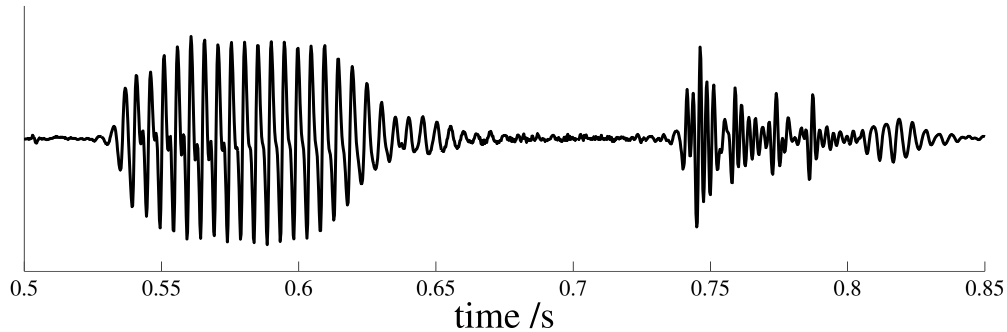
A new generative model for sounds: The Gaussian Modulation Cascade Process

Richard Turner (turner@gatsby.ucl.ac.uk)

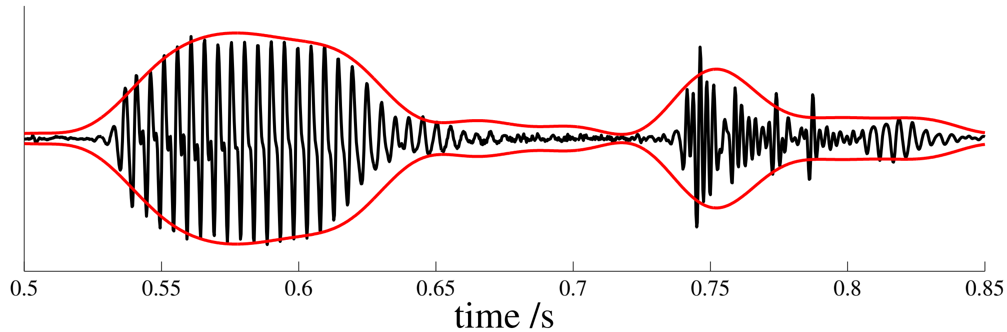
Maneesh Sahani (maneesh@gatsby.ucl.ac.uk)

