Learning mid-level codes for natural sounds

Wiktor Młynarski, Josh H. McDermott
Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

Auditory perception depends critically on abstract and behaviorally meaningful representations of natural auditory scenes. These representations are implemented by cascades of neuronal processing stages in which neurons at each stage recode outputs of preceding units. Explanations of auditory coding strategies must thus involve understanding how low-level acoustic patterns are combined into more complex structures. While models exist in the visual domain to explain how phase invariance is achieved by V1 complex cells, and how curvature representations emerge in V2, little is known about analogous grouping principles for mid-level auditory representations.

We propose a hierarchical, generative model of natural sounds that learns combinations of spectrotemporal features from natural stimulus statistics. In the first layer the model forms a sparse, convolutional code of spectrograms. Features learned on speech and environmental sounds resemble spectrotemporal receptive fields (STRFs) of mid-brain and cortical neurons, consistent with previous findings [1]. To generalize from specific STRF activation patterns, the second layer encodes patterns of time-varying magnitude (i.e. variance) of multiple first layer coefficients. Because it forms a code of a non-stationary distribution of STRF activations, it is partially invariant to their specific values. Moreover, because second-layer features are sensitive to STRF combinations, the representation they support is more selective to complex acoustic patterns. The second layer substantially improved the model's performance on a denoising task, implying a closer match to the natural stimulus distribution.

Quantitative hypotheses emerge from the model regarding selectivity of auditory neurons characterized by multidimensional STRFs [2] and sensitivity to increasingly more abstract structure [3]. The model also predicts that the auditory system constructs representations progressively more invariant to noise, consistent with recent experimental findings [4]. Our results suggest that mid-level auditory representations may be derived from high-order stimulus dependencies present in the natural environment.

References:

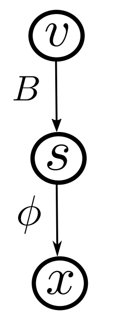
[1] Carslon, N.L., Ming V. L., DeWeese M.R. (2012) Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLOS Computational Biology*

[2] Atencio CA, Sharpee TO, Schreiner CE (2009) Hierarchical computation in the canonical auditory cortical circuit. *Proceedings of the National Academy of Sciences*

[3] Chechik, G., & Nelken, I. (2012). Auditory abstraction from spectro-temporal features to coding auditory entities. *Proceedings of the National Academy of Sciences*

[4] Rabinowitz, N. C., Willmore, B. D., King, A. J., & Schnupp, J. W. (2013). Constructing noise-invariant representations of sound in the auditory pathway. *PLOS Biology*

A)



B)

$$p(v_{j,t}) = Laplace(0,1)$$

$$\log\frac{\lambda_{i,t}}{\lambda_i} = \left[\sum_j B_{j,i} * v_j\right]_t$$

$$p(s_{i,t}|\lambda_{i,t}) = FoldedLaplace(0,\lambda_{i,t})$$

$$\hat{x}_{t,f} = \left[\sum_i \phi_{i,f} * s_i\right]_t$$

$$p(x_{t,f}) = \mathcal{N}(\hat{x}_{t,f},\sigma^2)$$

Fig 1 Model description. A) Graphical model presenting dependency structure between variables (deterministic dependencies are omitted). B) Detailed definition. $v_{j,t}$ - second layer activation, B - second layer features, defined by time-varying weights for different first layer STRFs, λ_i -marginal variance of s_i, $\lambda_{i,t}$ - instantaneous variance (magnitude) of $s_{i,t}$, $s_{i,t}$ - nonegative STRF activation, ϕ - STRFs, defined by weights on spectrogram bins, $x_{t,f}$ - spectrogram bins.

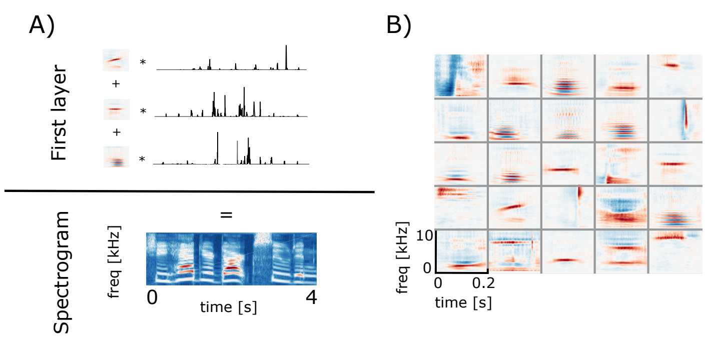


Fig 2 First layer of the model A) First layer convolutional sparse code of spectrograms. A spectrogram (at the bottom) is represented as a linear combination of spectrotemporal basis functions (ϕ) convolved with their non-negative activations (s). B) Example spectrotemporal features (STRFs) learned from natural sounds.

