be sparser than noise (see also Experiment 2b) (Attias and Schreiner, 1998).

Cochlear Cross-Band Envelope Correlation

$$C_{jk} = \sum_t \frac{w(t)(s_j(t) - \mu_j)(s_k(t) - \mu_k)}{\sigma_j \sigma_k}, \; j, \; k \in [1 \ldots 32]$$

such that $(k - j) \in [1, 2, 3, 5, 8, 11, 16, 21]$.

Our model included the correlation of each cochlear subband envelope with a subset of eight of its neighbors, a number that was typically sufficient to reproduce the qualitative form of the full correlation matrix (interactions between overlapping subsets of filters allow the correlations to propagate across subbands). This was also perceptually sufficient: we found informally that imposing fewer correlations sometimes produced perceptually weaker synthetic examples, and that incorporating additional correlations did not noticeably improve the results.

Figure 3B shows the cochlear correlations for recordings of fire, applause, and a stream. The broadband events present in fire and applause, visible as vertical streaks in the spectrograms of Figure 4B, produce correlations between the envelopes of different cochlear subbands. Cross-band correlation, or "comodulation," is common in natural sounds (Nelken et al., 1999), and we found it to be to be a major source of variation among sound textures. The stream, for instance, contains much weaker comodulation.

The mathematical form of the correlation does not uniquely specify the neural instantiation. It could be computed directly, by averaging a product as in the above equation. Alternatively, it could be computed with squared sums and differences, as are common in functional models of neural computation (Adelson and Bergen, 1985):

$$C_{jk} = \sum_t w(t) \frac{(s_j(t) - \mu_j + s_k(t) - \mu_k)^2 - (s_j(t) - \mu_j - s_k(t) + \mu_k)^2}{4\sigma_j \sigma_k}.$$

Modulation Power

For the modulation bands, the variance (power) was the principal marginal moment of interest. Collectively, these variances indicate the frequencies present in an envelope. Analogous quantities appear to be represented by the modulation-tuned neurons common to the early auditory system (whose responses code the power in their modulation passband). To make the modulation power statistics independent of the cochlear statistics, we normalized each by the variance of the corresponding cochlear envelope; the measured statistics thus represent the proportion of total envelope power captured by each modulation band:

$$M_{k,n} = \frac{\sum_t w(t) b_{k,n}(t)^2}{\sigma_k^2}, \; k \in [1 \ldots 32], \; n \in [1 \ldots 20].$$

Note that the mean of the modulation bands is zero (because the filters f_n are zero-mean). The other moments of the modulation bands were either uninformative or redundant (see Supplemental Experimental Procedures) and were omitted from the model.

The modulation power implicitly captures envelope correlations across time, and is thus complementary to the cross-band correlations. Figure 3A shows the modulation power statistics for recordings of swamp insects, lake shore waves, and a stream.

Modulation Correlations

These correlations were computed using octave-spaced modulation filters (necessitated by the C2 correlations), the resulting bands of which are denoted by $\tilde{b}_{k,n}(t)$.

The C1 correlation is computed between bands centered on the same modulation frequency but different acoustic frequencies:

$$C1_{jk,n} = \frac{\sum_t w(t)\tilde{b}_{j,n}(t)\tilde{b}_{k,n}(t)}{\sigma_{j,n}\sigma_{k,n}},\ j \in [1\ldots32],\ (k-j) \in [1,2],\ n \in [2\ldots7],$$

and

$$\sigma_{j,n} = \sqrt{\sum_t w(t)\tilde{b}_{j,n}(t)^2}.$$

We imposed correlations between each modulation filter and its two nearest neighbors along the cochlear axis, for six modulation bands spanning 3–100 Hz.

C1 correlations are shown in Figure 3C for the sounds of waves and fire. The qualitative pattern of C1 correlations shown for waves is typical of a number of sounds in our set (e.g., wind). These sounds exhibit low-frequency modulations that are highly correlated across cochlear channels, but high-frequency modulations that are largely independent. This effect is not simply due to the absence of high-frequency modulation, as most such sounds had substantial power at high modulation frequencies (comparable to that in pink noise, evident from dB values close to zero in Figure 3A). In contrast, for fire (and many other sounds), both high and low frequency modulations exhibit correlations across cochlear channels. Imposing the C1 correlations was essential to synthesizing realistic waves and wind, among other sounds. Without them, the cochlear correlations affected both high and low modulation frequencies equally, resulting in artificial sounding results for these sounds.

C1 correlations did not subsume cochlear correlations. Even when larger numbers of C1 correlations were imposed (i.e., across more offsets), we found informally that the cochlear correlations were necessary for high quality synthesis.

The second type of correlation, labeled C2, is computed between bands of different modulation frequencies derived from the same acoustic frequency channel. This correlation represents phase relations between modulation frequencies, important for representing abrupt onsets and other temporal asymmetries. Temporal asymmetry is common in natural sounds, but is not captured by conventional measures of temporal structure (e.g., the modulation spectrum), as they are invariant to time reversal (Irino and Patterson, 1996). Intuitively, an abrupt increase in amplitude (e.g., a step edge) is generated by a sum of sinusoidal envelope components (at different modulation frequencies) that are aligned at the beginning of their cycles (phase $-\pi/2$), whereas an abrupt decrease is generated by sinusoids that align at the cycle midpoint (phase $\pi/2$), and an impulse (e.g., a click) has frequency components that align at their peaks (phase 0). For sounds dominated by one of these feature types, adjacent modulation bands thus have consistent relative phase in places where their amplitudes are high. We captured this relationship with a complex-valued correlation measure (Portilla and Simoncelli, 2000).

We first define analytic extensions of the modulation bands:

$$\alpha_{k,n}(t) \equiv \tilde{b}_{k,n}(t) + iH\left(\tilde{b}_{k,n}(t)\right),$$

where H denotes the Hilbert transform and $i = \sqrt{-1}$.

The analytic signal comprises the responses of the filter and its quadrature twin, and is thus readily instantiated biologically. The correlation has the standard form, except it is computed between analytic modulation bands tuned to modulation frequencies an octave apart, with the frequency of the lower band doubled. Frequency doubling is achieved by squaring the complex-valued analytic signal:

$$d_{k,n}(t) = \frac{a_{k,n}^2(t)}{\|a_{k,n}(t)\|},$$

yielding

$$C2_{k,mn} = \frac{\sum_t w(t)d_{k,m}^*(t)a_{k,n}(t)}{\sigma_{k,m}\sigma_{k,n}},$$

$k \in [1\ldots32]$, $m \in [1\ldots6]$, and $(n - m) = 1$, where * and $\|\cdot\|$ denote the complex conjugate and modulus, respectively.

Because the bands result from octave-spaced filters, the frequency doubling of the lower-frequency band causes them to oscillate at the same rate, producing a fixed phase difference between adjacent bands in regions of large amplitude. We use a factor of 2 rather than something smaller because the operation of exponentiating a complex number is uniquely defined only for integer powers. See Figure S6 for further explanation.

$C2_{k,mn}$ is complex valued, and the real and imaginary parts must be independently measured and imposed. Example sounds with onsets, offsets, and impulses are shown in Figure 3D along with their C2 correlations.

In total, there are 128 cochlear marginal statistics, 189 cochlear cross-correlations, 640 modulation band variances, 366 C1 correlations, and 192 C2 correlations, for a total of 1515 statistics.

Imposition Algorithm

Synthesis was driven by a set of statistics measured for a sound signal of interest using the auditory model described above. The synthetic signal was initialized with a sample of Gaussian white noise, and was modified with an iterative process until it shared the measured statistics. Each cycle of the iterative process, as illustrated in Figure 4A, consisted of the following steps:

(1) The synthetic sound signal is decomposed into cochlear subbands.
(2) Subband envelopes are computed using the Hilbert transform.
(3) Envelopes are divided out of the subbands to yield the subband fine structure.
(4) Envelopes are downsampled to reduce computation.
(5) Envelope statistics are measured and compared to those of the original recording to generate an error signal.
(6) Downsampled envelopes are modified using a variant of gradient descent, causing their statistics to move closer to those measured in the original recording.
(7) Modified envelopes are upsampled and recombined with the unmodified fine structure to yield new subbands.
(8) New subbands are combined to yield a new signal.

We performed conjugate gradient descent using Carl Rasmussen's "minimize" MATLAB function (available online). The objective function was the total squared error between the synthetic signal's statistics and those of the original signal. The subband envelopes were modified one-by-one, beginning with the subband with largest power, and working outwards from that. Correlations between pairs of subband envelopes were imposed when the second subband envelope contributing to the correlation was being adjusted.

