SOUND TEXTURE SYNTHESIS VIA FILTER STATISTICS

Josh H. McDermott[1]*, Andrew J. Oxenham* [2]*, and Eero P. Simoncelli*[1]

[1] Center for Neural Science, New York University, USA
[2] Department of Psychology, University of Minnesota
Email: jhm@cns.nyu.edu

ABSTRACT

Many natural sounds, such as those produced by rainstorms, fires, or insects at night, consist of large numbers of rapidly occurring acoustic events. We hypothesize that humans encode these "sound textures" with statistical measurements that capture their constituent features and the relationship between them. We explored this hypothesis using a synthesis algorithm that measures statistics in a real sound and imposes them on a sample of noise. Simply matching the marginal statistics (variance, kurtosis) of individual frequency subbands was generally necessary, but insufficient, to yield good results. Imposing various pairwise envelope statistics (correlations between bands, and autocorrelations within each band) greatly improved the results, frequently producing synthetic textures that sounded natural and that listeners could reliably recognize. The results suggest that such statistical representations could underlie sound texture perception, and that the auditory system may use fairly simple statistics to recognize many natural sound textures.

Index Terms— texture, statistics, synthesis, envelope, correlations

1. INTRODUCTION

One approach to understanding the representation of natural sensory stimuli in the brain is to develop methods for synthesizing such stimuli. Successful synthesis implies that the perceptually relevant information is captured by the synthesis process, whereas failure indicates that the process is missing something important. Synthesis thus provides a strong test of a perceptual model [1].

We used synthesis to study the perception of sound textures – signals that result from multiple, rapidly occurring acoustic events whose temporal distribution is roughly stationary. These are analogous to visual textures, which have been studied for decades [2]. Sound textures are common in natural environments, but have been neglected in hearing science, though there has been some interest in the computational audio community [3, 4, 5, 6]. Their temporal homogeneity suggests they might be particularly amenable to statistical modeling.

Natural sounds are known to exhibit statistical properties that are distinct from those of noise [7, 8, 9, 10], including kurtotic amplitude histograms and long-term amplitude correlations. We tested the perceptual significance of these and other statistical properties by synthesizing stimuli that shared the statistics of different natural sound textures, and assessing whether they sounded like the real textures they were supposed to resemble. Our interest was to explore the representation of sound textures in the brain, so we focused on statistics of representations inspired by the peripheral auditory system.

2. SOUND STATISTICS

Sounds were analyzed using a bank of bandpass filters. We used filters with bandwidths and spacing that are similar to what is found in the ear (30 filters, equally spaced on an ERBN scale [11], center frequencies from 20 Hz to 14 kHz). Adjacent filters overlapped by 50%, and had frequency responses that were a half cycle of a cosine function. The summed frequency response of such a filter bank applied twice is flat; the filter bank can thus be applied repeatedly without altering the frequency content of the signal.

We obtained a large set of natural sound textures from commercially available CDs, and computed statistics from their subband representations. We examined three classes of statistics: the moments of the marginal distribution of the amplitude of each subband, the correlations between the Hilbert envelopes of neighboring subbands at the same point in time, and the autocorrelation of the envelope of each subband. Envelope correlations were computed on the log of the Hilbert envelope, to retain sensitivity to low amplitude events. Direct evidence for neural selectivity to such statistics is scant [12], but it seems at least plausible that the statistics we measured (generally, temporal averages of nonlinear functions of the filter responses) could be computed with simple neural circuitry.

Natural sound textures contain interesting statistical structure. Figure 1a compares an example subband histogram for a recording of rain to that of Gaussian noise with the same subband variance. They clearly differ in shape: the distribution for rain has long tails, presumably reflecting the discrete raindrop events within the sound, which yield high amplitude filter responses much more often than occurs in Gaussian noise. Long-tailed distributions of subband coefficients have been previously observed in natural images and sounds [7, 8, 9, 10], and can be quantified by the subband kurtosis. Figure 1b shows the average kurtosis of the five subbands with highest variance for each of an assortment of sounds. The red bar indicates the kurtosis of Gaussian noise (always equal to 3), and the others correspond to various natural sound textures (rain, wind, fire etc.). The kurtosis varies from sound to sound, but is always greater than that of the Gaussian noise.