Gatsby Computational Neuroscience Unit, 09/12/2006

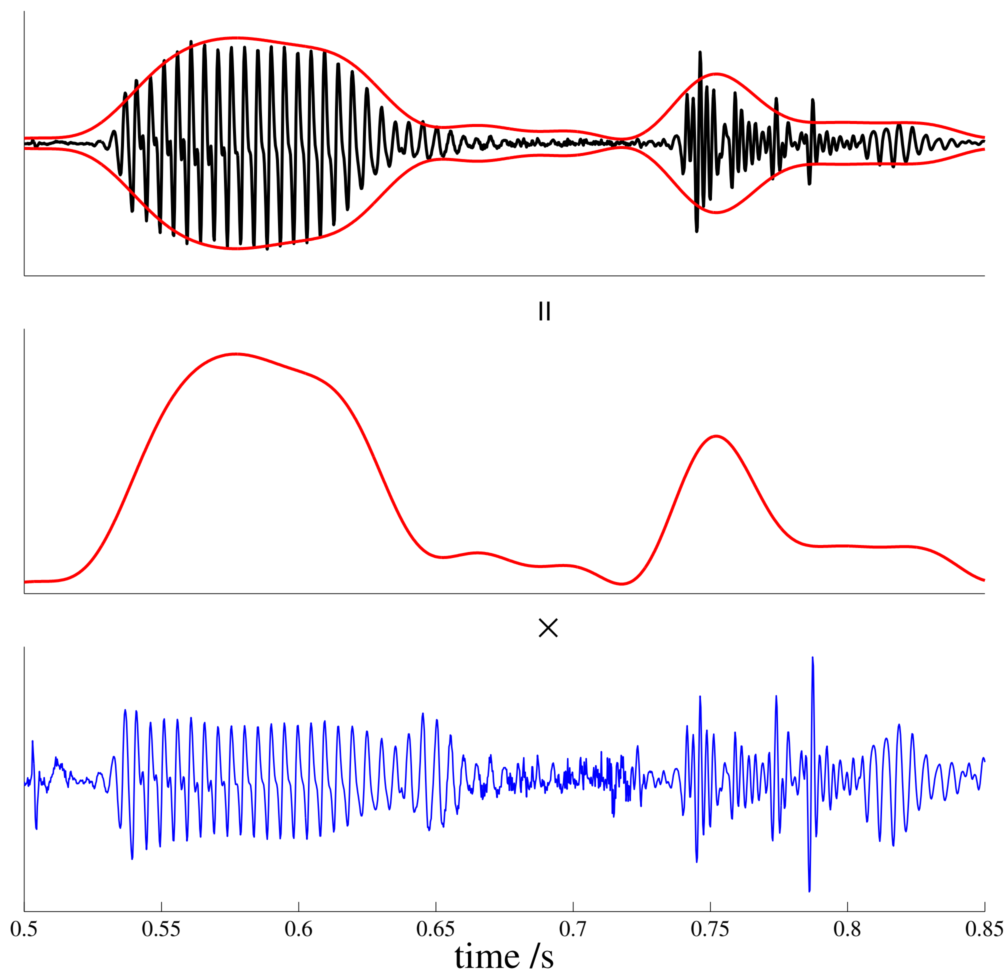
Motivation



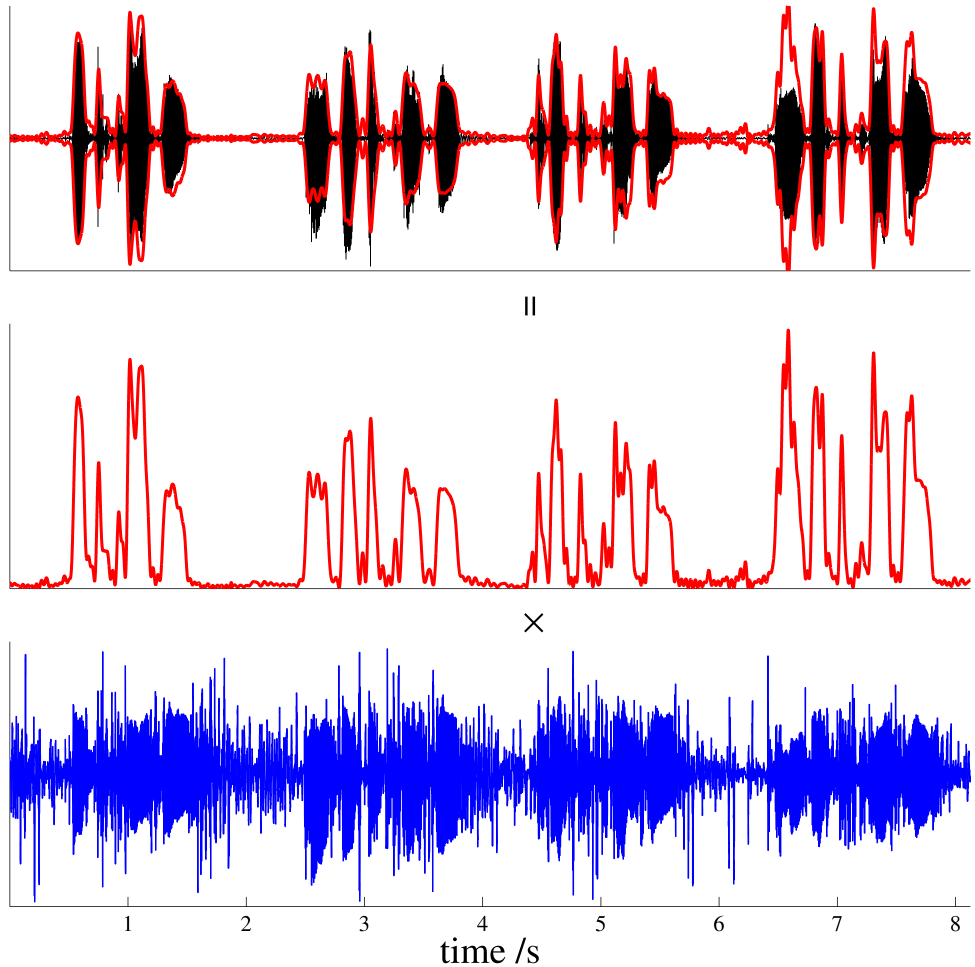
Motivation



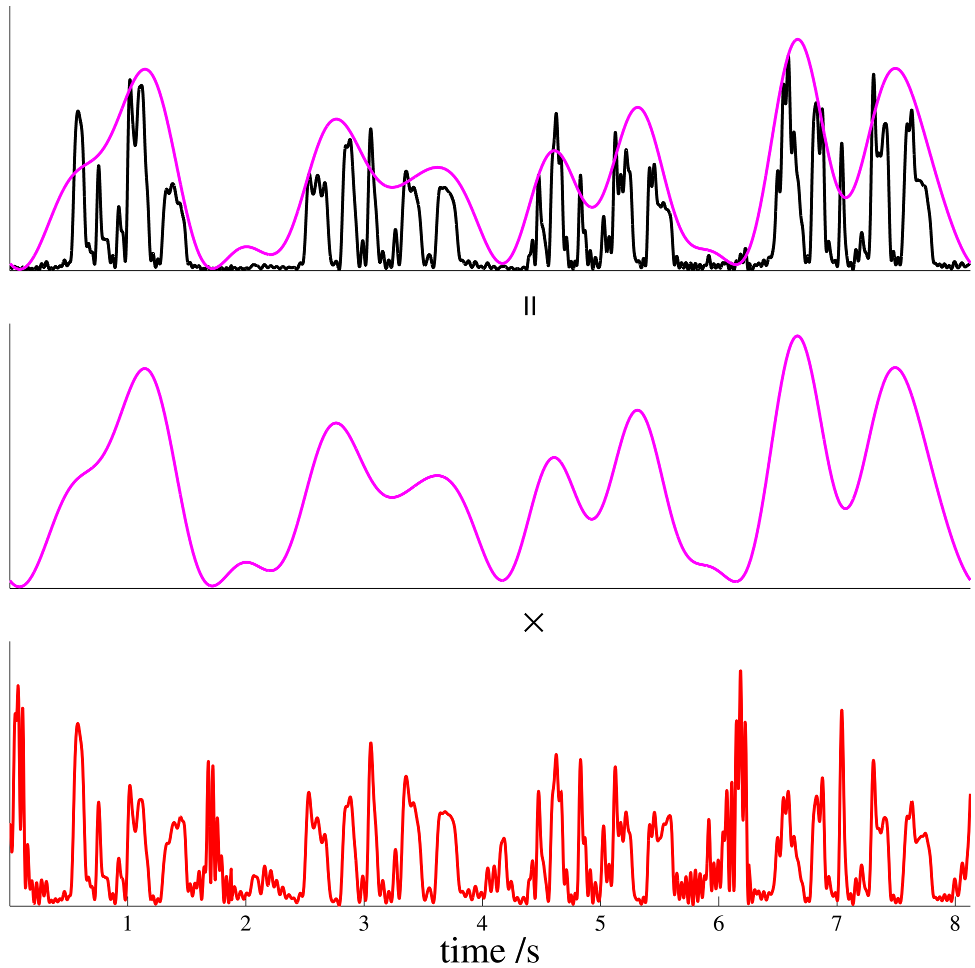
Motivation: Traditional AM



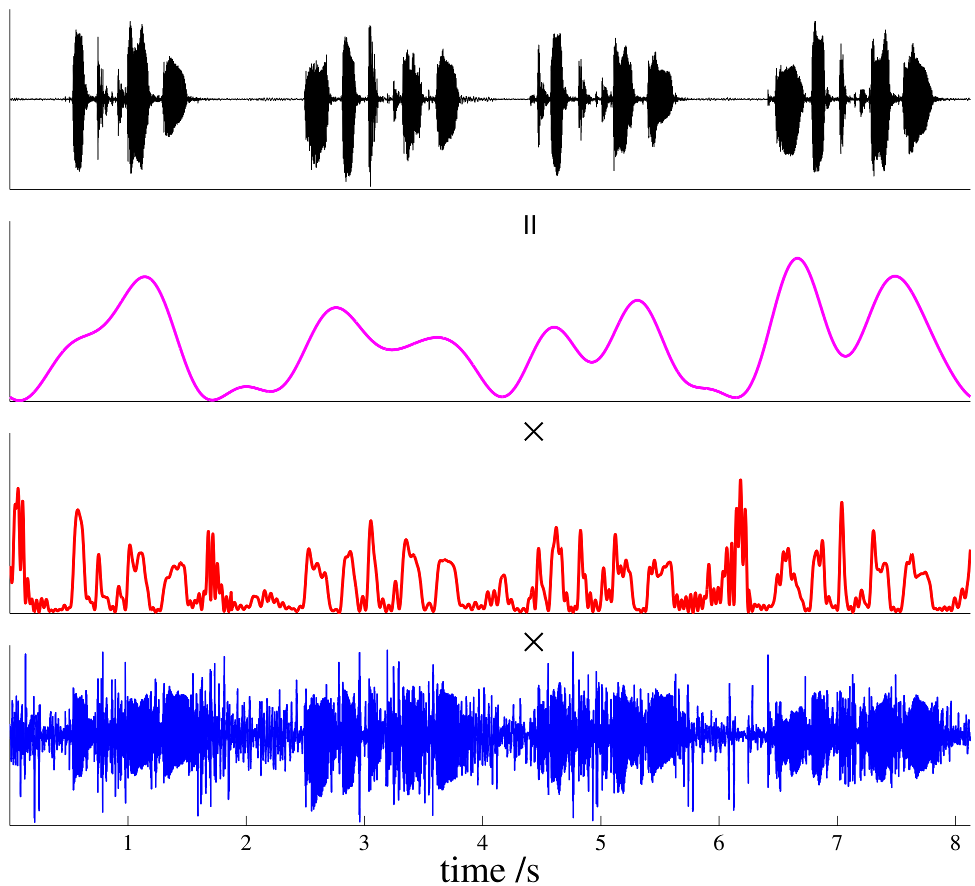
Motivation: Traditional AM



Motivation: Demodulate the Modulator

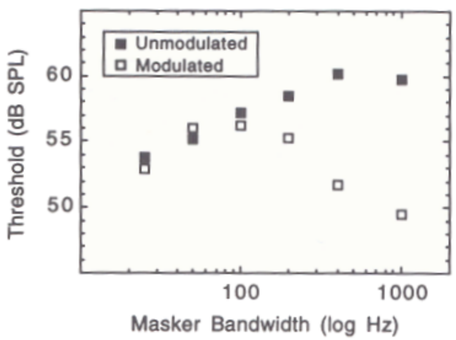


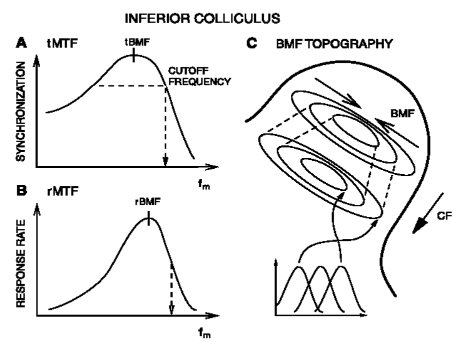
Motivation: Demodulation Cascade



AM: a candidate organising principle in the Auditory System?

- The auditory system listens attentively to Amplitude Modulation
- Examples include Comodulation Masking Release in psychophysics, and a possible topographic mapping of AM in the IC from electrophysiology.
- **Main goal**: discover computational principles underpinning auditory processing.
- But armed with a generative model you can do: sound denoising, source segregation, fill in missing data/remove artifacts etc.





What's up with current generative models?

Model	latents			learnable
	sparse	share power	slowly varying	
ICA	$\surd$	$\times$	$\times$	$\surd$
SC	$\surd$	$\times$	$\times$	$\surd$

Assumption: Latent variables are sparse.

What's up with current generative models?

Model	latents			learnable
	sparse	share power	slowly varying	
ICA	$\surd$	$\times$	$\times$	$\surd$
SC	$\surd$	$\times$	$\times$	$\surd$
GSM	$\surd$	$\surd$	$\times$	$\surd$

Assumption: Latents are sparse, and share power.