Each episode of gradient descent resulted in modified subband envelopes that approached the target statistics. However, there was no constraint forcing the envelope adjustment to remain consistent with the subband fine structure (Ghitza, 2001), or to produce new subbands that were mutually consistent (in the sense that combining them would produce a signal that would yield the same subbands when decomposed again). It was thus generally the case that during the first few iterations, the envelopes measured at the beginning of cycle n + 1 did not completely retain the adjustment imposed at cycle n, because combining envelope and fine structure, and summing up the subbands, tended to change the envelopes in ways that altered their statistics. However, we found that with iteration, the envelopes generally converged to a state with the desired statistics. The fine structure was not directly constrained, and relaxed to a state consistent with the envelope constraints.

Convergence was monitored by computing the error in each statistic at the start of each iteration and measuring the signal-to-noise ratio (SNR) as the ratio

of the squared error of a statistic class, summed across all statistics in the class, to the sum of the squared statistic values of that class. The procedure was halted once all classes of statistics were imposed with an SNR of 30 dB or higher or when 60 iterations were reached. The procedure was considered to have converged if the average SNR of all statistic classes was 20 dB or higher. Occasionally the synthesis process converged to a local minimum in which it failed to produce a signal matching the statistics of an original sound according to our criterion. This was relatively rare, and such failures of convergence were not used in experiments.

Although the statistics in our model constrain the distribution of the sound signal, we have no explicit probabilistic formulation and as such are not guaranteed to be drawing samples from an explicit distribution. Instead, we qualitatively mimic the effect of sampling by initializing the synthesis with different samples of noise (as in some visual texture synthesis methods) (Heeger and Bergen, 1995; Portilla and Simoncelli, 2000). An explicit probabilistic model could be developed via maximum entropy formulations (Zhu et al., 1997), but sampling from such a distribution is generally computationally prohibitive.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and one table and can be found with this article online at doi:10.1016/j.neuron.2011.06.032.

ACKNOWLEDGMENTS

We thank Dan Ellis for helpful discussions and Mark Bee, Mike Landy, Gary Marcus, and Sam Norman-Haignere for comments on drafts of the manuscript.

Accepted: June 27, 2011
Published: September 7, 2011

REFERENCES

Adelson, E.H., and Bergen, J.R. (1985). Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A *2*, 284–299.

Alvarez, G.A., and Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. Proc. Natl. Acad. Sci. USA *106*, 7345–7350.

Asari, H., Pearlmutter, B.A., and Zador, A.M. (2006). Sparse representations for the cocktail party problem. J. Neurosci. *26*, 7477–7490.

Athineos, M., and Ellis, D. (2003). Sound texture modelling with linear prediction in both time and frequency domains. Proc. ICASSP-03, Hong Kong. 10.1109/ASPAA.2003.1285816.

Attias, H., and Schreiner, C.E. (1998). Coding of naturalistic stimuli by auditory midbrain neurons. In Advances in Neural Information Processing Systems, M.I. Jordan, M.J. Kearns, and S.A. Solla, eds. (Cambridge, MA: MIT Press), pp. 103–109.

Bacon, S.P., and Grantham, D.W. (1989). Modulation masking: effects of modulation frequency, depth, and phase. J. Acoust. Soc. Am. *85*, 2575–2580.

Barbour, D.L., and Wang, X. (2003). Contrast tuning in auditory cortex. Science *299*, 1073–1075.

Baumann, S., Griffiths, T.D., Sun, L., Petkov, C.I., Thiele, A., and Rees, A. (2011). Orthogonal representation of sound dimensions in the primate midbrain. Nat. Neurosci. *14*, 423–425.

Bee, M.A., and Micheyl, C. (2008). The cocktail party problem: what is it? How can it be solved? And why should animal behaviorists study it? J. Comp. Psychol. *122*, 235–251.

Bell, A.J., and Sejnowski, T.J. (1996). Learning the higher-order structure of a natural sound. Network *7*, 261–267.

Crochiere, R.E., Webber, S.A., and Flanagan, J.L. (1976). Digital coding of speech in sub-bands. Bell Syst. Tech. J. *55*, 1069–1085.

Cusack, R., and Carlyon, R.P. (2003). Perceptual asymmetries in audition. J. Exp. Psychol. Hum. Percept. Perform. *29*, 713–725.

Dau, T., Kollmeier, B., and Kohlrausch, A. (1997). Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. *102*, 2892–2905.

de Cheveigne, A. (2004). Pitch perception models. In Pitch, C.J. Plack and A.J. Oxenham, eds. (New York: Springer Verlag).

Dean, I., Harper, N.S., and McAlpine, D. (2005). Neural population coding of sound level adapts to stimulus statistics. Nat. Neurosci. *8*, 1684–1689.

Depireux, D.A., Simon, J.Z., Klein, D.J., and Shamma, S.A. (2001). Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J. Neurophysiol. *85*, 1220–1234.

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. (Regul. Ed.) *11*, 333–341.

Dubnov, S., Bar-Joseph, Z., El-Yaniv, R., Lischinski, D., and Werman, M. (2002). Synthesizing sound textures through wavelet tree learning. IEEE Comput. Graph. Appl. *22*, 38–48.

Elhilali, M., and Shamma, S.A. (2008). A cocktail party with a cortical twist: how cortical mechanisms contribute to sound segregation. J. Acoust. Soc. Am. *124*, 3751–3771.

Field, D.J. (1987). Relations between the statistics of natural images and the response profiles of cortical cells. J. Opt. Soc. Am. A. *4*, 2379–2394.

Garcia-Lazaro, J.A., Ahmed, B., and Schnupp, J.W. (2006). Tuning to natural stimulus dynamics in primary auditory cortex. Curr. Biol. *16*, 264–271.

Gardner, W.G. (1998). Reverberation algorithms. In Applications of Digital Signal Processing to Audio and Acoustics, M. Kahrs and K. Brandenburg, eds. (Norwell, MA: Kluwer Academic Publishers).

Ghitza, O. (2001). On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. J. Acoust. Soc. Am. *110*, 1628–1640.

Glasberg, B.R., and Moore, B.C.J. (1990). Derivation of auditory filter shapes from notched-noise data. Hear. Res. *47*, 103–138.

Gygi, B., Kidd, G.R., and Watson, C.S. (2004). Spectral-temporal factors in the identification of environmental sounds. J. Acoust. Soc. Am. *115*, 1252–1265.

Heeger, D.J., and Bergen, J. (1995). Pyramid-based texture analysis/synthesis. Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, 229–238. 10.1145/218380.218446.

Irino, T., and Patterson, R.D. (1996). Temporal asymmetry in the auditory system. J. Acoust. Soc. Am. *99*, 2316–2331.

Joris, P.X., Schreiner, C.E., and Rees, A. (2004). Neural processing of amplitude-modulated sounds. Physiol. Rev. *84*, 541–577.

Julesz, B. (1962). Visual pattern discrimination. IRE Trans. Inf. Theory *8*, 84–92.

Kayser, C., Petkov, C.I., Lippert, M., and Logothetis, N.K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. Curr. Biol. *15*, 1943–1947.

Kvale, M.N., and Schreiner, C.E. (2004). Short-term adaptation of auditory receptive fields to dynamic stimuli. J. Neurophysiol. *91*, 604–612.

Lesica, N.A., and Grothe, B. (2008). Efficient temporal processing of naturalistic sounds. PLoS One *3*, e1655.

Lorenzi, C., Berthommier, F., and Demany, L. (1999). Discrimination of amplitude-modulation phase spectrum. J. Acoust. Soc. Am. *105*, 2987–2990.

Lorenzi, C., Simpson, M.I.G., Millman, R.E., Griffiths, T.D., Woods, W.P., Rees, A., and Green, G.G.R. (2001). Second-order modulation detection thresholds for pure-tone and narrow-band noise carriers. J. Acoust. Soc. Am. *110*, 2470–2478.

McDermott, J.H. (2009). The cocktail party problem. Curr. Biol. *19*, R1024–R1027.

McDermott, J.H., and Oxenham, A.J. (2008). Spectral completion of partially masked sounds. Proc. Natl. Acad. Sci. USA *105*, 5939–5944.

McDermott, J.H., Oxenham, A.J., and Simoncelli, E.P. (2009). Sound texture synthesis via filter statistics. Proceedings of IEEE Workshop on Applications

of Signal Processing to Audio and Acoustics, 297–300. 10.1109/ASPAA.2009.5346467.

McDermott, J.H., Wrobleski, D., and Oxenham, A.J. (2011). Recovering sound sources from embedded repetition. Proc. Natl. Acad. Sci. USA *108*, 1188–1193.

Mesgarani, N., David, S.V., Fritz, J.B., and Shamma, S.A. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J. Neurophysiol. *102*, 3329–3339.

Miller, L.M., Escabí, M.A., Read, H.L., and Schreiner, C.E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J. Neurophysiol. *87*, 516–527.

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

Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607–609.

Overath, T., Kumar, S., von Kriegstein, K., and Griffiths, T.D. (2008). Encoding of spectral correlation over time in auditory cortex. J. Neurosci. *28*, 13268–13273.

Portilla, J., and Simoncelli, E.P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. Int. J. Comput. Vis. *40*, 49–71.