Structure is also apparent in spectral and temporal correlations. The sound texture produced by fire contains crackles, pops, and other broadband features. These are visible as vertical streaks in the spectrogram (Fig. 2a), and result in non-zero correlations between neighboring subband envelopes (Fig. 2b). Temporal structure is also present in many sound textures, such as that produced by insects at night (Fig. 3a), the calls of which occur with temporal regularity. This periodicity is apparent in the subband envelope autocorrelation function (Fig. 3b). Examples such as these raise the possibility that when we recognize sound textures, we may be recognizing their statistical properties.

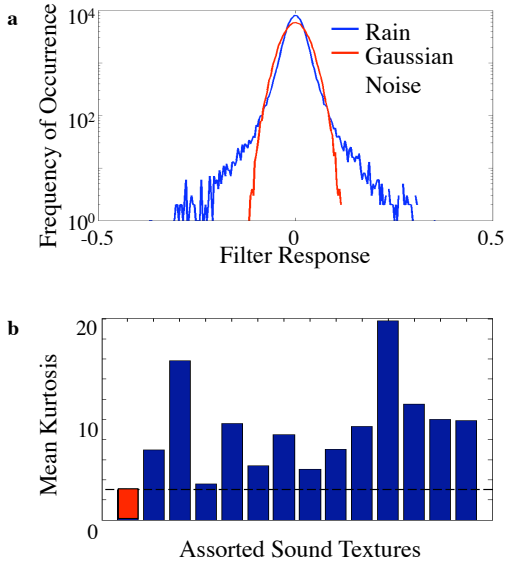

Figure 1: Marginal subband statistics of sound textures. a) Comparison of response histograms of a bandpass filter (4-5 kHz) applied to a recording of rain, and to Gaussian noise with the same subband variance. b) Subband kurtosis of Gaussian noise (in red), and an assortment of natural sound textures. Bars represent average kurtosis of the five subbands with highest variance.

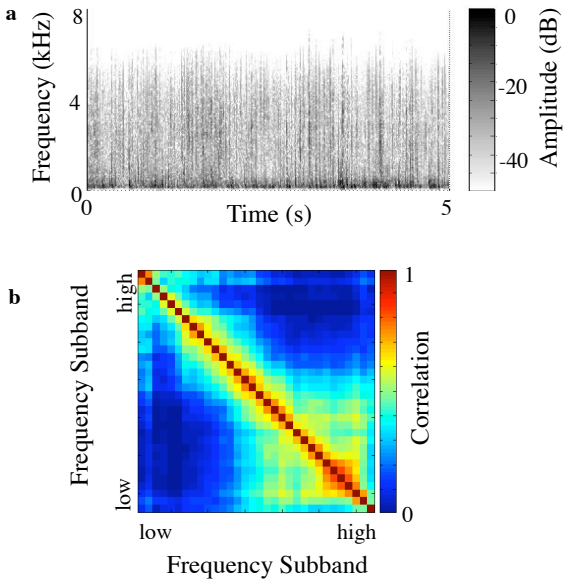

Figure 2: Statistics of a recording of fire. a) Spectrogram. b) Matrix of correlations between subband envelopes. Each cell represents the correlation coefficient for a pair of subbands.

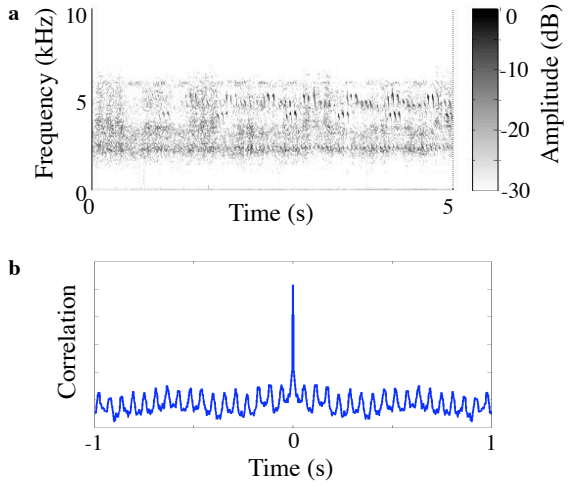

Figure 3: Statistics of a recording of swamp insects. a) Spectrogram. b) Autocorrelation of envelope for one subband (2.2-2.8 kHz).

3. SYNTHESIS ALGORITHM

To test whether these statistics are sufficient to capture the perceptual experience of naturally occurring sound textures, we designed an algorithm to synthesize new textures with particular statistics. Synthetic textures were generated by imposing the statistics of a particular real sound on a sample of (initially) Gaussian noise. Our method was inspired by visual texture synthesis algorithms [13, 1] in which statistics act as constraints on a noise signal.

