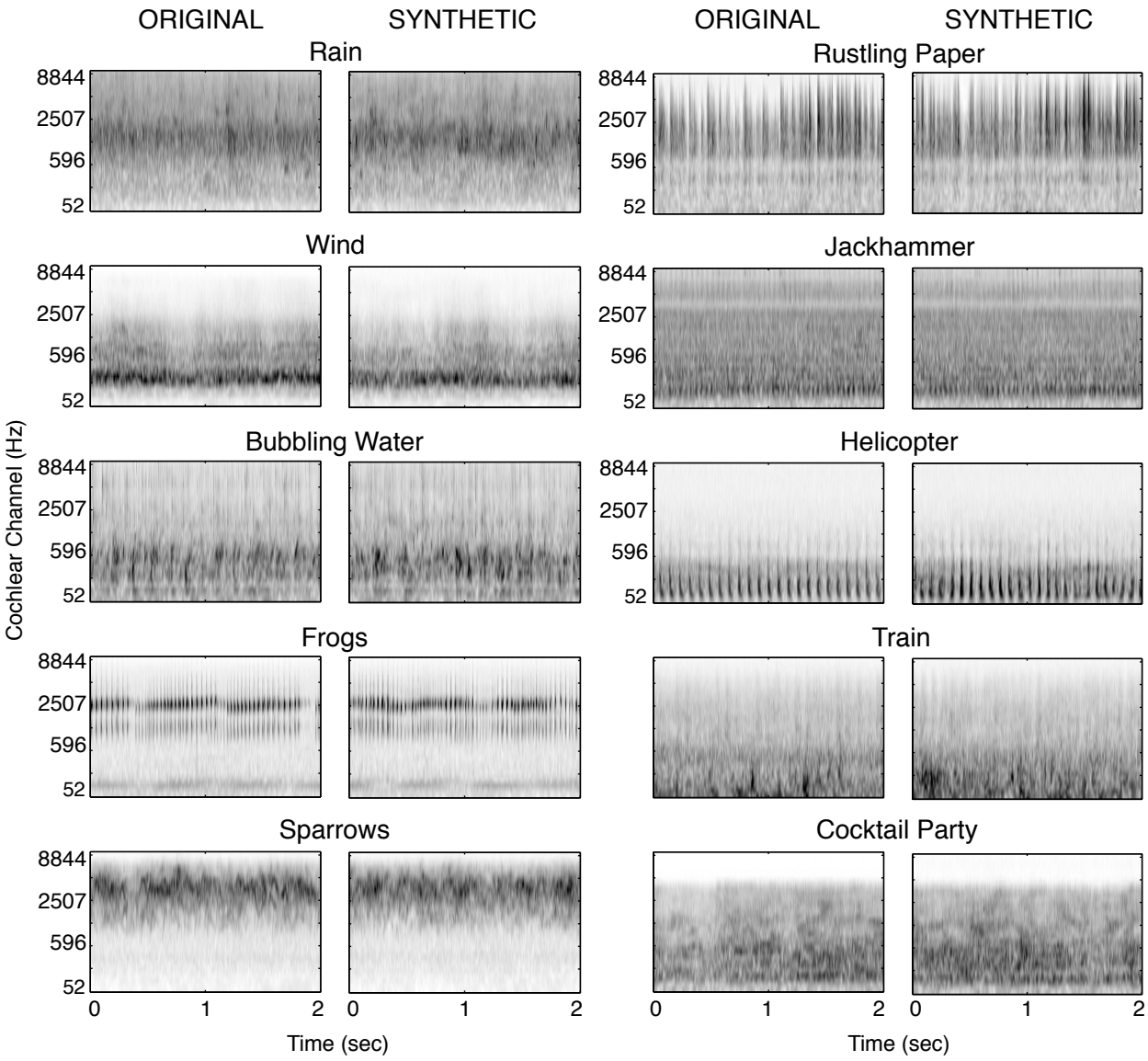
Neuron, Volume *71*
Supplemental Information

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

Josh H. McDermott & Eero P. Simoncelli

Supplemental Figures and Tables

Supplemental Experimental Procedures



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.

