Sparse Representations for the Cocktail Party Problem

Barak A. Pearlmutter,*1 Hiroki Asari,*2 and Anthony M. Zador*3

*1: Hamilton Institute, National University of Ireland Maynooth, Co. Kildare, Ireland. Ollscoil na hÉireann Má Nuad, Éire.

*2: Cold Spring Harbor, Watson School of Biological Sciences, One Bungtown Road, Cold Spring Harbor, NY 11724, USA

*3: Cold Spring Harbor, One Bungtown Road, Cold Spring Harbor, NY 11724, USA

asari@cshl.edu http://zadorlab.cshl.edu/

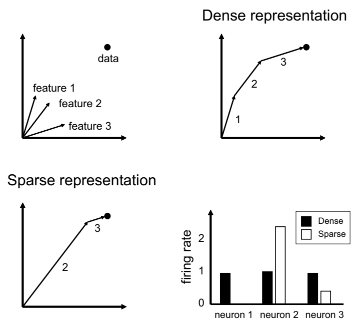
Sparse Overcomplete Representation

Why does the cortex have so many neurons?

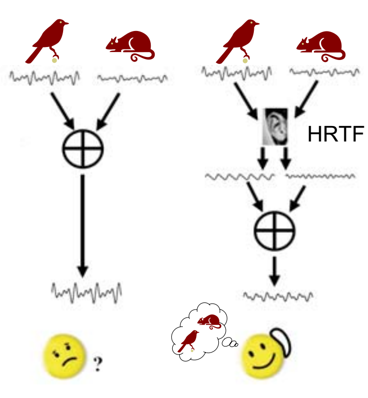
30,000 cochlear neurons

>100 million neurons in the auditory cortex

Intuition



Problem formulation

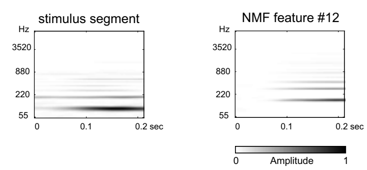


Source Separation

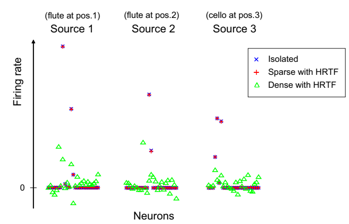
Assumptions

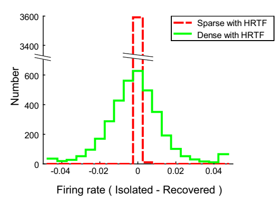
- Sparse representations
- Prior knowledge of HRTF

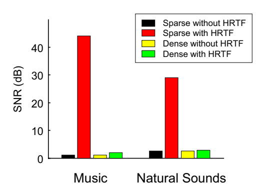
Spectral basis for sources via NMF



Performance of different separation approaches





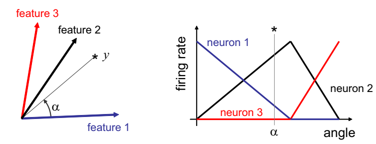


Conclusions

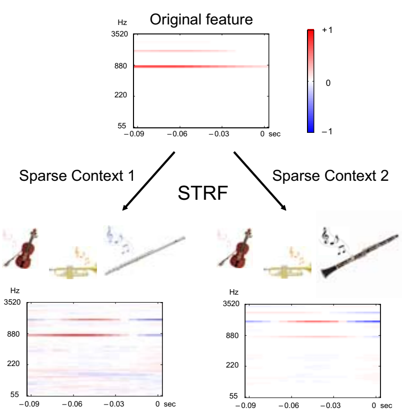
- We propose that the cortex exploits the excess neural representations by selecting the sparsest representation within an overcomplete set of features.
- Sparseness can be used to separate sources perceived monaurally, by exploiting the differential filtering imposed by the HRTF.
- Our results support the idea that sparse representations may underlie efficient computations in the auditory cortex.

Predictions

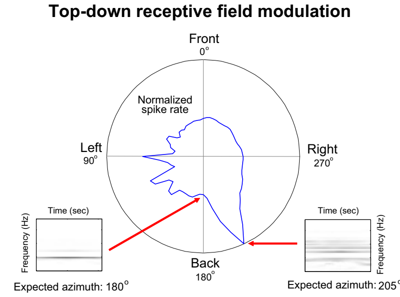
Decoding is linear whereas encoding is nonlinear.



Context dependence of receptive field



Top-down receptive field modulation

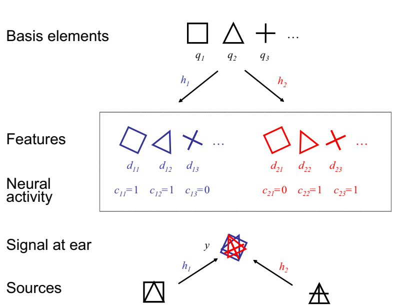


Prediction Summary

- Neural representations should be sparse.
- Neural decoding is linear whereas neural encoding is nonlinear.
- Representations should be dynamically influenced by acoustic context (bottom-up).
- Representations should be dynamically modulated by top-down influences, including spatial expectation.

Methods

Cartoon of the model



Problem formulation

- **Sound Source** : $x_i(t) = \sum_j c_{ij} q_j(t)$ ($i \ll j$, c_{ij} : sparse)

  [Sounds ($i = 1,\ldots,N$)] = [Coefficients] × [Basis elements]

- **Filter** (Head-Related Transfer Function) : $h_i(t)$

- **Features** : $d_{ij}(t) = h_i(t) * q_j(t)$

  [Features] = [Filter (HRTF)] ∗ [Basis elements]

- **Received signal** : $y(t) = \sum_i h_i(t) * x_i(t) = \sum_{ij} c_{ij} d_{ij}(t)$

  [Signal (Monaural)] = [Filter (HRTF)] ∗ [Sounds ($i = 1,\ldots,N$)]

  = [Coefficients] × [Features]

Question

How to recover the underlying sources $x_i(t)$ from the signal $y(t)$ using the knowledge of the directional filters $h_i(t)$?

Given **y** and **D**, how can we solve **y** = **D c** where the number of columns in **D** is larger than that of rows ?

Sparse representation

L1-minimization: $\min \sum_{i,j} \|c_{ij}\|_{L_1}$ subject to $y(t) = \sum_{i,j} c_{ij} d_{ij}(t)$

Dense representation

L2-minimization: (pseudoinverse) $\mathbf{c}_{L2} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}$

NMF (non-negative matrix factorization)

Algorithm for factorizing a data matrix under an elementwise non-negativity constraint