The synthesis process begins by decomposing a sample real texture into its subband representation, and measuring the statistics of interest (Fig. 4). A noise sample is then generated, and each statistical constraint is imposed on its subbands. To produce a sound signal from these modified subbands, the filters used to generate the subbands are applied to each subband once more, and the results are summed. Such a scheme has the advantage of ensuring that the subbands remain band-limited despite the statistical adjustments (which are not guaranteed to respect the subband band limits). However, the refiltering and recombination of the subbands generally alters the subband statistics such that they no longer match their desired values. We thus iterate the process of imposing the statistical constraints and reconstructing the signal until the statistics converge to the desired values (see Fig. 4).

Imposing a particular value of a statistic (the kurtosis of a subband, for instance) involves changing one of the subbands until that statistic matches the desired value (as measured from the original sound texture). There are in principle many ways this could be accomplished; we chose to move in the direction of the statistic's gradient until the statistic reached the desired value [1]. The gradient direction has the attractive property of producing the largest change in the statistic for a given step size in signal-space, and it is readily derived for all of our statistics.

Gradient descent could be used to move through the space of signals, but we find that there is typically an analytic solution for the step size λ_k needed to reach the desired value of a statistic ϕ_k:

$$\vec{s}\,' = \vec{s} + \lambda_k \nabla \phi_k(\vec{s})$$

where $\vec{s}$ and $\vec{s}\,'$ are the signals before and after the adjustment, and

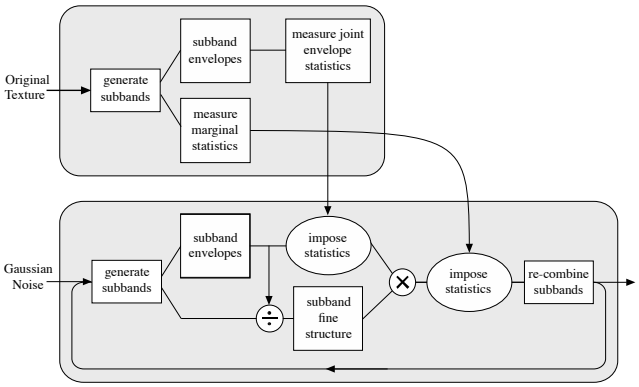

Figure 4: Schematic of synthesis algorithm.

∇ denotes the gradient. The adjustments for the subband moments are straightforward (details can be found in [1]). The kurtosis, for instance, has a gradient that is positive at the samples in a signal that are large in magnitude, and negative at the samples that are small in magnitude. The kurtosis may thus be increased by making the large samples larger and the small samples smaller.

The cross-band correlation adjustment is slightly more complex because pairs of subbands must be adjusted simultaneously. The correlation between subbands $\vec{s}_m$ and $\vec{s}_n$ is:

$$C_{n,m} = \sum_t \vec{s}_m(t)\vec{s}_n(t)$$

whose partial derivative with respect to a sample of $\vec{s}_n$ is proportional to the corresponding sample of $\vec{s}_m$. The gradient projection for each subband thus takes the form:

$$\vec{s}\,'_n = \vec{s}_n + \sum_k \lambda_{n,k}\vec{s}_k$$

Thus, the update to each subband is simply a linear combination of the other subbands. The adjustment procedure involves solving for the weights $\lambda_{n,k}$ that will produce the desired $C_{n,m}$. Details are given in [1]. One can in principle impose the entire matrix of cross-correlations. In practice, we find it is usually sufficient to enforce the correlations of each subband with the four nearest subbands above and below it. This can be done iteratively, proceeding from the low- to the high-frequency subbands.