Rieke, F., Bodnar, D.A., and Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. Proc. Biol. Sci. *262*, 259–265.

Rodríguez, F.A., Chen, C., Read, H.L., and Escabí, M.A. (2010). Neural modulation tuning characteristics scale to efficiently encode natural sound statistics. J. Neurosci. *30*, 15969–15980.

Ruggero, M.A. (1992). Responses to sound of the basilar membrane of the mammalian cochlea. Curr. Opin. Neurobiol. *2*, 449–456.

Saint-Arnaud, N., and Popat, K. (1995). Analysis and synthesis of sound texture. Proceedings of AJCAI Workshop on Computational Auditory Scene Analysis, pp. 293–308.

Schönwiesner, M., and Zatorre, R.J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. USA *106*, 14611–14616.

Schwartz, O., and Simoncelli, E.P. (2001). Natural signal statistics and sensory gain control. Nat. Neurosci. *4*, 819–825.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. Science *270*, 303–304.

Singh, N.C., and Theunissen, F.E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. J. Acoust. Soc. Am. *114*, 3394–3411.

Slaney, M. (1995). Pattern playback in the 90's. In Advances in Neural Information Processing Systems 7, G. Tesauro, D. Touretsky, and T. Leen, eds. (Cambridge, MA: MIT Press).

Smith, E.C., and Lewicki, M.S. (2006). Efficient auditory coding. Nature *439*, 978–982.

Smith, Z.M., Delgutte, B., and Oxenham, A.J. (2002). Chimaeric sounds reveal dichotomies in auditory perception. Nature *416*, 87–90.

Strickland, E.A., and Viemeister, N.F. (1996). Cues for discrimination of envelopes. J. Acoust. Soc. Am. *99*, 3638–3646.

Verron, C., Pallone, G., Aramaki, M., and Kronland-Martinet, R. (2009). Controlling a spatialized environmental sound synthesizer. Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 321–324. 10.1109/ASPAA.2009.5346504.

Warren, R.M., Obusek, C.J., and Ackroff, J.M. (1972). Auditory induction: perceptual synthesis of absent sounds. Science *176*, 1149–1151.

Woolley, S.M., Fremouw, T.E., Hsu, A., and Theunissen, F.E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. Nat. Neurosci. *8*, 1371–1379.

Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. J. Acoust. Soc. Am. *111*, 1832–1846.

Zhu, X.L., and Wyse, L. (2004). Sound texture modeling and time-frequency LPC. Proceedings of Conference on Digital Audio Effects, 345–349.

Zhu, S.C., Wu, Y.N., and Mumford, D.B. (1997). Minimax entropy principle and its applications to texture modeling. Neural Comput. *9*, 1627–1660.

Neuron, Volume *71*
Supplemental Information

Sound Texture Perception via Statistics of The Auditory Periphery: Evidence from Sound Synthesis

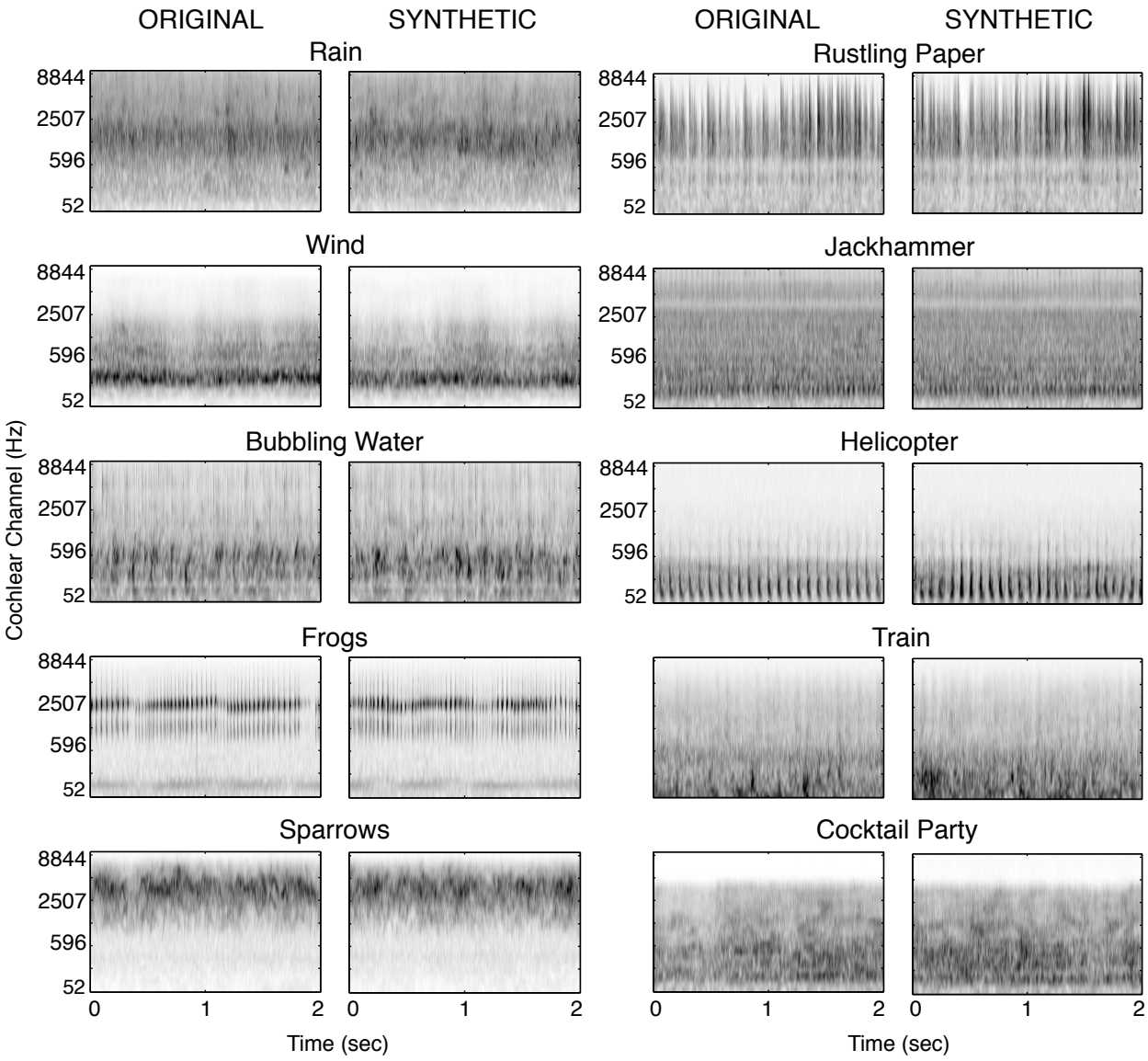
Josh H. McDermott & Eero P. Simoncelli

Supplemental Figures and Tables

Figure S1. Additional examples of synthetic textures
Figure S2. Multiple synthetic texture exemplars generated from the same statistics
Figure S3. Realism of synthesis with filters narrower or broader than those in the cochlea
Figure S4. Realism of synthesis with more cochlear channels or full marginal histograms
Table S1. Synthetic textures ranked by realism ratings
Figure S5. Spectrograms of original and synthetic versions of artificial sounds

Supplemental Experimental Procedures

Figure S6. Stages involved in computing the C2 correlation



Figures S1. Additional examples of synthetic textures. Spectrograms for 10 additional examples of synthetic textures and the real-world sound textures whose statistics they were generated from. Two-second excerpts are shown, to make the rapid temporal structure more visible.