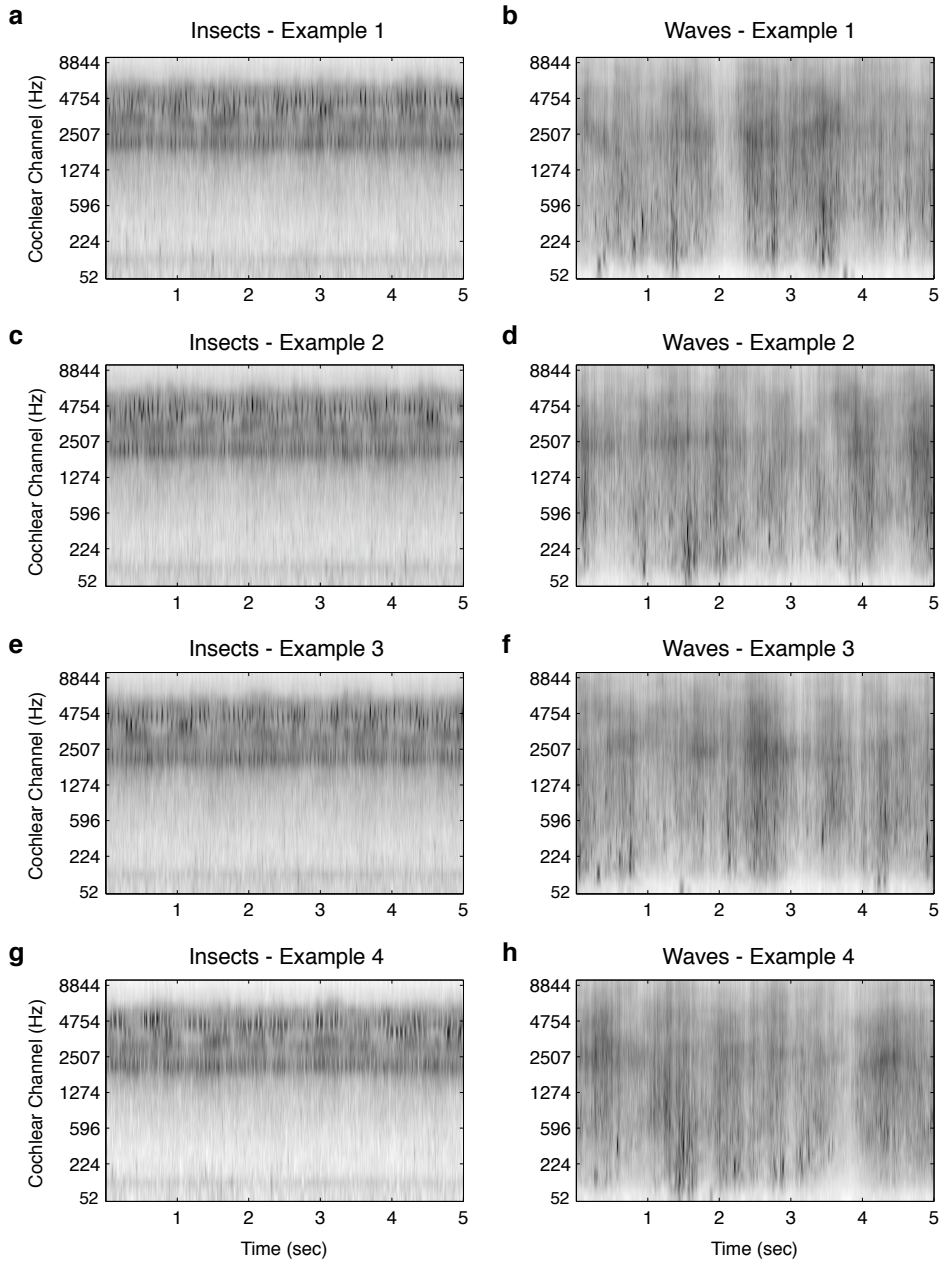


Figure S2. Multiple synthetic texture exemplars generated from the same statistics. Each of the four examples of each sound was generated from a new sample of random noise using the same set of statistics (measured from the same sound recording of swamp insects (a, c, e, g), or seaside waves (b, d, f, 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.