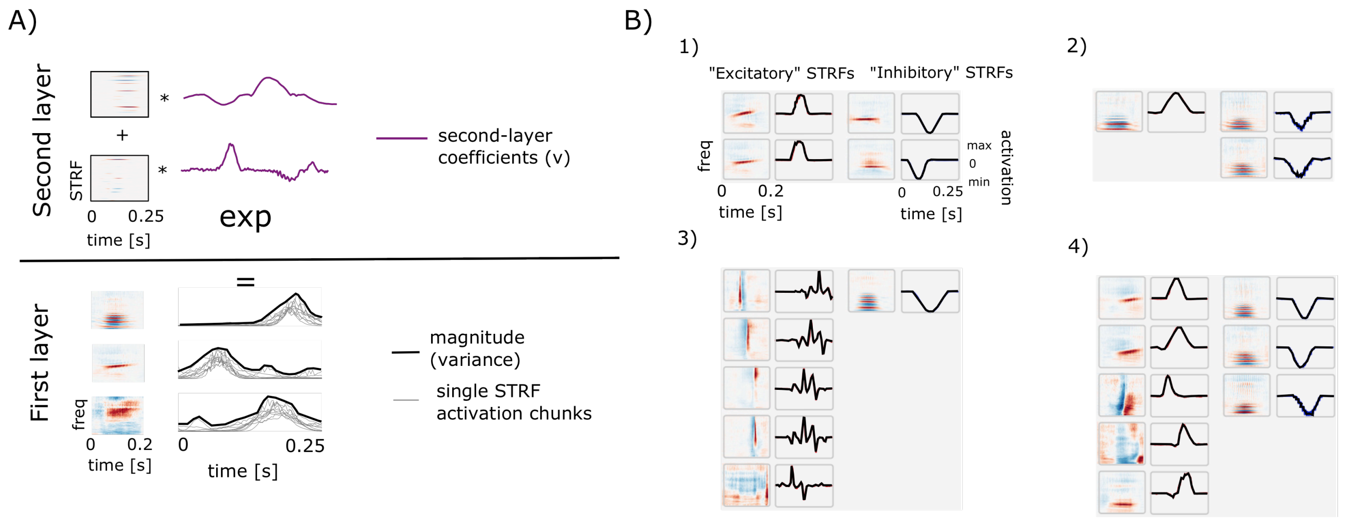


Fig 3 Second layer of the model. A) Similar patterns of first-layer STRF activations (gray lines) are interpreted as different samples from a distribution with a particular time varying instantaneous variance (i.e magnitude, λ, thick black lines). The log-variance is encoded by a population of second layer units (*B*, visualized in Fig. 3B), convolved with sparse activation time courses (v). B) Example second-layer units. Panels 1-4 depict first-layer STRFs together with their temporal variance patterns encoded by a single second-layer unit. Negative ("inhibitory") STRF activations are plotted in blue, while positive ("excitatory") activations are plotted in red. Some interpretable patterns are evident - e.g. unit C becomes activated by a sequence of impulsive events (clicks) and suppressed by a harmonic STRF.

Fig 4 Invariance signature. Histograms of correlations among optimal stimuli for model units. Optimal stimuli are spectrogram chunks which elicit maximally positive (red line) or negative (blue line) responses in the second layer, or maximal responses in the first layer (black line, first layer coefficients are non-negative). While first-layer STRFs respond most strongly to highly correlated stimuli, optimal stimuli for the second layer units are much less correlated. This is evidence of increased invariance of the representation relative to the first layer.

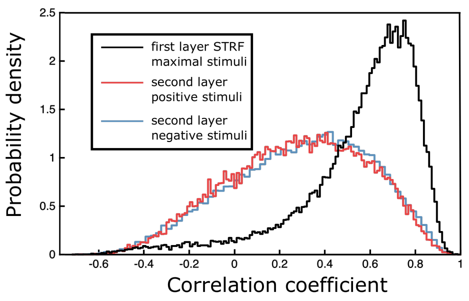
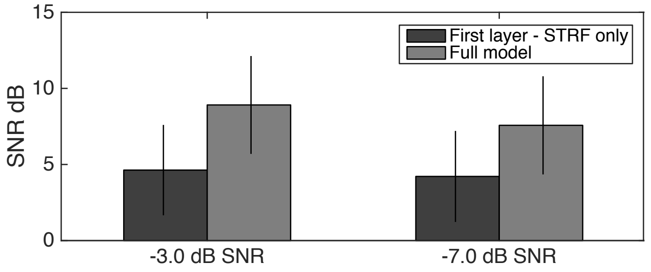


Fig 5 Denoising results. 300 ms long speech spectrograms from the TIMIT database were distorted with Gaussian white noise (-3 and -7 dB SNR). Single-layer STRF model (black bars) is capable of recovering the signal from noise to some extent, but the full, 2-layer model (gray bars) substantially increases the denoising performance.


SNR dB
First layer - STRF only
Full model
-3.0 dB SNR
-7.0 dB SNR