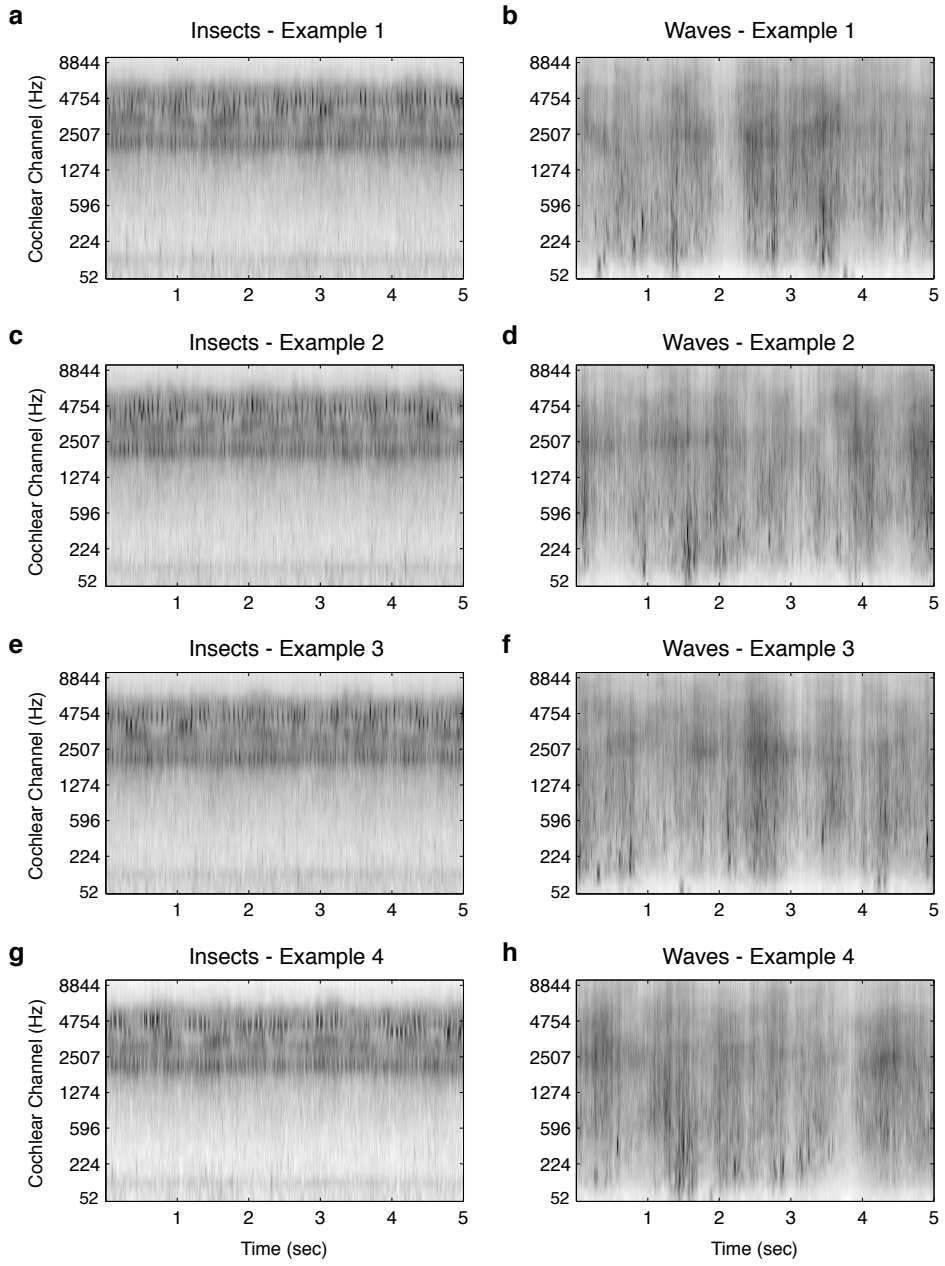


Figure S2. Multiple synthetic texture exemplars generated from the same statistics. Each of the four examples of each sound was generated from a new sample of random noise using the same set of statistics (measured from the same sound recording of swamp insects (a, c, e, g), or seaside waves (b, d, f, h)). Spectrograms of the full 5 second synthetic excerpt are shown to make the slow fluctuations of the waves visible. It is visually apparent that the examples have similar texture qualities, but are nonetheless physically distinct. The texture statistics thus describe a large set of sounds (united by their texture qualities), and the synthesis process generates a different member of the set each time it is run.

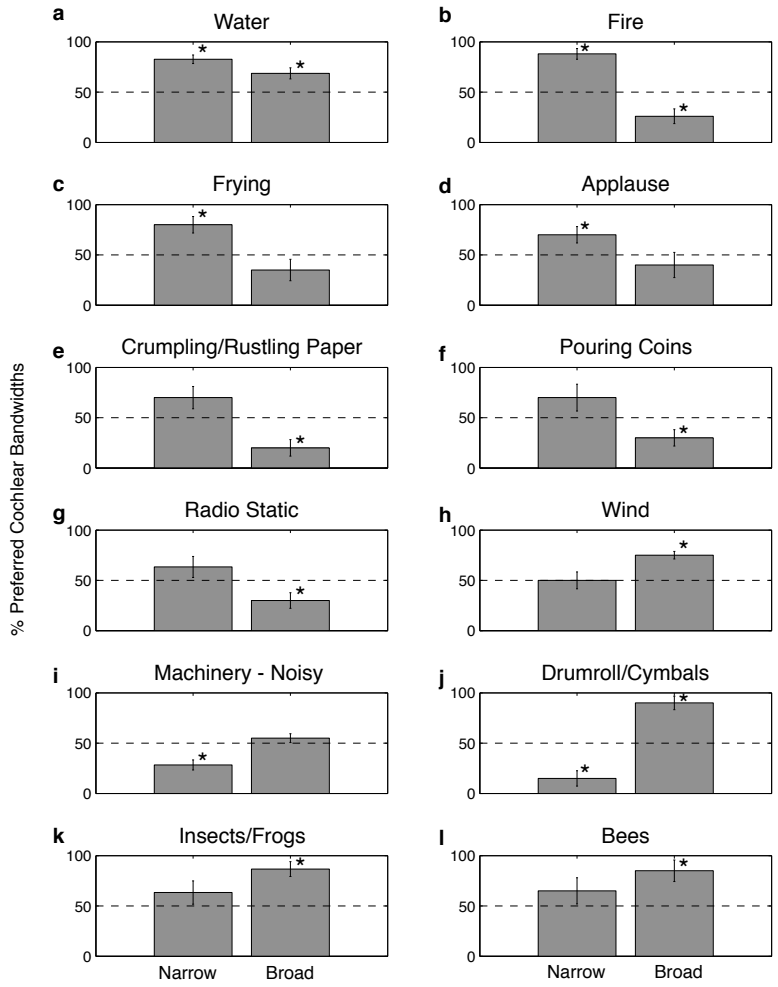


Figure S3. Realism of synthesis with filters narrower or broader than those in the cochlea. Results shown are from conditions 1 and 2 of Experiment 1c. Listeners heard a real-world texture followed by two synthetic versions, and chose which was a more realistic example of the original sound. Both synthetic sounds were generated from "cochlear" marginal moments, either using filters with bandwidths comparable to those in the cochlea, or filters four times narrower (cond. 1) or four times broader (cond. 2). See Supp. Methods for details. The graphs plot the proportion of trials on which the synthesis using cochlear bandwidths was preferred to that using either broader or narrower filters, subdivided according to sound class. Asterisks denote significant differences from chance, uncorrected. Because synthesis with marginal statistics generates sounds with largely independent bandpass events, the filter bandwidths that are preferred provide an indication of the bandwidths of the acoustic generative process. Water is the only sound class for which synthesis with the correct cochlear filter bandwidths was significantly preferred over that with both broader and narrower filters. The bandwidth of water events thus seems to be comparable to the bandwidths of cochlear filters. Other classes of sounds exhibit alternative patterns. Fire, for instance, as well as other sounds with broadband events (frying, applause, rustling paper, pouring coins, radio static) tend to sound better when synthesized with filters broader than those found in the ear. Noise-like sounds (e.g. machinery, cymbals), whose perception is dominated by the shape of the power spectrum, appear to be better synthesized with a larger number of narrower filters, which can better recreate the original spectrum.

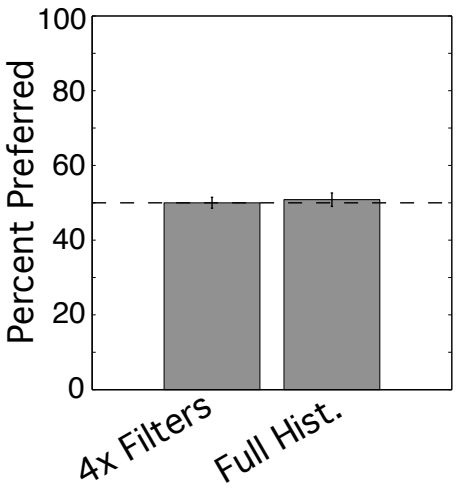


Figure S4. Realism of synthesis with four times as many filters or full marginal histograms. Results shown are from conditions 3 and 4 of Experiment 1c. Listeners heard a real-world texture followed by two synthetic versions, and chose which was a more realistic example of the original sound. One of the synthetic versions was synthesized from our canonical model. The other was synthesized from a model with four times as many filters with the same bandwidth, or from the canonical model, but using the full marginal histogram in addition to the marginal moments. See Supplementary Methods for details. Y-axis plots the percent of trials on which synthesis with the canonical model was preferred. Unlike in Fig. S3, there were not clear differences across sound classes, and thus the results are collapsed across classes to show the overall difference between synthesis conditions. Neither synthesizing from the marginal statistics of four times as many filters nor from the full marginal histogram produced noticeably better synthetic results. This further supports the conclusions of Experiment 1b – that the failure of marginal statistics to capture the sound of non-water sounds reflects the importance of qualitatively different statistics that capture dependencies over time and frequency. Simply adding more marginal constraints (as one might be inclined to, given that the ear contains roughly 3000 hair cells rather than the 30 that we simulate in our model) does not serve to better capture sound structure.