We impose the subband envelope autocorrelation at a set of time lags Δt_j. We express the correlation at Δt_j as:

$$a_{\Delta t_j}(\vec{s}_n) = \sum_t \vec{s}_n(t)\vec{s}_n(t+\Delta t_j)$$

where $\vec{s}_n$ is zero-padded to reduce edge artifacts. The partial derivative of $a_{\Delta t_j}$ with respect to $\vec{s}_n(t_i)$ is proportional to $\vec{s}_n(t_i+\Delta t_j)+\vec{s}_n(t_i-\Delta t_j)$, and thus the gradient is proportional to the sum of two shifted copies of $\vec{s}_n$. For efficiency we impose all the autocorrelation coefficients simultaneously. The adjustment to $\vec{s}_n$ thus takes the form:

$$\vec{s}\,'_n(t) = \vec{s}_n(t) + \sum_j \lambda_{n,j}[\vec{s}_n(t+\Delta t_j) + \vec{s}_n(t-\Delta t_j)]$$

where we must solve for λ that produce correlations A = $\{a_{\Delta t_j}\}$. In lieu of an exact solution we Taylor-expand the coefficients A about $\vec{s}_n$ and solve for the adjustment via the pseudoinverse of the matrix of gradient vectors. We obtain good results using a set of 25 time lags ranging from 2 to 500 ms.

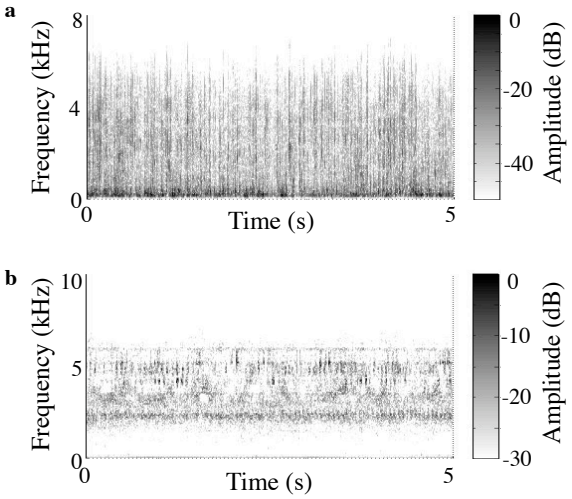

Figure 5: Spectrograms of synthetic examples produced from statistics of fire (a) and swamp insects (b).

The statistics in our constraint set are imposed sequentially. The envelope statistics are imposed first, after which the modified envelopes are combined with the old fine structure (Fig. 4). The subband marginal statistics are then imposed on the resulting new subbands. These modified subbands are then combined to yield a modified noise signal.

Each statistical constraint pushes the signal in a different direction. As these directions are generally not orthogonal, the adjustments interfere with each other. The filtering that occurs prior to combining the subbands also has the potential to partially undo the effect of the statistical adjustments. However, we find that with iteration, the process typically converges to a signal whose statistics are close to the desired values (the ratio of the squared statistic magnitude to the squared error in the statistic typically surpasses 30-40 dB).

We emphasize that the imposition of these statistical constraints does not simply recreate the original signal. Because the synthesis starts with a sample of random noise, the resulting signal is different every time, and shares only the statistical properties of the original sound. Fig. 5 shows spectrograms of synthetic fire and swamp insects sounds. Inspection reveals that they are distinct from the originals, despite having some qualitative similarities.

4. RESULTS

We find that the imposition of this set of statistics produces compelling synthetic examples of many natural sound textures. A set of example synthetic sounds may be found at `http://www.cns.nyu.edu/~jhm/texture.html`. Much can be learned simply from listening to results of the synthesis. Although such assessments are subjective, we have confirmed our observations in groups of listeners at conferences and seminars. We found that imposing only the marginal statistics (variance and kurtosis) of the subbands was sufficient to produce compelling synthetic examples of many water textures (rain, streams etc.), but not much else. It

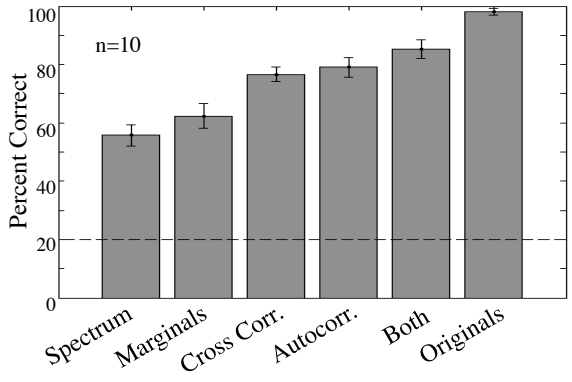

Figure 6: Results of texture recognition experiment (see text). Dashed line represents chance performance. Error bars denote standard errors.

was important to constrain the kurtosis in addition to the variance. If we imposed Gaussian kurtosis, or a value that was too high by a factor of 8, even the water sounds were generally not synthesized well. Synthesizing non-water sounds using only the marginal statistics tended to produce textures that sounded like water, and that sounded categorically different from the originals. Consistent with these observations, we found that many water sounds had cross-band correlations and autocorrelations that were near zero, suggesting they are produced by multiple, independent bandpass events. Sound examples illustrating these effects can be found at the website given above.