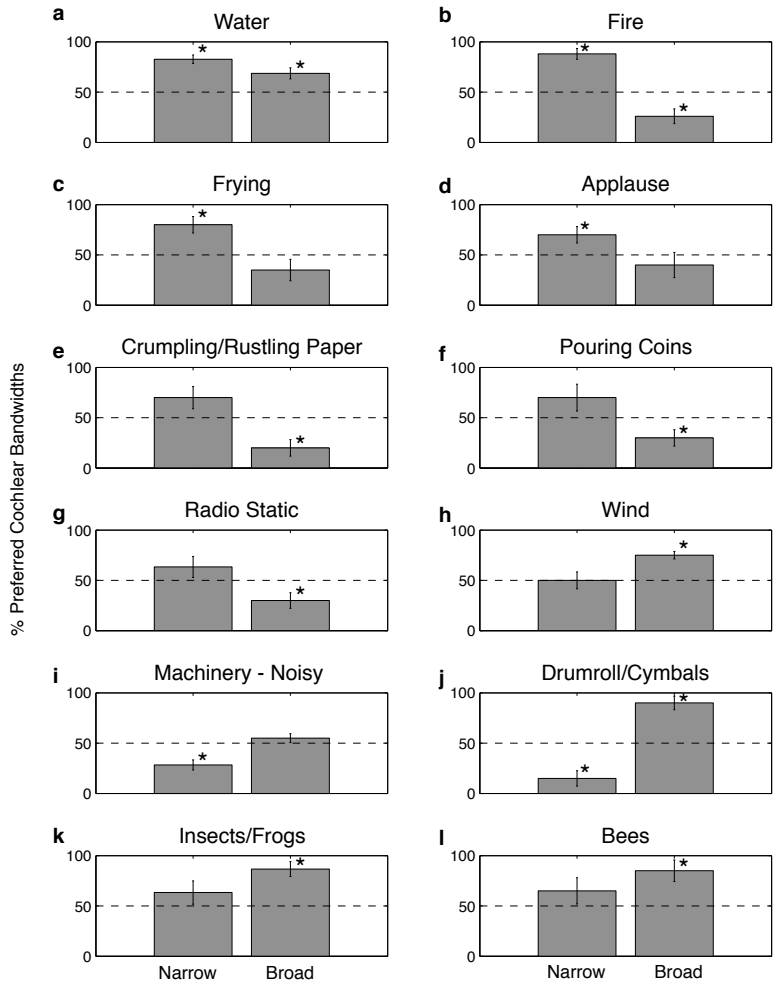


Figure S3. Realism of synthesis with filters narrower or broader than those in the cochlea. Results shown are from conditions 1 and 2 of Experiment 1c. Listeners heard a real-world texture followed by two synthetic versions, and chose which was a more realistic example of the original sound. Both synthetic 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.

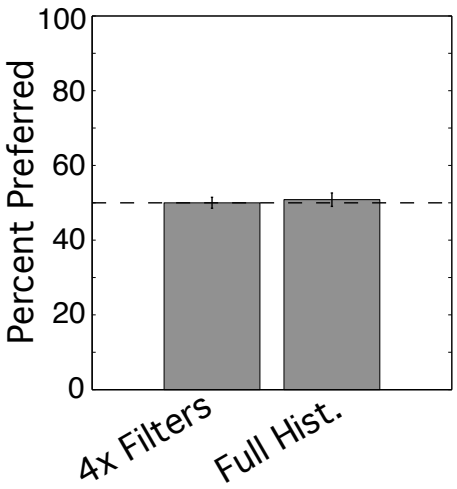


Figure S4. Realism of synthesis with four times as many filters or full marginal histograms. Results shown are from conditions 3 and 4 of Experiment 1c. Listeners heard a real-world texture followed by two synthetic versions, and chose which was a more realistic example of the original sound. One of 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.