6.57 – Insects in swamp
6.57 – Heavy rain on hard surface
6.53 – Frogs
6.53 – Rain
6.47 – Applause – big room
6.43 – Radio static
6.43 – Stream
6.43 – Jungle rain
6.40 – Air conditioner
6.40 – Stream near small waterfall
6.37 – Frogs
6.37 – Frogs and insects
6.37 – Frying eggs
6.33 – Frogs
6.33 – Wind – blowing
6.33 – Wind – whistling
6.33 – Insects during day in South
6.30 – Radio static
6.30 – Frogs
6.30 – Heavy rain falling and dripping
6.27 – Applause – large crowd
6.27 – River running over shallows
6.27 – Construction site ambience
6.23 – Waterfall
6.20 – Sparrows – large excited group
6.17 – Pneumatic drills
6.17 – Small river
6.17 – Fast running river
6.17 – Rain in woods
6.13 – Water trickling into pool
6.10 – Bathroom sink
6.10 – Water running into sink
6.03 – Frying bacon
6.03 – Rain in ihe woods
6.00 – Fire – forest inferno
5.97 – Birds in forest
5.90 – Linotype
5.90 – Bee swarm
5.90 – Applause
5.90 – Bath being drawn
5.90 – Rustling paper
5.87 – Train speeding down railroad tracks – steam
5.87 – Rattlesnake rattle
5.83 – Fire – burning room
5.83 – Bubbling water
5.83 – Fire – burning room
5.83 – Thunder and rain
5.73 – Fire
5.70 – Wind – moaning
5.70 – Bulldozer
5.70 – Babble
5.70 – Fire
5.70 – Wind – spooky
5.70 – Water lapping gently
5.67 – Shaking coins
5.67 – Helicopter
5.67 – Seagulls
5.63 – Crunching cellophane
5.63 – Sander
5.60 – Radio static
5.60 – Teletype – city room
5.57 – Steam shovel
5.53 – Pigeons cooing
5.50 – Metal lathe
5.47 – Bee swarm
5.47 – Lapping waves
5.43 – Geese cackling
5.40 – Train speeding down railroad tracks – Diesel
5.30 – Lake shore
5.30 – Sanding by hand
5.30 – Blender
5.30 – Teletype
5.30 – Birds in tropical forest
5.27 – Drumroll
5.27 – Surf hitting beach
5.23 – Industrial machinery
5.20 – Crowd noise
5.20 – Rolling coin
5.20 – Ducks quacking
5.20 – WWII bomber plane
5.17 – Applause
5.17 – Idling boat
5.17 – Jackhammer
5.10 – Brushing teeth
5.10 – Horse trotting on cobblestones
5.07 – Scratching beard
5.07 – Printing press
5.07 – Writing with pen on paper
5.00 – Train locomotive – steam engine
5.00 – Helicopter fly by
4.97 – Pouring coins
4.97 – Motorcycle idling
4.97 – Fire
4.93 – Crumpling paper
4.87 – Ship anchor being raised
4.87 – Jingling keys
4.87 – Electric adding machine
4.80 – Horse walking in snow
4.73 – Cymbals shaking
4.70 – Fire – in chimney
4.67 – Tambourine shaking
4.67 – Pouring coins
4.63 – Rhythmic applause
4.63 – Cat lapping milk
4.57 – Seaside waves
4.43 – Rustling paper
4.37 – Horse pulling wagon
4.37 – Vacuum cleaner
4.37 – Horse and carriage
4.30 – Power saw
4.30 – Tire rolling on gravel
4.27 – Horse and buggy
4.27 – Steam engine
4.23 – Cement mixer
4.23 – Power saw
4.23 – Castanets
4.23 – Ox cart
4.20 – Battle explosions
4.17 – Chickens squawking
4.10 – Rubbing cloth
4.03 – Rain beating against window panes
3.97 – Typewriter – IBM electric
3.90 – Lawn mower
3.77 – Gargling
3.77 – Horse gallop on soft ground
3.73 – Applause – foreground clapper
3.67 – Sawing by hand
3.67 – Crumpling paper
3.60 – Wolves howling
3.60 – Fast breathing
3.57 – Dogs
3.40 – Out of breath
3.23 – Windshield wipers
3.20 – Pile driver
3.13 – Silly mouth noise
3.10 – Large diner
3.00 – Filing metal
2.90 – Typewriter – manual
2.83 – Fire alarm bell
2.83 – Knife sharpening
2.83 – Typewriter – old
2.70 – Pile driver
2.70 – Clock ticking
2.67 – Jogging on gravel
2.67 – Castanets
2.57 – Hammering copper
2.47 – Laughter
2.47 – Tapping rhythm
2.37 – Running up stairs
2.27 – Typewriter – IBM selectric
2.17 – Men marching together
2.00 – Tapping on hard surface
1.93 – Railroad crossing
1.90 – Tapping 1–2
1.77 – Wind chimes
1.77 – Corkscrew against desk edge
1.70 – Reverse drum beats – snare
1.70 – Tapping 1–2–3
1.67 – Snare drum beats
1.63 – Walking on gravel
1.60 – Snare rimshot sequence
1.60 – Music – Apache drum break
1.50 – Music – mambo
1.50 – Bongo loop
1.47 – Firecrackers
1.40 – Person speaking French
1.37 – Church bells
1.20 – Person speaking English

Table S1. Synthetic textures ranked by realism ratings. Table displays the complete results of Experiment 4, in which listeners compared the results of our synthesis algorithm to the original sounds from which their statistics were measured. All 168 sounds are ranked by the average realism rating of the resulting synthetic signals. It is apparent that a wide range of natural environmental sounds are well synthesized. The lowest rated sounds provide indications of sound qualities that are not well-captured by such statistics, and that likely implicate more sophisticated acoustic measurements.

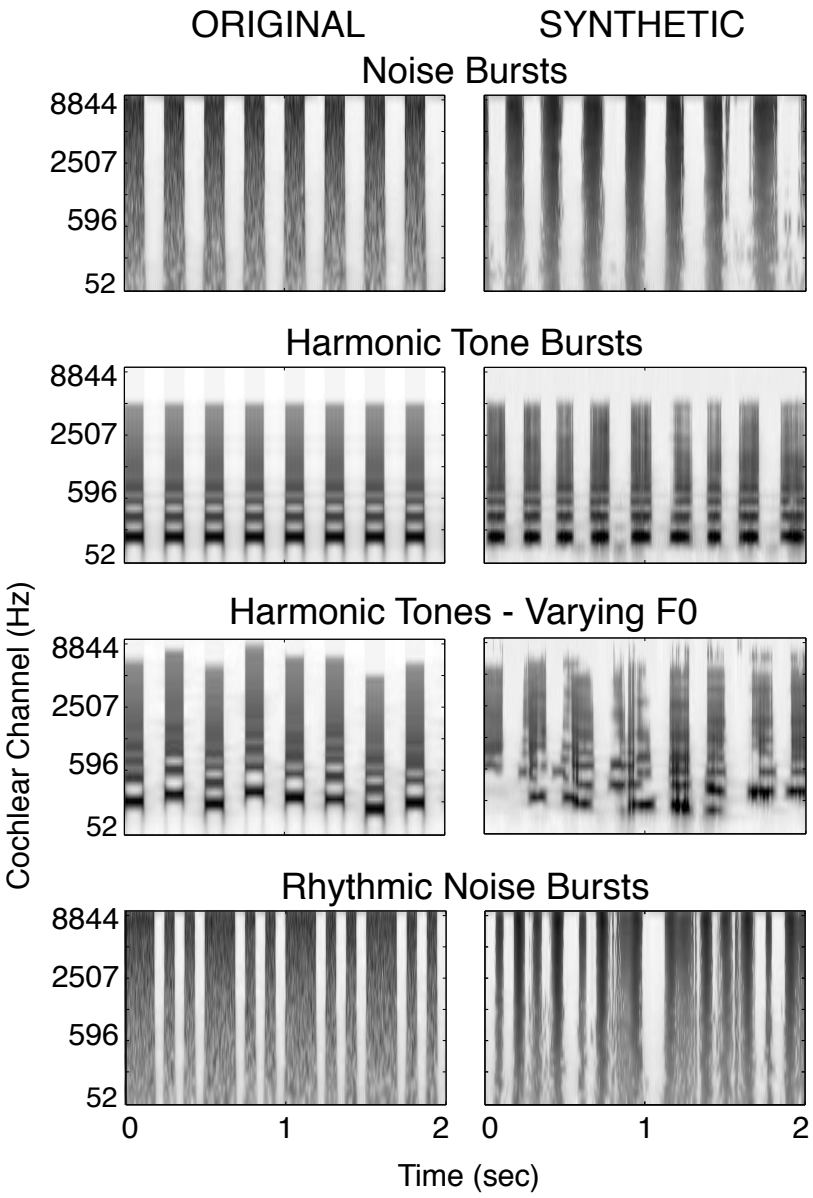


Figure S5. Spectrograms of original and synthetic versions of artificial sounds. Format is as in Fig. S1. Two-second excerpts are shown for clarity. The structure of regularly spaced noise bursts (white noise modulated with a square wave) is largely replicated, as shown in the first row. Regular tone bursts are also largely captured so long as the pitch remains constant, as in the second row. However, when the pitch varies, as in the third row, the harmonic relations are lost, indicating that they are not captured by the model statistics. This result is consistent with the failures documented in Fig. 7 and Table S1 for textures containing pitched sounds. A slightly more complex patterns of rhythmic modulation (bottom row) is also not captured by the model statistics, and is again consistent with the failures of Fig. 7 and Table S1. We note that the synthesis process sometimes failed to completely converge for simple artificial examples, instead getting stuck in local optima in which the statistics were not sufficiently close to the desired values. This rarely happened for real-world sound recordings, but the symmetry and binary nature of some artificial sounds made them more prone to this behavior.