- $\mathbf{x} = \lambda \mathbf{u}$
  - $\lambda \geq 0$ a positive scalar random variable, $\mathbf{u} \sim G(0, Q)$

What's up with current generative models?

Model	latents			learnable
	sparse	share power	slowly varying	
ICA	$\surd$	$\times$	$\times$	$\surd$
SC	$\surd$	$\times$	$\times$	$\surd$
GSM	$\surd$	$\surd$	$\times$	$\surd$
SFA	$\times$	$\times$	$\surd$	$\surd$

Assumption: Latents are slow.

What's up with current generative models?

Model	latents			learnable
	sparse	share power	slowly varying	
ICA	$\surd$	$\times$	$\times$	$\surd$
SC	$\surd$	$\times$	$\times$	$\surd$
GSM	$\surd$	$\surd$	$\times$	$\surd$
SFA	$\times$	$\times$	$\surd$	$\surd$
Bubbles	$\surd$	$\surd$	$\surd$	$\times$

Assumption: Latents are sparse, slow (and share power).

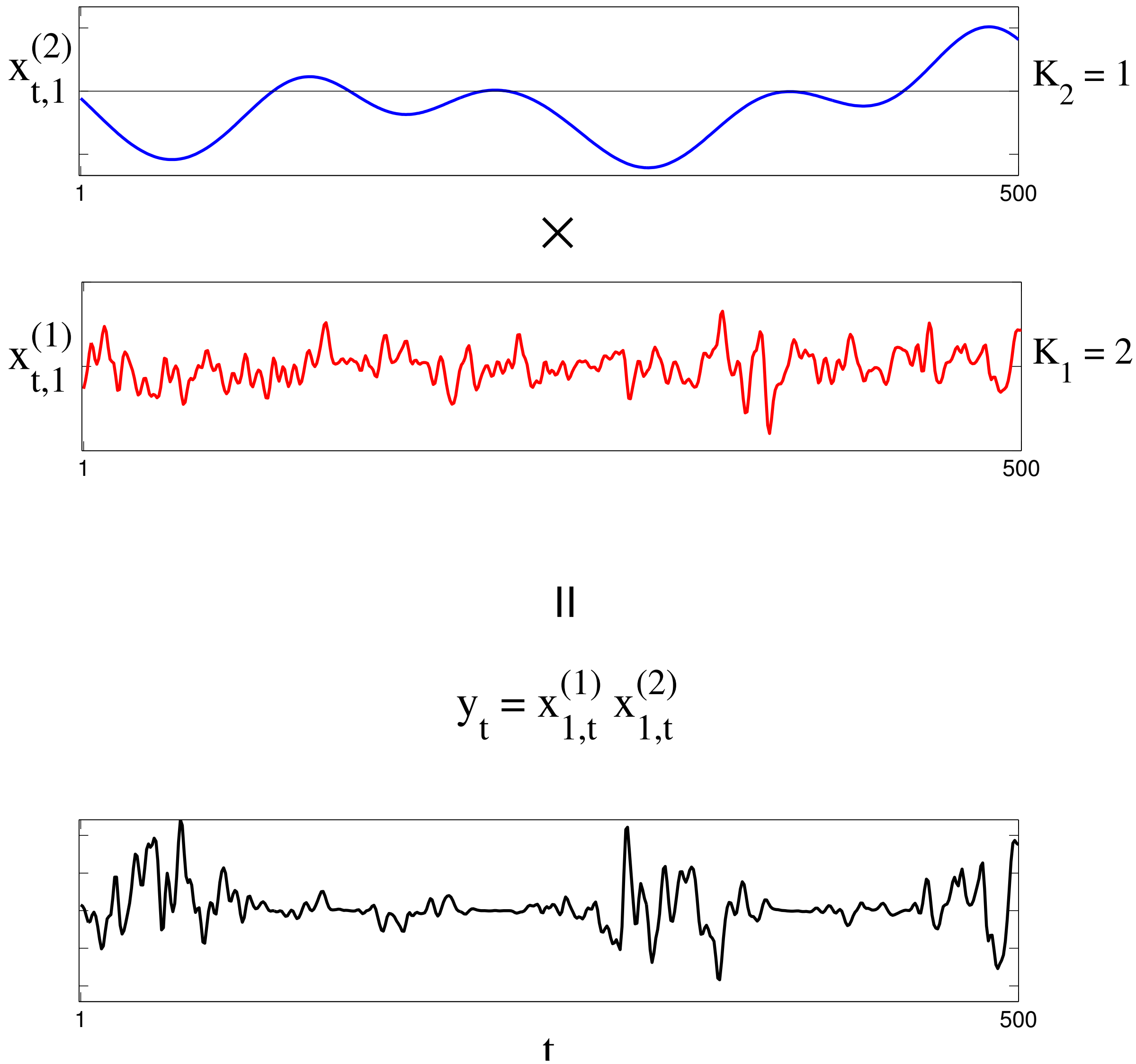
Desirable features of a new generative model