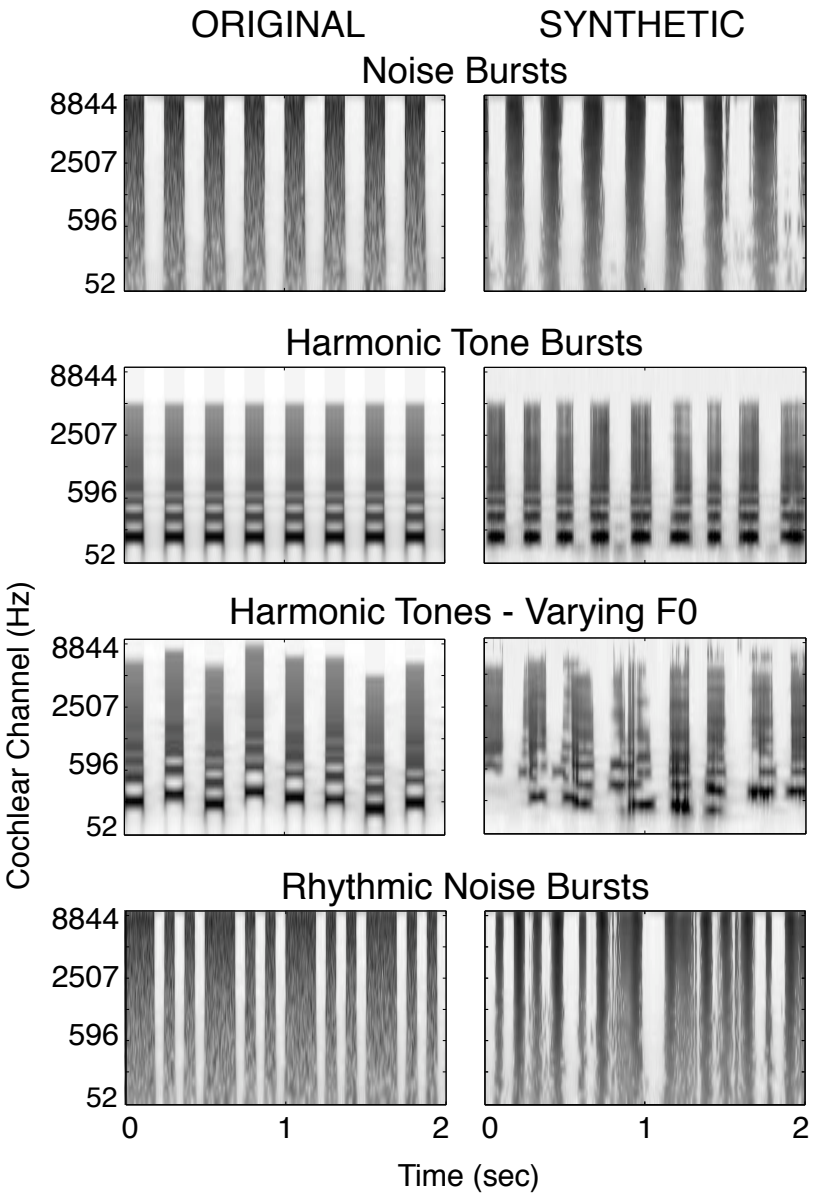


Figure S5. Spectrograms of original and synthetic versions of artificial sounds. Format is as in Fig. S1. Two-second excerpts are shown for clarity. The structure of regularly spaced noise bursts (white noise modulated with a square wave) is largely replicated, as shown in the first row. Regular tone 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.