Supplemental Experimental Procedures

Sparsity and cochlear marginal statistics

Because the envelope is positive-valued, its moments are not independent – increasing the variance and kurtosis generally increases the skew as well, as the envelope can be pushed arbitrarily high, but not arbitrarily low. We thus found that these three moments were correlated across sounds, but that all three were nonetheless necessary to replicate the shape of marginal histograms when synthesizing sounds. Moreover, removing any of the three impaired the quality of the synthesis of some sounds (see Fig. 6b).

Modulation band moments

In contrast to the cochlear marginal statistics, which included each of the first four normalized moments, the modulation band marginal statistics were restricted to include only the variance (power). The other moments were omitted because they were either uninformative or redundant with other statistics. Because the modulation bands are computed with bandpass filters, their mean is always zero, and conveys no information about sound content. However, the skewness of the bands generally varied from sound to sound, and we initially thought it could play a role in capturing temporal asymmetry. The reasoning was as follows: Because anti-symmetric bandpass filters may be viewed as computing derivatives (Farid and Simoncelli, 2004), the modulation bands can encompass derivatives of an envelope (at a particular time scale). Derivatives reflect asymmetry - an abrupt onset followed by a more gradual decay, for instance, produces large-magnitude positive derivatives and small-magnitude negative derivatives, yielding a distribution that is positively skewed. The skewness of a signal's derivative thus seemed a promising way to capture temporal asymmetry in sound. However, we found in pilot experiments that its effect on the synthetic results was weak, and that the C2 correlations were more effective. We thus omitted it from the model, keeping only the modulation band variance.

Alternative Statistics

Other statistics that (a priori) seemed plausibly important also failed to produce noticeable perceptual effects. For instance, in pilot studies we examined the effect of multi-band "correlations" based on the expected product of three or more cochlear envelopes. Although these statistics varied across sounds, and although the synthetic and original sound signals often had different values of these statistics if they were not imposed, their inclusion failed to improve the synthesis of any of the sounds for which it was tested (as judged subjectively by the authors). This indicates that not every statistic that exhibits variation across sounds has noticeable perceptual consequences, and underscores the importance of testing the perceptual effect of statistics with the synthesis methodology.

Autocorrelation

Preliminary versions of our texture model were based strictly on subband statistics, and included the autocorrelation of each subband envelope (computed at a set of lags) as a means of capturing temporal structure (McDermott, Oxenham, & Simoncelli, 2009). The present model omitted the autocorrelation in lieu of modulation power statistics (computed from modulation bands not present in our earlier model). Because of the equivalence between the autocorrelation and power spectrum, these two types of statistic capture qualitatively similar information, and if the entire autocorrelation function and modulation power spectrum were used, their effect would be fully

equivalent. In our implementations, however, we used a smaller set of statistics to approximate the full function/spectrum - for the autocorrelation, we used a small subset of all possible lags, and for the modulation spectrum, we used the power in each of a small set of modulation frequency bands. Although not formally equivalent, the two formulations had similar effects on the synthesis of most sounds. Modulation bands were used in the present model both because they are consistent with the known neurobiology of the auditory system, and because they allowed additional acoustic properties to be captured via correlations between bands (C1 and C2).

Detailed explanation of C2 correlation

The C2 correlation has the standard form of a correlation coefficient:

$$C2_{k,mn} = \frac{\sum_t w(t) d_{k,m}(t)^* a_{k,n}(t)}{\sigma_{k,m}\sigma_{k,n}}$$

but is computed with analytic signals (complex-valued):

$a_{k,n}(t) \equiv \tilde{b}_{k,n}(t) + iH\left(\tilde{b}_{k,n}(t)\right)$ and $d_{k,n}(t) = \frac{a_{k,n}^2(t)}{\left\|a_{k,n}(t)\right\|}$ where $\tilde{b}_{k,n}(t)$ is the *n*th modulation band of the *k*th cochlear envelope and H is the Hilbert transform. Because the signals in the correlation are complex-valued, their product has four terms, two real and two imaginary:

$$C2_{k,mn} = \frac{\sum_t w(t)\left[d_{k,m}^R(t)a_{k,n}^R(t) + d_{k,m}^I(t)a_{k,n}^I(t) + id_{k,m}^R(t)a_{k,n}^I(t) - id_{k,m}^I(t)a_{k,n}^R(t)\right]}{\sigma_{k,m}\sigma_{k,n}}$$

where the superscripts *R* and *I* denote real and imaginary parts. Because the real and imaginary parts of each signal are in quadrature phase, the temporal expectation of $d_{k,m}^R(t)a_{k,n}^R(t)$ is approximately equal to that of $d_{k,m}^I(t)a_{k,n}^I(t)$, and the same relation holds for $d_{k,m}^R(t)a_{k,n}^I(t)$ and $-d_{k,m}^I(t)a_{k,n}^R(t)$. The C2 correlation can thus be written as the sum of one real and one imaginary expectation:

$$C2_{k,mn} = 2\frac{\sum_t w(t)d_{k,m}^R(t)a_{k,n}^R(t)}{\sigma_{k,m}\sigma_{k,n}} + 2i\frac{\sum_t w(t)d_{k,m}^R(t)a_{k,n}^I(t)}{\sigma_{k,m}\sigma_{k,n}}$$

This formulation is similar to that of a statistic developed by Portilla and Simoncelli (2000) to capture phase relations between image subbands.

Supplemental Figure 6a shows example subband envelopes for three types of abrupt events that are common in sound: an onset, an offset, and a transient. Plotted below (Fig. S6b) are two modulation bands of each envelope, tuned to frequencies an octave apart. The three types of events are characterized by alignment of the bands in amplitude and in phase. The amplitude alignment is common to all three event types. The phase alignment, however, distinguishes the three event types, because the bands become aligned at different phase values (evident in Fig. S6b as well as in the phase angles, shown in Fig. S6c).

It would seem natural to capture this phase alignment by computing a correlation between bands. However, because the bands oscillate at different rates, the phase alignment is momentary – the two bands align and then move away from the point of alignment at different rates (evident in the different slopes of the phase plots in Fig. S6c).

The phase alignment can be transformed into a constant phase difference by doubling the frequency of the lower frequency band. This is accomplished by squaring the analytic version of the band (doubling its frequency and squaring its magnitude), and then dividing by the magnitude to preserve the frequency doubling but retain the original magnitude: $d_{k,n}(t) = \frac{a_{k,n}^2(t)}{\left\| a_{k,n}(t) \right\|}$

Fig. S6d plots the original band, the magnitude of its analytic signal, and the real part of the frequency-doubled analytic signal. It is apparent that the magnitude is preserved but that the new signal oscillates at twice the rate.

