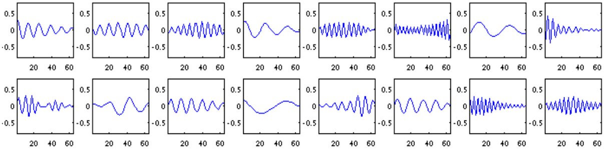
Learning Overcomplete Subspace Structures on Natural Speech Signal

Jimmy Wang
CIFAR

Neural Computation & Adaptive Perception Summer School
August 9, 2007

Sparse Coding on Natural Sound

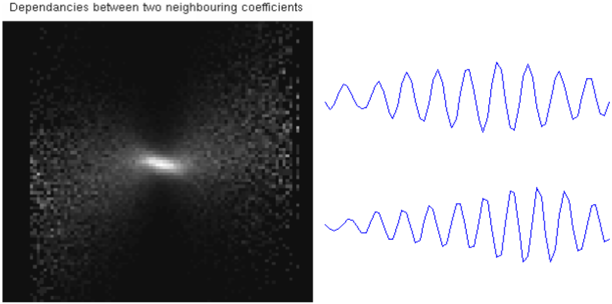
- Any N-dimensional signal can be represented by N orthogonal basis functions.
- Natural sound/image signals only occupy a small subset of the N dimensional space.
- Sparse coding (Olshausen and Field, 1996) learns an overcomplete set of basis functions that are optimal in representing natural image patches.
- Under similar principle, basis functions are learnt from natural auditory signals (Lewicki 2002).



- The learned basis functions are similar to the receptive fields found in cat auditory nerves.

Motivation for a Subspace Model

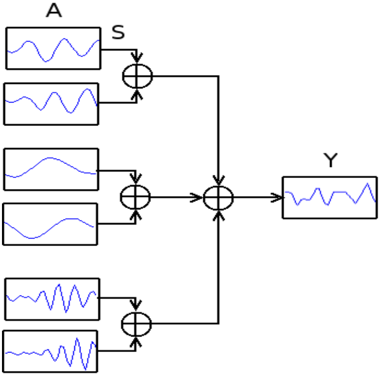
- Coefficients from neighboring basis functions are highly dependant.

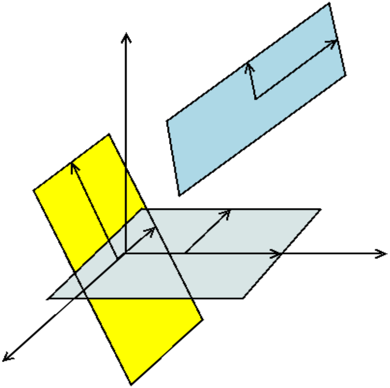


- Sparse coding forces these coefficients to be independent during sparsification.
- A better model should capture these dependencies, describing signal structure while facilitating these dependencies.

The subspace signal model

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





Model Learning

- The learning is carried out by maximize the log likelihood of the model over the data

$$L =< \log P(Y|A) >$$

- The update rule for the basis function is

$$\Delta A \propto \frac{\partial L}{\partial A} \propto \left\langle \left\langle (Y - A\mathbf{s})\mathbf{s}^T \right\rangle_{P(\mathbf{s}|Y,A)} \right\rangle$$

- The update rule requires us to sample from the posterior. ““
- If the distribution is sparse, the density of the posterior can be approximated by its maximum

$$\begin{aligned} \mathbf{s}^* &= \operatorname*{argmax}_{\mathbf{s}} \log P(Y|A,\mathbf{s})P(\mathbf{s}) \\ &= \operatorname*{argmin}_{\mathbf{s}} ||Y - A\mathbf{s}||_2^2 + \lambda C\left(\sum_{m=1}^{M} (\mathbf{s}^m)^2\right) \end{aligned}$$

