Auditory texture synthesis from task-optimized convolutional neural networks

Jenelle Feather (jfeather@mit.edu)
MIT Department of Brain and Cognitive Sciences
77 Massachusetts Ave, Cambridge, MA 02143, USA

Josh H. McDermott (jhm@mit.edu)
MIT Department of Brain and Cognitive Sciences
77 Massachusetts Ave, Cambridge, MA 02143, USA

Abstract

Models of sensory systems have traditionally been hand designed from engineering principles, but modern-day machine learning allows models to be learned from data. We sought to compare hand-engineered and learned models of the auditory system by generating synthetic sound textures. We synthesized sounds that produce the same time-averaged values in each model's representation as those measured from a natural texture using gradient-based optimization. Such stimuli should evoke the same texture percept if the model replicates the representations underlying auditory texture perception. Previous texture models involved statistics measured from multiple stages of standard visual or auditory processing cascades. We found that auditory textures generated simply from the time-averaged power in the first layer activations of a task-optimized convolutional neural network were as realistic and recognizable as the best previous auditory texture model. Unlike textures generated from traditional models, the textures from task-optimized filters did not require statistics from earlier stages in the sensory model (i.e., the cochlear stage). Further, the textures generated from the task-optimized CNN filters were more realistic than textures generated from a widely used hand-engineered model of primary auditory cortex. The results demonstrate that better sensory models can be obtained by task-optimizing sensory representations.

Keywords: neural networks; textures; sound

Introduction

Textures are typically generated by superpositions of large numbers of similar elements, and are distinguished from other sensory signals by homogeneity in time or space. As such, textures are believed to be represented in the brain with statistics that average information across space or time (McDermott, 2013; Ziemba, Freeman, Movshon, & Simoncelli, 2016). Textures have been a fruitful avenue to explore sensory representations in part because they are the only class of signals for which we have signal-computable models that come close to accounting for perception. Texture models are commonly evaluated with synthesis: a set of statistics is measured from a natural signal's representation in the model, and a synthetic signal is produced that has the same statistics as the natural signal (Heeger & Bergen, 1995; Portilla & Simoncelli, 2000; McDermott & Simoncelli, 2011). A model that replicates perceptual representations of textures should produce synthetic signals that replicate the perceptual attributes of the natural signals to which they are matched.

The statistics in traditional visual and auditory texture models are somewhat ad-hoc, assembled through a process of intuition-guided trial-and-error. Moreover, multiple classes of statistic are required to attain synthetic textures that replicate the qualities of natural textures. For instance, the texture model in McDermott and Simoncelli (2011) relies on the correlations between pairs of filters in addition to marginal statistics. These traditional models also rely on statistics measured from multiple stages of the underlying sensory cascade (for instance, statistics from cochlear filters as well as subsequent stages of modulation filters) in order to produce realistic textures. Relying on statistics from multiple stages is in some cases difficult to justify. For instance, it may be implausible to suppose that decisions could be based directly on the output of the cochlea. We were interested in whether a single, simple class of statistic (variance) measured at a single stage of an appropriate auditory model could replicate the multi-stage, multi-statistic representation of traditional texture models (McDermott & Simoncelli, 2011).

Task-optimized convolutional neural networks have been shown to outperform traditional hand-engineered models of both the auditory and visual system for predicting neural activity (Yamins & DiCarlo, 2016; Kell, Yamins, Shook, Norman-Haignere, & McDermott, 2018). Prior literature suggests that the representations learned by these models can produce realistic visual textures (Gatys, Ecker, & Bethge, 2015; Ustyuzhaninov*, Brendel*, Gatys, & Bethge, 2017; Wallis et al., 2017), and we sought to explore the relevance of comparable audio representations for sound texture. The first layer of 2D-filters of a CNN trained on a cochleagram representation can be thought of as spectrotemporal modulation filters (filters that act in frequency and time), and thus can be compared to prior models of primary auditory cortex (such as that in Chi, Ru, and Shamma (2005)). Here, we compared textures generated from: (1) the first layer of filters from three convolutional neural networks optimized for different tasks (2) the randomly initialized filters from the same architecture (3) a model of primary auditory cortex consisting of spectrotemporal filters, and (4) the McDermott and Simoncelli (2011) texture model (consisting of marginal moments and correlations from cochlear and temporal modulation filters).