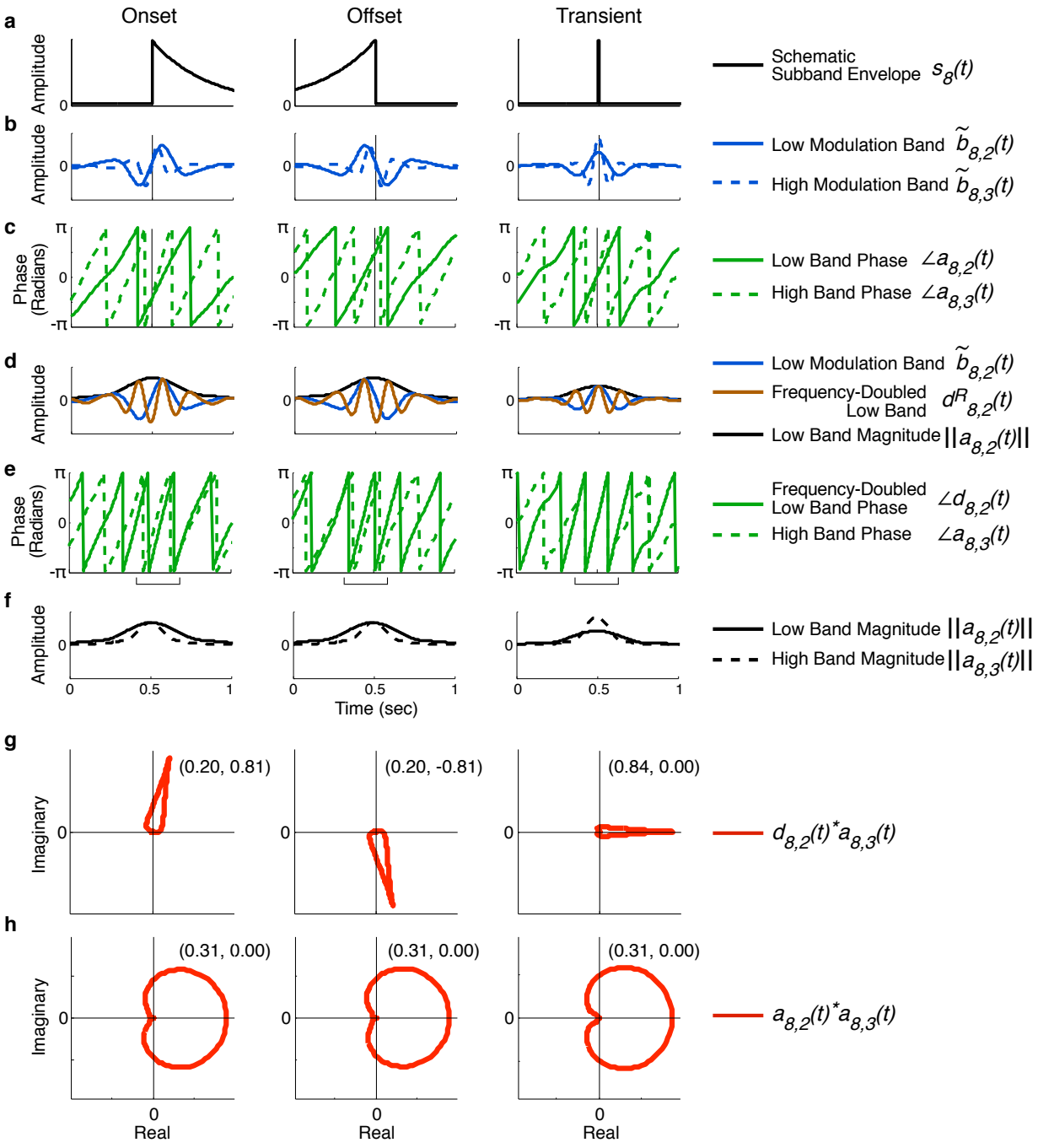
Doubling the frequency of the low band alters its phase at the point of alignment, but because the two bands are an octave apart, the phase of the frequency-doubled low band now advances at the same rate as the high band. This produces a constant phase offset in the vicinity of the original event.

Fig. S6e illustrates this relationship, plotting the phase of the original high frequency band along with that of the frequency-doubled low frequency band. Note that there is now an extended region in which the phase offset between the two signals is relatively constant, and that the offset is different for each of the three event types – a positive step, a negative step, and a brief pulse, respectively. The constant phase offset occurs in the region where the amplitude is high (plotted in Fig. S6f for each band). The phase offset can be made explicit through the product of the two complex signals: $d_{k,m}(t)^* a_{k,n}(t)$, which multiplies the amplitudes and subtracts the phases. When this complex product is plotted in polar coordinates (Fig. S6g), it is apparent that the high amplitudes occur at particular phase values that are different for each of the three event types. The time-average of this complex product thus yields different values in the three cases. When normalized by the band variances, this time-averaged complex product is the C2 correlation: $C2_{k,mn} = \frac{\sum_t w(t) d_{k,m}(t)^* a_{k,n}(t)}{\sigma_{k,m}\sigma_{k,n}}$, the real and imaginary components of which are shown on the plots in parentheses.

For comparison, Fig. S6h plots the same complex product but without the frequency doubling: $a_{k,m}(t)^* a_{k,n}(t)$. Note that the patterns are now similar in the three cases. A correlation of the bands without doubling: $\frac{\sum_t w(t) a_{k,m}(t)^* a_{k,n}(t)}{\sigma_{k,m}\sigma_{k,n}}$ thus yields the same values in each case (again shown on the plots). A standard correlation thus indicates phase alignment, but not the phase

value at the point of alignment that is needed to distinguish event types. Our C2 correlation, in contrast, reveals the structure of interest.

Figure S6. The stages involved in computing the C2 correlation (on next page). The stages are illustrated for cochlear channel 8 and modulation bands 2 and 3. (a) Schematic subband envelopes for three event types: an onset, an offset, and a transient. (b) Modulation bands 2 and 3 (tuned to frequencies an octave apart). (c) The phase angles of each band. The location of the event is indicated by the thin black vertical line. (d) The frequency-doubled low band, plotted on top of the original (undoubled) band, and their magnitude. (e) The phase angles of the frequency-doubled low frequency band, and the original high frequency band. The black bracket below the graph indicates the region in which there is a constant phase offset between the bands. (f) Magnitude of each band. (g) Polar plot of the complex product of the frequency-doubled low band and the original high band. The C2 correlation (the vector average of this product, normalized by the standard deviations of the two bands) is shown in parentheses on each plot. (h) The same product computed without frequency doubling, with the corresponding correlation shown in parentheses on each plot.


a
Onset
Offset
Transient
Amplitude
Schematic Subband Envelope $s_8(t)$
b
Low Modulation Band $\tilde{b}_{8,2}(t)$
High Modulation Band $\tilde{b}_{8,3}(t)$
c
Phase (Radians)
Low Band Phase $\angle a_{8,2}(t)$
High Band Phase $\angle a_{8,3}(t)$
d
Frequency-Doubled Low Band $d^R_{8,2}(t)$
Low Band Magnitude $||a_{8,2}(t)||$
e
Frequency-Doubled Low Band Phase $\angle d_{8,2}(t)$
f
High Band Magnitude $||a_{8,3}(t)||$
Time (sec)
g
Imaginary
(0.20, 0.81)
(0.20, -0.81)
(0.84, 0.00)
$d_{8,2}(t)^* a_{8,3}(t)$
h
(0.31, 0.00)
$a_{8,2}(t)^* a_{8,3}(t)$
Real

Sounds

Our set of natural sound textures was a mix of in-house recordings made by the authors, files downloaded from online sound repositories, and excerpts from sound effects CDs. The main criterion for inclusion in the set was that the sounds be approximately stationary (as judged subjectively by the authors), because the algorithm was not expected to successfully synthesize non-stationary sounds. See Supp. Table 1 for the full list of sounds. We used 7 sec segments to measure statistics. All sounds were resampled with a sampling rate of 20 kHz and normalized to a fixed rms amplitude.

Psychophysical Experiments

Subjects performed experiments seated in a sound-attenuating booth (Gretchken Industries). Sounds were presented diotically at 70 dB SPL over Sennheiser HD580 headphones, via a LynxStudio Lynx22 24-bit D/A converter with a sampling rate of 48 kHz. Sounds were synthesized to be 5 sec in duration. The middle 4 sec was excerpted for use in experiments, with 10 ms half-Hanning windows applied to each end. The only exception to this was condition 7 of Experiment 1b, in which 15 sec sounds were synthesized, from which the middle 4 seconds was excerpted for use in the experiment. Multiple synthetic versions were generated for all experimental conditions, (two per condition for Experiments 1-3, and three for Experiment 4), one of which was randomly chosen on each trial of an experiment (except for Experiment 4, in which listeners were presented with all three versions on separate trials). All participants were non-expert listeners and were naïve as to the purpose of the experiments.

Experiment 1a: Identification

On each trial, participants heard a 4 sec excerpt of either an original sound recording, or a synthetic signal with some subset of our model's statistics matched to those of an original recording. They then selected a name for the sound from a list of five, by clicking a mouse. The sounds were drawn from a subset of 96 of the 168 sounds in our set, the sound of which we thought likely to be familiar to undergraduate subjects. These sounds were organized into groups whose sounds we thought were likely to be confusable (e.g. rain and river, WWII bomber plane and construction noise), and the incorrect choices on a trial were constrained not to be drawn from the sound's confusion group. All sounds were used once per condition, for a total of 864 trials completed in pseudo-random order. Ten subjects participated (9 female), averaging 22.2 years of age. Subjects in Experiments 1a and 1b were not given practice trials, and had not participated in any of our other experiments, such that they had never heard any of the sounds or their synthetic versions prior to starting the experiment.

Experiment 1b: Identification, Part 2

The procedure for Experiment 1b was identical to that of Experiment 1a. Ten subjects participated (8 female) averaged 20.2 years of age.

Experiments 1b and 1c utilized alternative models to test various hypotheses of interest. Two conditions featured sounds synthesized using a model with broader or narrower filters than those in our canonical model (whose filters were approximately matched to those of the human

auditory system). In these cases we used 7 and 120 filters, respectively, covering the same frequency range, again equally spaced on an ERB scale, with adjacent filters overlapping by 50%, and with lowpass and highpass filters on each end of the spectrum to assure perfect tiling (and thus invertibility). Another condition used a model with the same filter bandwidths as the canonical model, but with neighboring filters that overlapped by 87.5%, producing four times as many filters over the same spectral range. These filter banks also had lowpass and highpass filters on the ends to produce perfect tiling. We also included a condition in both experiments in which our canonical texture model was supplemented with marginal histogram matching, using 128 bins per histogram (Heeger and Bergen, 1995).