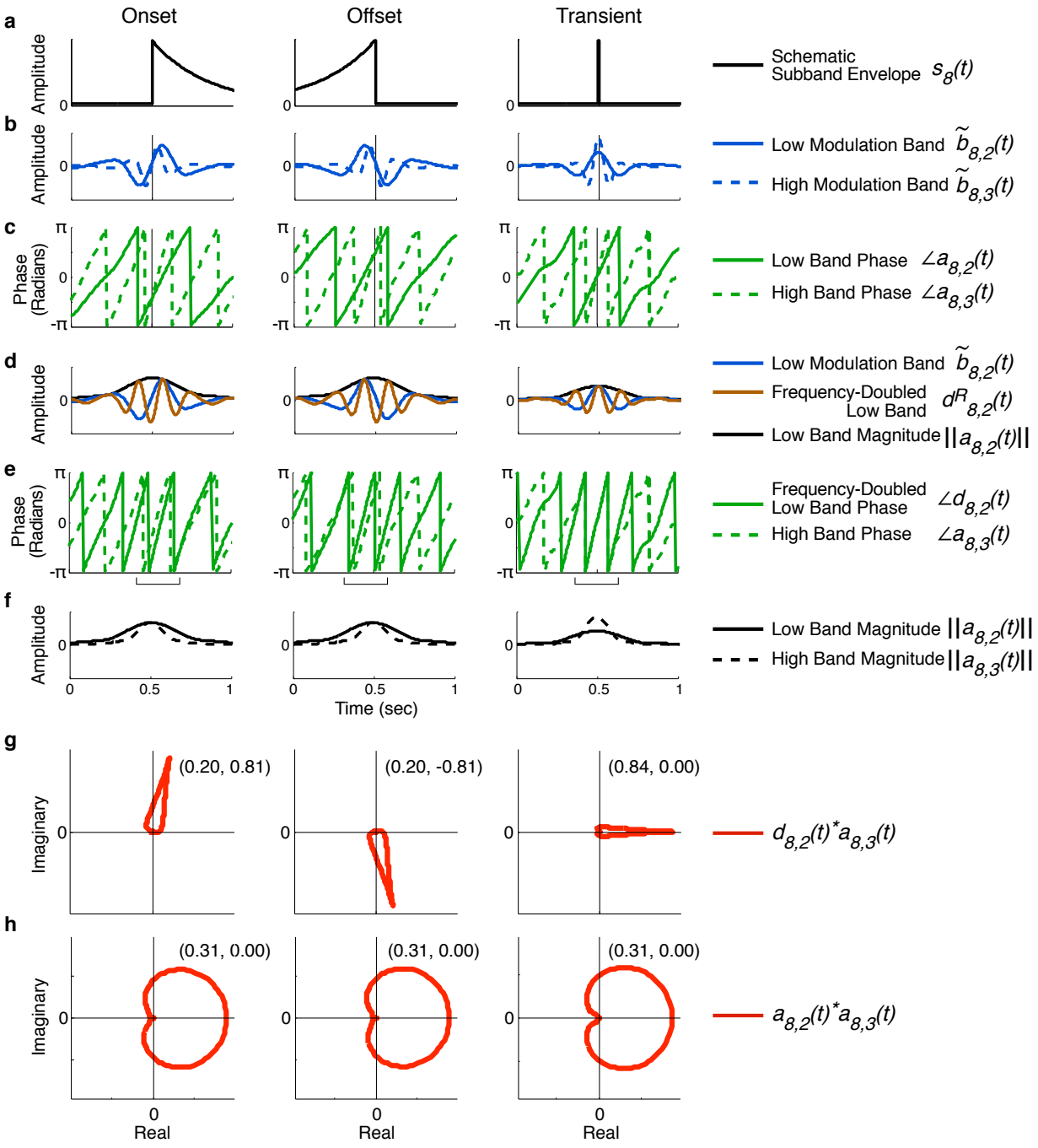
Doubling the frequency of the low band alters its phase at the point of alignment, but because the two bands are an octave apart, the phase of the frequency-doubled low band now advances at the same rate as the high band. This produces a constant phase offset in the vicinity of the original event.

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.