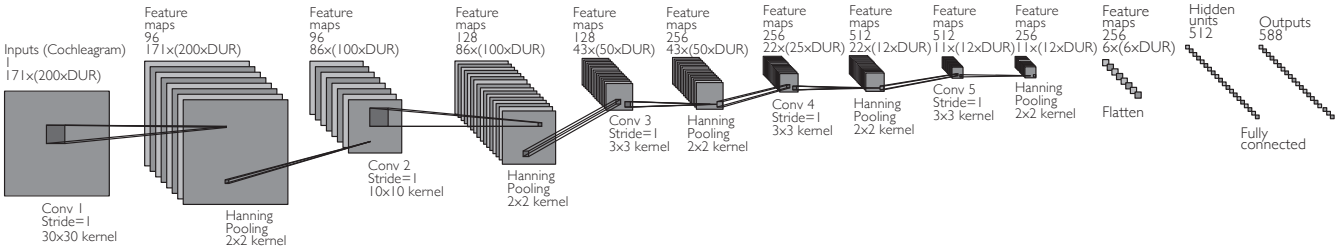

Figure 1: Neural network architecture. The time-averaged power from the activations of the first convolutional layer was used as the texture model.

Methods

CNN training

The convolutional network consisted of 5 rectified convolutional layers, 5 pooling layers, a fully connected layer, and a final softmax classification layer (architecture details are in Figure 1). All pooling layers downsampled by a factor of 2, using a weighted average with a hanning window. Convolutions were performed in two dimensions, such that the learned convolutional filters were shared across frequency and time.

The input to the first convolutional layer of the network was a "cochleagram". To generate the cochleagram, a natural sound was sent through a filter bank modeled after the human cochlea. The filterbank consisted of 171 filters, spaced between 20Hz-8kHz. The envelope of each audio subband was extracted via the Hilbert transform, downsampled to 200Hz, and passed through a compressive nonlinearity. This yields a cochleagram representation, similar to a conventional spectrogram with frequency on the y-axis and time on the x-axis, but with frequency resolution based on the human cochlea. We consider the transformations from waveform to cochleagram as a cochlear model, and compare the synthesis with and without statistics measured from these stages (which are critical to traditional texture models).

Other statistics used for model comparisons are measured from the activations of the first convolutional layer (consisting of 96 30x30 filters), which acts on the cochleagram. We refer to these as "cortical" features, as primary auditory cortex is commonly modeled with spectrotemporal filters that act on a cochleagram Chi et al. (2005).

The same architecture was trained on three different tasks for comparison: (a) classification of the word in the middle of a clip from 588 possibilities (b) classification of the speaker in clip from 789 possibilities (c) classification of the genre of music from 42 possibilities. The target sound was embedded in background noise for each of the tasks. The training sounds and task parameters were the same as those described in Kell et al. (2018). Performance of the trained networks on the tasks was comparable to humans, with top 1 accuracy of 78% for the word task, 91% for the speaker identification task, and 45% for the genre task. Sound were also synthesized from a random network with untrained weights to disambiguate effects of training from those inherent to the model architecture.

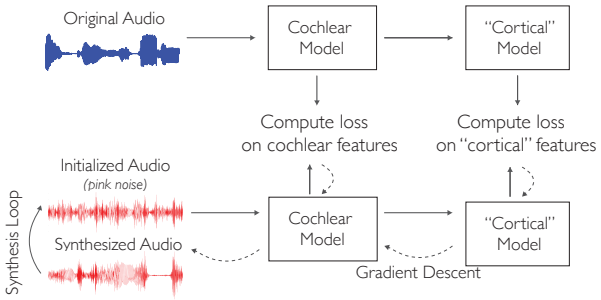

Figure 2: Diagram of sound synthesis pipeline.

Hand-engineered filter bank

A set of hand-engineered spectrotemporal filters were taken from a commonly-used model of primary auditory cortex (Chi et al., 2005). They consisted of a set of gabor-like spectrotemporal filters tuned to specific spectral and temporal modulations. Ninety-six filters were used to match the number in the first layer of the CNN. Eight temporal filters were spaced between $\pm$ 0.5-64 Hz and six spectral filters were spaced between 0.25-8 cycles/octave. Each of the temporal filters at 0.5Hz were lowpass to capture DC components in the cochleagram representation.

We also compared the results directly to textures generated from the full model in McDermott and Simoncelli (2011), which contains the moments and correlations of a sound decomposition similar to that in the subcortical auditory system.

Sound synthesis

Synthetic stimuli were synthesized to have the same measured filter statistics as a natural sound. Measured "cortical" features comprised the time-averaged power in each of the spectrotemporal filters (either task-optimized for hand-engineered). Cochlear model features, when included, were the first four moments of the cochlear subbands and their envelopes, as have been used in previous texture models McDermott and Simoncelli (2011). The synthetic signal was initialized with pink noise. Gradient descent was performed on the waveform to minimize the difference between its model responses and those of the target natural sound, and iterated until the statistics were the same (Figure 2).