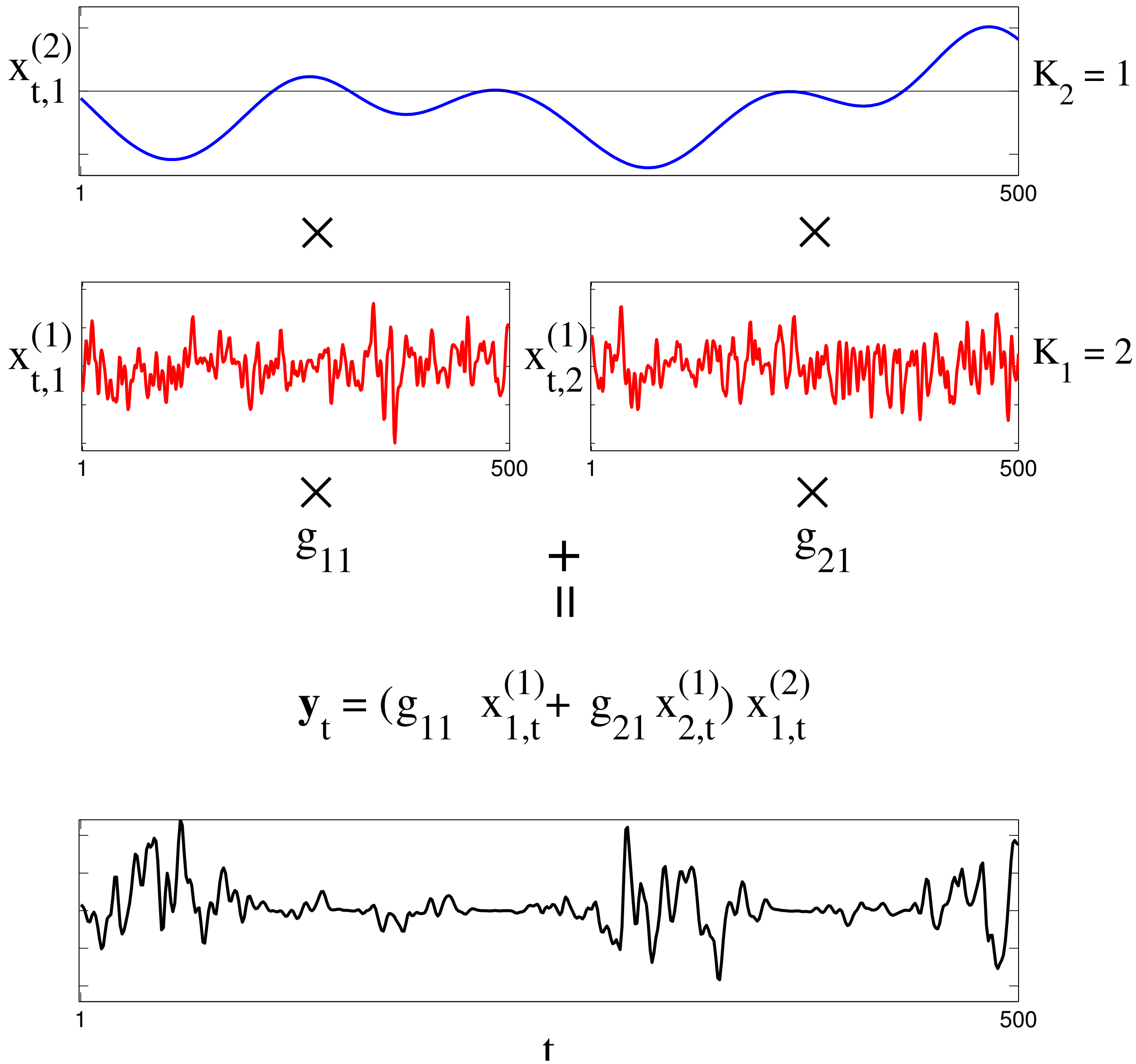
1. **Sparse outputs**.