As a crude means of quantifying the quality of the synthesis, we ran 10 subjects in a texture recognition task. Subjects were presented with both synthetic and original samples of various natural sound textures (25 different textures in total, each 5 seconds in duration), and had to choose an identifying name for the sound from a set of five. We measured the percentage of correct choices for synthetic samples generated according to five different sets of statistical parameters: 1) the subband variances (approximately equivalent to matching the power spectrum); 2) the full marginal statistics; 3) the envelope cross-correlation; 4) the envelope autocorrelation; and 5) the envelope cross- and auto- correlations. The last three also included the marginal statistics. As shown in Fig. 6, subjects performed above chance levels with the spectrum alone (as many of the examples differed dramatically in frequency composition), but steadily improved as additional statistical constraints were imposed. Filter statistics can thus support the identification of sound textures.

5. CONCLUSIONS

We find that rudimentary statistics of bandpass filter responses suffice to produce realistic synthetic examples of many naturally occurring sound textures. The results support the notion that the auditory system represents textures through statistical measurements. Our method also provides a means to test the perceptual significance of different statistics, and we find that marginal subband statistics, as well as cross- and auto-correlations of their envelopes, can each capture perceptually important information.

There are numerous cases where the synthesis is clearly inadequate (documented in the online examples), suggesting the need for additional statistics. Some of the failures are for obvious reasons: our current set of statistics does not capture frequency modulation, for instance. Another limitation is that our measure of temporal correlation is invariant to the directionality of time. We thus cannot capture the asymmetric temporal envelopes that characterize many sounds. Harmonic sounds that change in pitch, as are found in many mammalian vocalizations, are also poorly captured by the present set of statistics, as the harmonics are scattered across filters, and the filter outputs change in complex ways as pitch is modulated. These failures can be used to identify additional statistics that may be important to the auditory system, and that can be incorporated into future versions of the algorithm to improve the quality of synthesis.

6. REFERENCES

[1] J. Portilla and E. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. Journal of Computer Vision*, vol. 40, pp. 49–71, 2000.

[2] B. Julesz, "Textons, the elements of texture perception, and their interactions." *Nature*, vol. 290, no. 5802, pp. 91–97, Mar 1981.

[3] N. Arnaud and K. Popat, "Analysis and synthesis of sound texture," in *AJCAI workshop on Computational Auditory Scene Analysis*, 1995, pp. 293–308.

[4] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Synthesizing sound textures through wavelet tree learning," vol. 22, no. 4, pp. 38–48, July 2002.

[5] M. Athineos and D. P. W. Ellis, "Sound texture modelling with linear prediction in both time and frequency domains," in *Proc. IEEE ICASSP*, vol. 5, 2003, pp. 648–51.

[6] X. Zhu and L. Wyse, "Sound texture modeling and time-frequency lpc," in *Proceedings of the 7th International Conference on Digital Audio Effects (DAFX-04)*, 2004.

[7] A. J. Bell and T. J. Sejnowski, "Learning the higher-order structure of a natural sound." *Network*, vol. 7, no. 2, pp. 261–267, May 1996.

[8] H. Attias and C. Schreiner, "Temporal low-order statistics of natural sounds," in *Advances in Neural Information Processing*, M. Mozer, M. Jordan, and T. Petsche, Eds., 1997.

[9] O. Schwartz and E. Simoncelli, "Natural sound statistics and divisive normalization in the auditory system," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds. MIT Press, 2001.

[10] M. S. Lewicki, "Efficient coding of natural sounds." *Nat Neurosci*, vol. 5, no. 4, pp. 356–363, Apr 2002.

[11] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data." *Hear Res*, vol. 47, no. 1-2, pp. 103–138, Aug 1990.

[12] M. N. Kvale and C. E. Schreiner, "Short-term adaptation of auditory receptive fields to dynamic stimuli." *J Neurophysiol*, vol. 91, no. 2, pp. 604–612, Feb 2004.

[13] D. Heeger and J. Bergen, "Pyramid-based texture analysis/synthesis," in *Computer Graphics (ACM SIGGRAPH Proceedings)*, 1995, pp. 229–238.