Inference via Gradient Descent

- This optimization problem can be solved through gradient ascent

$$\Delta \mathbf{s}_p = (Y - A\mathbf{s})A_p^T - \lambda \frac{\partial C(\mathbf{s}_p)}{\partial \mathbf{s}_p}$$

$$= Y^T A - \sum_{q \neq p}^{N} A_q^T A_q \mathbf{s}_q - g(\mathbf{s}_p)$$

where $g(\mathbf{s}_p) = \mathbf{s}_p + \lambda \frac{\partial C(\mathbf{s}_p)}{\partial \mathbf{s}_p}$

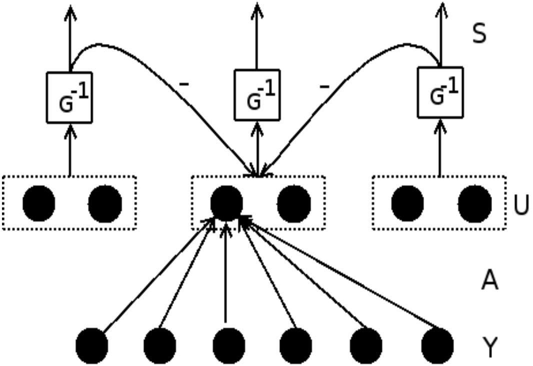
- If we let $\mathbf{u}_p = g(\mathbf{s}_p)$ and let $\mathbf{u}_p$ follow the energy gradient with respect to $\mathbf{s}_p$, we have

$$\mathbf{u}_p(t) + \tau \dot{\mathbf{u}}_p = Y^T A - \sum_{q \neq p}^{N} A_q^T A_q \mathbf{s}_q(t)$$

$$\mathbf{s}_p(t+1) = g^{-1}(\mathbf{u}_p(t))$$

Subspace Thresholding Circuit

- A circuit implementation of the non-linear differential equation

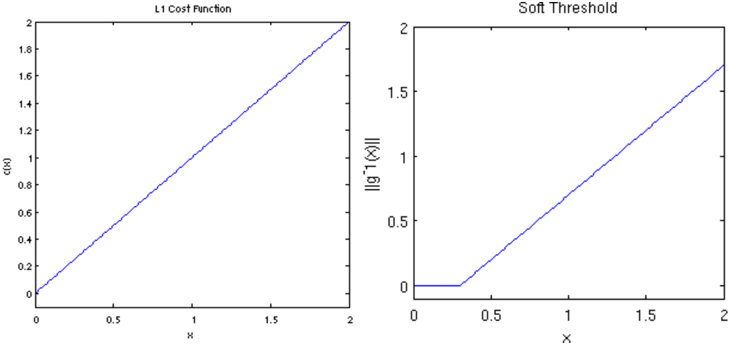


- This circuit will be efficient if $\mathbf{s}$ is sparse, i.e. no need to compute $A_q^T A_q \mathbf{s}_q$ if $\mathbf{s}_q = 0$.
- A biological plausible implementation.

- Let $C(\mathbf{s}_p) = ||\mathbf{s}_p||_1$, the thresholding function g^{-1} is

$$||\mathbf{s}_p|| = ||g^{-1}(\mathbf{u}_p)|| = \begin{cases} 0 & \text{if } ||\mathbf{u}_p|| \le \lambda \\ ||\mathbf{u}_p|| - \lambda & \text{if } ||\mathbf{u}_p|| > \lambda \end{cases}$$