2. **Explicit temporal dimension**, **smooth latent variables**.

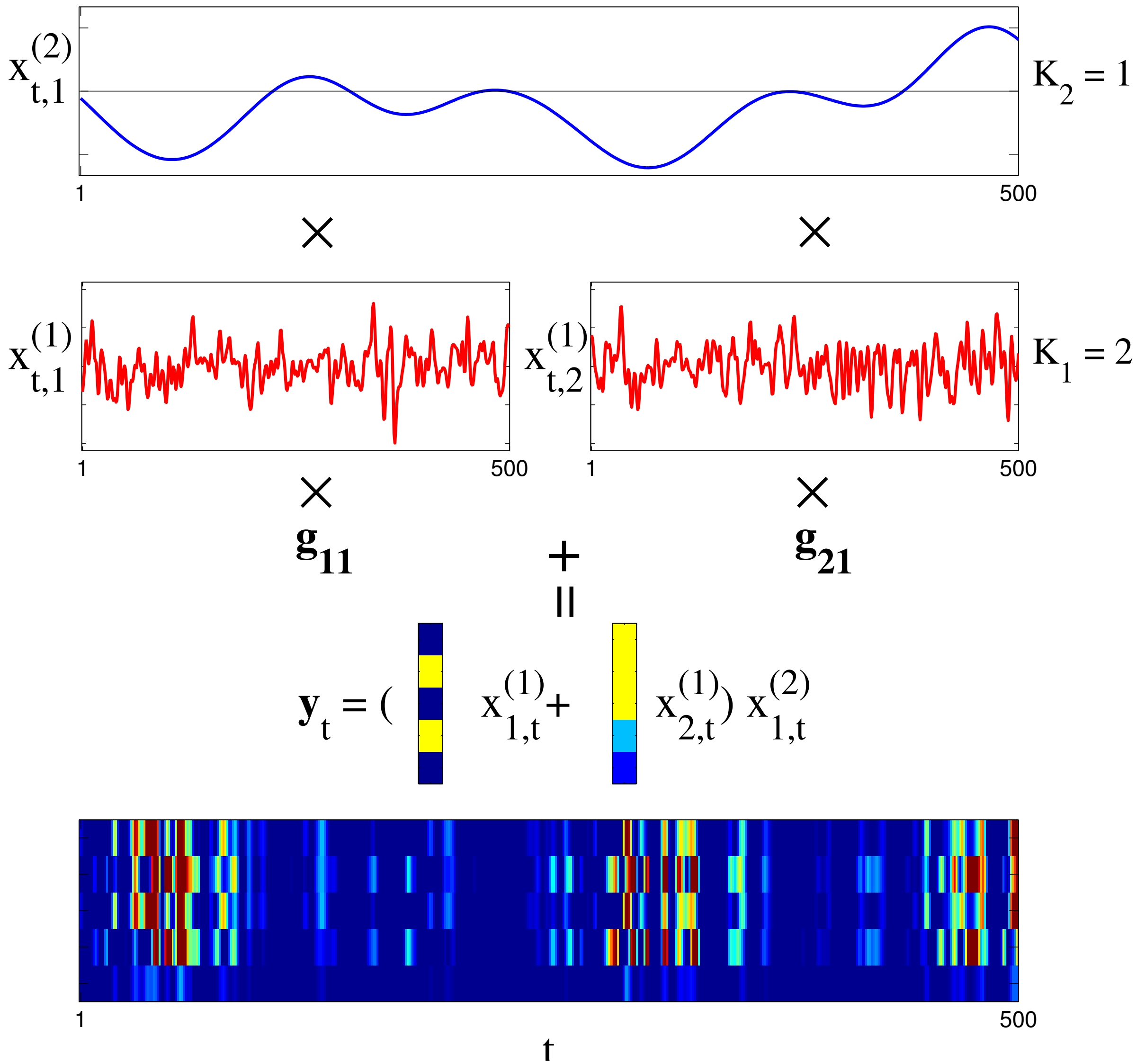
3. **Hierarchical prior** that captures the AM statistics of sounds at different time scales: **cascade of modulatory processes**, with slowly varying processes at the top modulating more rapidly varying signals towards the bottom.