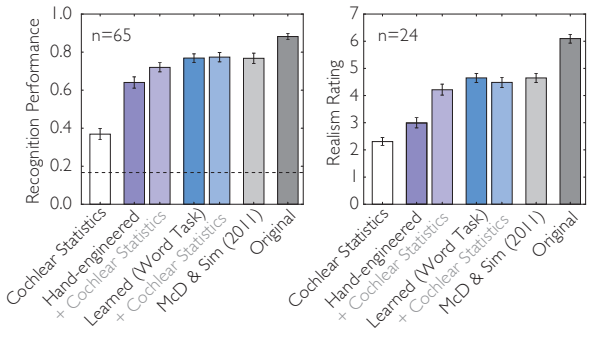

Figure 3: Recognition (left) and realism (right) of synthetic textures generated from the filters optimized for work recognition or from random filters, with and without the inclusion of cochlear statistics.

Recognition and realism experiments

We performed two experiments - a sound texture recognition task and a realism rating task - to assess whether the synthetic results generated from each model were perceived similarly to natural textures. Sounds included in these experiments were drawn from the texture set used in McDermott and Simoncelli (2011). A seven-second clip of each texture was synthesized, and two-second segments were extracted to avoid any boundary effects from the synthesis algorithm. For the recognition experiment, subjects were presented with a two-second sound and identified the sound from 6 candidate descriptions. For the realism experiment, subjects first listened to the original sound, and then were presented with synthesized examples from the different models and asked to rate the sounds from 1-7 based on their similarity to the original (a MUSHRA paradigm). All experiments were run on Amazon Mechanical Turk, and subjects were screened for headphone usage (Woods, Siegel, Traer, & McDermott, 2017). Error bars and significance tests were obtained by bootstrapping across subjects.

Results

Textures generated from power statistics of the first layer filters of a CNN were as recognizable and realistic as those from the previous McDermott and Simoncelli model (Figure 3). Moreover, the task-optimized filters produced more realistic and recognizable textures than the hand-engineered spectro temporal filters, suggesting a benefit to learned filters. To get insight into the source of the difference between the two sets of filters, we tested the effect of also including statistics from the cochlear filter model stage (as are present in the original McDermott and Simoncelli model). Cochlear marginals alone do not produce realistic or recognizable textures, but their inclusion increased the realism and recognizability of the textures from the the hand-engineered filters. This increase did not occur for the task-optimized filters, suggesting that they are implicitly encoding the perceptually relevant statistics from lower stages in the sensory model. Cochleagrams corresponding to sounds from these model comparisons are shown in Figure 4.

Textures generated from random filters were less recognizable and realistic than those generated from the filters optimized for the word-in-noise task (Figure 5), though they similarly did not benefit from the inclusion of cochlear statistics. This suggests that task-optimization helps to generate filters that capture perceptually relevant information, and that the CNN architecture alone is not enough.

The exact task that the network was trained on did not greatly influence the recognizability of the synthetic textures (Figure 6) – first layer filters from networks trained on the genre or speaker id tasks produced comparably realistic textures to those from the word-in-noise task. In all cases, cochlear statistics did not improve recognizability for any of the task-optimized networks, suggesting that all of the tasks force the filters to implicitly encode perceptually relevant cochlear features.

Discussion

The results demonstrate that realistic and recognizable sound textures can be generated simply by measuring and matching power statistics from a single layer of spectro-temporal filters acting on a cochleagram representation. These statistics produce textures that are as recognizable and realistic as previous textures models that measure multiple classes of statistics at multiple stages of a sensory cascade (McDermott & Simoncelli, 2011). A commonly used hand-engineered model as well as a set of random filters produced lower quality textures. This suggests that the representations learned through task-optimization may more closely resemble biological sensory systems than traditional hand-engineered auditory models.

Acknowledgments

Work was supported by a Department of Energy Computational Science Graduate Fellowship to J.F., McDonnell Scholar Award to J.H.M. and NSF grant BCS-1634050 to J.H.M.

References

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). Texture Synthesis Using Convolutional Neural Networks. , 1–10. Retrieved from http://arxiv.org/abs/1505.07376 doi: 10.1109/CVPR.2016.265

Heeger, D. J., & Bergen, J. R. (1995). Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd annual conference on computer graphics and interactive techniques* (pp. 229–238).