$$\angle \mathbf{s}_p = \angle g^{-1}(\mathbf{u}_p) = \angle \mathbf{u}_p$$



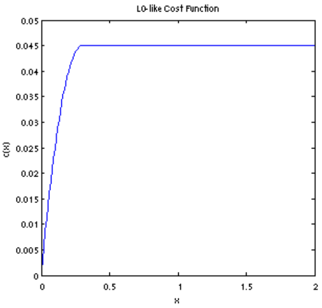
Hard Thresholding Functions

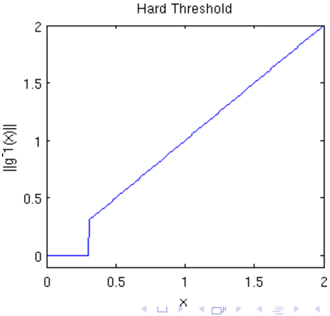
- $$C(\mathbf{s}_p) = \begin{cases} \frac{1}{2}(\lambda^2 - (||\mathbf{s}_p|| - \lambda))^2 & \text{if } ||\mathbf{s}_p|| \le \lambda \\ \frac{1}{2}\lambda^2 & \text{if } ||\mathbf{s}_p|| > \lambda \end{cases}$$

  the thresholding function g^{-1} is

  $$||\mathbf{s}_p|| = ||g^{-1}(\mathbf{u}_p)|| = \begin{cases} 0 & \text{if } ||\mathbf{u}_p|| \le \lambda \\ ||\mathbf{u}_p|| & \text{if } ||\mathbf{u}_p|| > \lambda \end{cases}$$

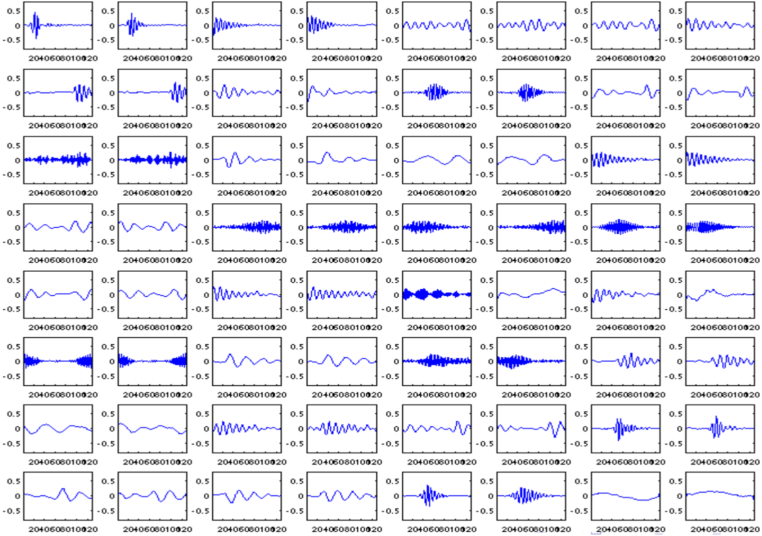
  $$\angle \mathbf{s}_p = \angle g^{-1}(\mathbf{u}_p) = \angle \mathbf{u}_p$$





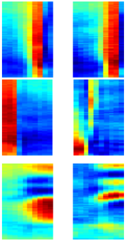
Result - Amplitude Waveform

Signals are 128 sample. Basis contains 256, 2-d subspaces (i.e., 4x overcomplete)



Result - Spectrograms

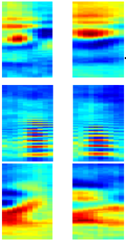
- 512 sample windowed FFT. 10 time steps with 50% overlap.
- The spectrogram is converted into log-frequency and log amplitude.
- The spectrogram is whitened.
- PCA is performed on the vectors and kept 90% of the variance.
- Trained subspace on whitened PCA data



Broadband Onset

Low frequency burst

Bandwidth Change



Time Frequency Shift

Harmonic Stack

Formant

Conclusion and Future Works

Conclusion

- Learned an overcomplete subspace model of natural sound.
- A new inference method was developed.
- Learned subspaces show shift and phase invariance.
- Some subspaces also show novel speech features such as formant invariance.

Future Directions

- Identify sounds/speech features that drive each subspace.
- Learning subspaces for a convolution model.
- Learning the appropriate dimension of the subspace for sound.

Acknowledgments

- Vivienne Ming, Bruno Olshausen