A sparse, subspace model of higher-level sound structure

Jimmy Wang[1], **Bruno Olshausen**[1] **& Vivienne L'Ecuyer Ming**[1,2]

[1]Redwood Center for Theoretical Neuroscience, UC Berkeley
[2]Mind, Brain & Computation/MBC, Stanford University

An understanding of auditory processing requires an understanding of the statistics of natural sounds. It has been shown in previous work, for example, that the response properties of neurons in the cochlea are well matched to the statistics of the raw sound waveform so as to provide a sparse code of natural sound[1]. Learning higher-level structure and evaluating its relevance to processing along the auditory pathway will require analyzing more abstract properties of sound rather than the raw sound waveform. Here we present a model for learning a representation of sound in terms of multidimensional subspaces that are adapted to the statistics of natural sound. The signal is represented as a linear combination of subspaces

$$y(t) = \sum_{n=1}^{N} \sum_{m=1}^{M} s_n^m A_n^m(t) + \eta(t) \tag{1}$$

where A_n^m is the m^{th} component of the n^{th} subspace, s_n^m is its corresponding coefficient and $\eta(t)$ is additive Gaussian noise. A sparse prior is imposed on the norm of the subspace coefficients that encourages independence between different subspaces, but other than the norm there are no constraints on the values of the coefficients within a subspace. The learning is done via a variational EM method that iterates between maximizing the log-probability of the coefficients and the dimensions of the subspaces. Specifically, we alternate between

$$S^{t+1} = \underset{S}{\operatorname{argmin}} ||Y - A^t S||_2^2 \quad s.t. \quad \sum_{n=1}^{N} C\left(\sum_{m=1}^{M} (\mathbf{s}^m)^2\right) < \lambda \tag{2}$$

$$A^{t+1} = \underset{A}{\operatorname{argmin}} ||Y - A S^{t+1}||_2^2 \quad s.t. \quad ||A_n|| < 1, \forall n = 1, ..., N \tag{3}$$

where $C(.)$ is a sparse penalty function. To optimize the coefficients given the subspaces (equation 2), we employ the *subspace thresholding circuit*, a computationally efficient and neurally plausible gradient descent based method[2]. It implements a dynamical system inducing local competition to achieve a sparse decomposition of the signal. Optimizing the subspaces given the coefficients (equation 3) is a quadratic programming problem which we solve using its Lagrangian dual[3]. Solving the dual problem is much faster as the number of dual variables to be optimized equals the number of basis functions, N, while the number of primal variables equals the total number of basis elements, $N \times M$. Using these methods we adapt the model to both waveform and spectrogram representations of spoken English. The resulting subspaces learn a variety of acoustic invariances, including phase- and bandwidth-invariance in waveform subspaces, formant- and limited "pitch"-invariance in spectrogram subspaces.

References

[1] Efficient auditory coding. E. Smith & M. Lewicki, *Nature* 439(7079), 2006.

[2] Neurally plausible sparse coding via thresholding and local competition. C. Rozell, D. Johnson, R. Baraniuk & B. Olshausen, *Neural Computation* in press.

[3] Efficient sparse coding algorithms. H. Lee, A. Battle, R. Rajat & A. Ng, *Advances in Neural Information Processing* 19, 2007.