Experiment 1c: Realism of Synthesis with Alternative Marginal Statistics

Experiment 1c extended the identification results of Experiment 1b by comparing the realism of sounds synthesized with different kinds of marginal statistics – those measured from our canonical model, or from alternative models, either with different filters or with a more comprehensive description of the filter marginal distributions. The results are shown in Figs. S3 and S4.

The procedure was identical to that of Experiments 2 and 3 (described in full below). On each trial, participants heard an excerpt of an original recording followed by two synthetic excerpts, and selected the synthetic example that sounded like a more realistic example of the original. All synthetic excerpts were synthesized by imposing the "cochlear" marginal statistics of the original recording. One of the synthetic examples was always generated using biologically faithful cochlear bandwidths and the four marginal moments of our canonical model. The other synthetic example was generated using filters four times narrower (condition 1), filters four times broader (condition 2), four times as many filters with the same bandwidth (condition 3), or the filters of the canonical model, but using the full marginal histogram in addition to the marginal moments (condition 4).

Experiment 1c used a smaller subset of 48 sound recordings that were subjectively judged to lack strong temporal structure (because the marginal statistics did not greatly constrain temporal structure, and we wanted to avoid large numbers of trials where both synthetic examples bore little resemblance to the original sound). All sounds were used in all conditions, yielding 192 trials, completed in random order. For the analysis of Fig. S3, sounds were grouped into 12 classes, each containing at least two sounds. Ten subjects participated (8 female), averaging 23.7 years of age.

Experiment 2a: Omission

On each trial, participants heard a 4 sec excerpt of an original sound recording followed by two synthetic versions, with 400 ms of silence between sounds. One synthetic version was synthesized with the full set of statistics, and the other with all but one class of statistics. In the marginal condition, the envelope variance, skew, and kurtosis were omitted (the mean was left in to ensure the correct spectrum). The order in which the two versions were presented was randomized. Listeners selected which version sounded like a more realistic example of the original. Ninety-eight original sounds (and their synthetic versions) were used. Each sound was

presented once per condition, for a total of 490 trials, completed in random order. All subjects were given 20 practice trials prior to starting the experiment. Ten subjects participated (8 female), averaging 24.4 years of age.

The sound set was slightly different from that used in Experiment 1 – some of the multiple versions of certain sound classes were removed to reduce redundancy, replaced by sounds that were omitted from Experiment 1 for reasons of their unfamiliarity (e.g. a "teletype", which most undergraduates are unfamiliar with, but for which original and synthetic versions are readily compared). The asymmetric sounds included in the analysis of C2 correlation omission were: Typewriter – manual, Typewriter – IBM electric, Drumroll, Battle explosions, Tapping on hard surface, Hammering copper, Snare drum beats, Bongo loop, Reverse drum beats – snare, Teletype, Firecrackers, and Rhythmic applause. 30000 other randomly chosen subsets of sounds were evaluated for comparison.

Experiment 2b :Marginal Variants

The trial format and set of sounds was identical to that of Experiment 2a. In every condition, one of the two synthetic versions was synthesized with the full set of statistics measured in the original sound. In condition 1, the other synthetic sound was given the envelope variance, skew and kurtosis of pink noise (measured from a 30 sec excerpt), with the other statistics taken from the original recording, including the envelope mean, which ensured that the spectrum was faithful to the original. The synthesis process succeeded in synthesizing signals with the desired statistics despite the artificial combination (as verified with the same SNR measurements used in other experimental stimuli). In condition 2, the marginal moments were omitted from synthesis but the other statistics were set to the values of the original sound (this conditions was equivalent to condition 1 of Experiment 2a, but was repeated because the subjects in the two experiments were different). In condition 3, only the skew and kurtosis were omitted from synthesis. Nine female subjects participated, averaging 24.1 years of age.

Experiment 3: Nonbiological Models

The format of this experiment was identical to that of Experiment 2a&b, and the same sounds were used. The participant group had not participated in the other experiments, to avoid the possibility that participants might have learned the sound of our original model in previous experiments. Eight subjects participated (5 female), averaging 25 years of age.

Linearly-spaced filters were substituted for the acoustic (cochlear) and modulation filterbanks in some of the conditions. The linear acoustic filterbank had the same number of filters as that in the original model, and was identically generated except that the frequency responses were half-cosines on a linear scale rather than an ERB scale, with a fixed bandwidth of 321.9 Hz. The linearly-spaced filters thus tiled the spectrum as did the ERB filter bank, and produced the same number of statistical measurements, but divided up the spectrum differently.

The linear modulation filterbank also had 20 filters, with peak frequencies ranging from .5 to 200 Hz in 10.5 Hz intervals. The frequency responses were half-cosines with a fixed bandwidth of

17.04 Hz, which produced the same degree of overlap between the passbands (defined by the 3dB-down points) of adjacent filters (38.4%) as was present in the constant Q filterbank (averaging the overlap on the low and high end of a filter). The two lowest filters had a slightly different frequency response to avoid including DC – they increased with a cosine ramp from 0 Hz to their peak frequency. The highest filter cut off at the peak of its frequency response; i.e. it was a quarter-cosine (this was to ensure that all modulation frequencies were represented in the bank).

Because the C2 correlations could only be computed with octave-spaced filters, it was not possible to alter the filter bank used to measure and impose them, and we omitted them from the conditions using a linear modulation filterbank (as well as in the comparison stimuli generated with the biologically plausible model). However, the C1 correlations presented no such limitation, and for them we used a linear filter bank that tiled the spectrum and had comparable overlap to the octave-spaced modulation filter bank in our standard model. The peak frequencies ranged from 3.1 to 167.2 Hz in steps of 32.8 Hz (half-cosine frequency responses with bandwidths of 32.8 Hz, except for the lowest frequency filter, which ramped from 0 to its peak frequency, again to avoid including DC).

Note that the superior realism we observed for the biological texture model could not be explained simply by a difference in how well the biological and nonbiological statistics were imposed. SNRs were comparable between conditions. Synthesis with the nonbiological model averaged 36.26, 45.31, 33.64, and 45.06 dB (conditions 1-4), compared to 35.75 (condition 1) and 38.01 dB (conditions 2-4; no C2 correlations) for the original syntheses used as a comparison.

The choice of which cochlear correlations (i.e., which offsets) to include in the model was informally optimized in pilot tests using the biological model. It might be argued that different choices would be optimal for the nonbiological model with linearly spaced filters, and that the chosen offsets, even if themselves imposed faithfully, would thus be less likely to instantiate the full correlation structure between channels. To address this possibility, we checked the fidelity with which the full correlation matrix was imposed for both models (in dB SNR). There was no significant difference (paired t-test, $p=.06$), and if anything, the SNR for the full correlation matrix was slightly higher for the nonbiological model with the linearly spaced filter bank than for our canonical biologically plausible model (18.04 dB vs. 18.60 dB, SE=.70 and .75). We performed the same sort of analysis for the C1 correlations, and obtained a similar result: slightly higher SNRs for the nonbiological model (8.75 dB vs. 10.36 dB, SE=.55 and .63; the lower SNRs here are due to the fact that only two offsets were imposed for this statistic). Although we cannot exclude the possibility that some alternative non-biological model would produce better synthetic results, the differences in synthesis quality we observed do not appear to be due to the choices that were made about the number of statistics to include.

Experiment 4: Realism Ratings

On each trial, participants heard a 4 sec excerpt of an original sound recording followed by a synthetic version of the original, with 400 ms of silence between sounds. The synthetic version was synthesized with the full set of statistics. Participants were instructed to judge the extent to which the synthetic version sounded like another example of the original sound. They selected a rating on a scale of 1-7 by clicking a mouse, with 7 indicating the highest degree of realism and 1 the lowest. The full set of 168 original sounds (and their synthetic versions) was used. The experiment cycled through the set of sounds three times, each time in a different random order and with a different synthetic exemplar. Ten subjects participated (7 female), averaging 20 years of age.

Supplemental References

Farid, H., and Simoncelli, E.P. (2004). Differentiation of multi-dimensional signals. IEEE Transactions on Image Processing *13*, 496-508.

Heeger, D.J., and Bergen, J. (1995). Pyramid-based texture analysis/synthesis. Computer Graphics (ACM SIGGRAPH Proceedings), 229-238.

McDermott, J.H., Oxenham, A.J., and Simoncelli, E.P. (2009). Sound texture synthesis via filter statistics. In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (New Paltz, New York), pp. 297-300.

Portilla, J., and Simoncelli, E.P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. International Journal of Computer Vision *40*, 49-71.