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 1–15. Retrieved from http://www.cell.com/neuron/fulltext/

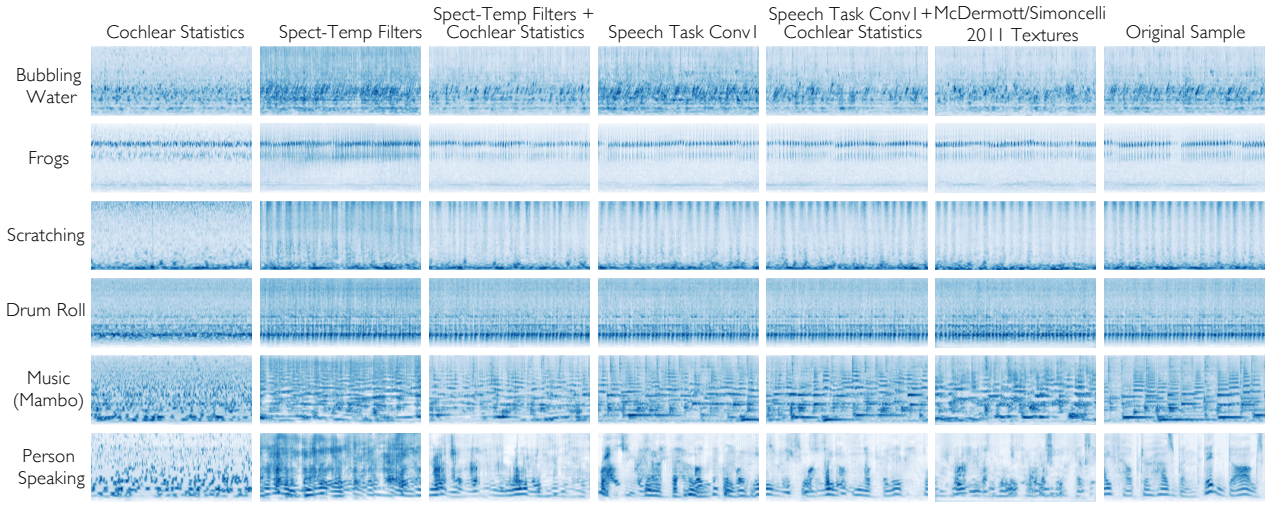

Figure 4: Example cochleagrams for the original and synthetic sounds. Music and speech are included for comparison with the stationary textures.

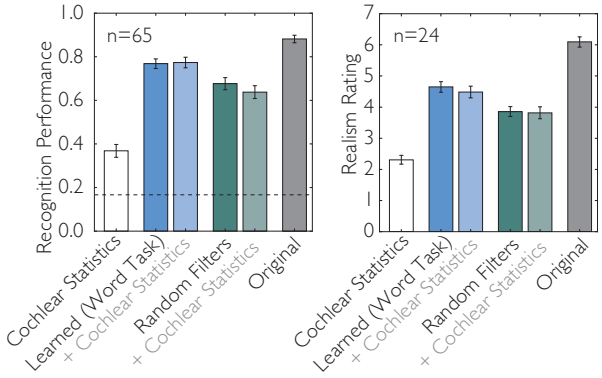

Figure 5: Recognition (left) and realism (right) results comparing the word-optimized filters to randomly initialized filters.

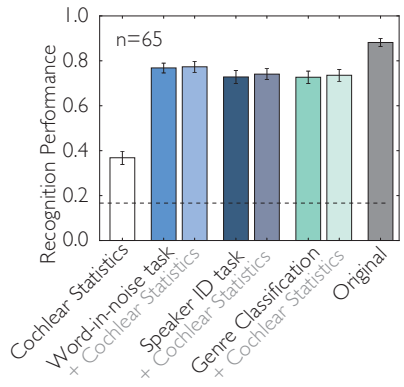

Figure 6: Recognition results comparing sounds generated from the three task-optimized networks.

S0896-6273(18)30250-2 doi: 10.1016/j.neuron.2018.03.044

McDermott, J. H. (2013). Audition. *The Oxford Handbook of Cognitive Neuroscience*, 135–170.

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, *71*(5), 926–940. doi: 10.1016/j.neuron.2011.06.032

Portilla, J., & Simoncelli, E. P. (2000). Parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–71. doi: 10.1023/A:1026553619983

Ustyuzhaninov*, I., Brendel*, W., Gatys, L., & Bethge, M. (2017, Apr). What does it take to generate natural textures?. Retrieved from https://openreview.net/forum?id=BJhZeLsxx

Wallis, T. S. A., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., & Bethge, M. (2017). A parametric texture model based on deep convolutional features closely matches texture appearance for humans. *Journal of Vision*, *17*(12), 5. Retrieved from http://jov.arvojournals.org/article.aspx?doi=10.1167/17.12.5 doi: 10.1167/17.12.5

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064–2072.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. doi: 10.1038/nn.4244

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque v2. *Proceedings of the National Academy of Sciences*, *113*(22), E3140–E3149.