4. **Learnable**; and we would like to **preserve information about the uncertainty**, and possibly correlations, in our inferences.

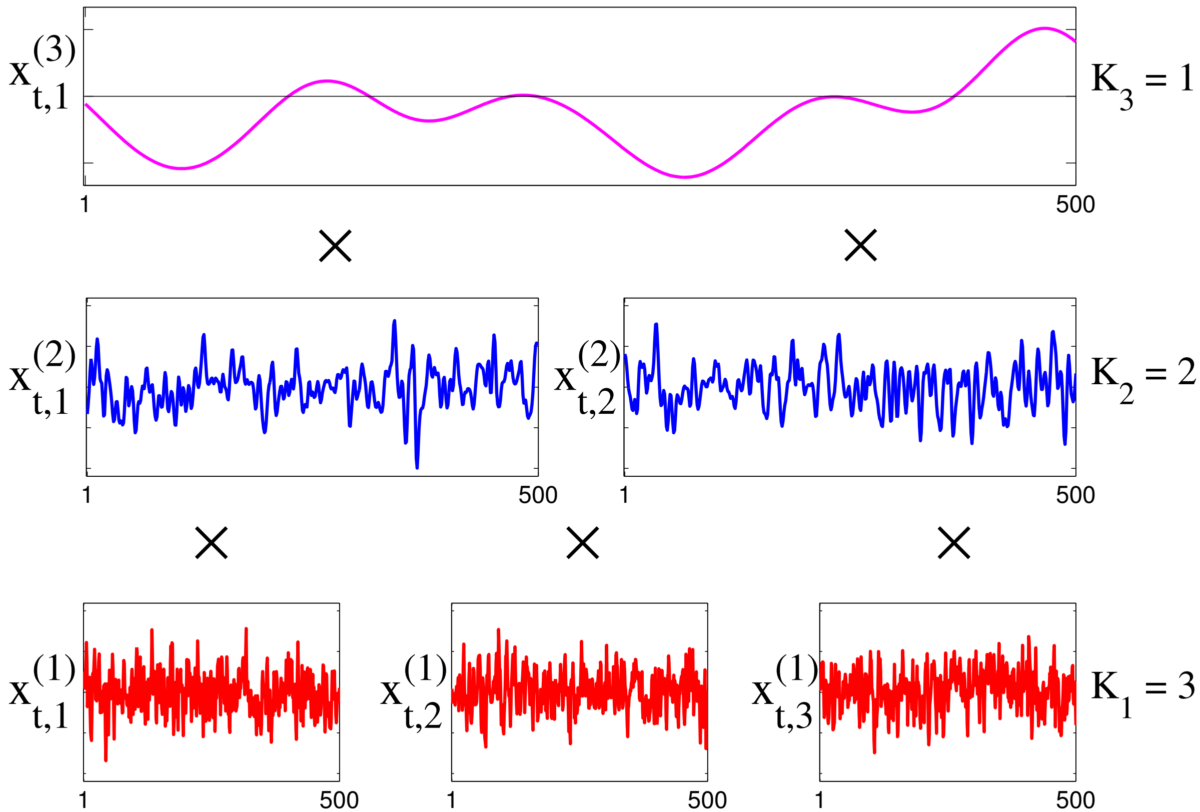

$x^{(2)}_{t,1}$
$K_2 = 1$
1
500
×
$x^{(1)}_{t,1}$
$K_1 = 2$
1
500
=
$y_t = x^{(1)}_{1,t}\, x^{(2)}_{1,t}$
1
500
t



$x^{(2)}_{t,1}$
$K_2 = 1$
1
500
$x^{(1)}_{t,1}$
$x^{(1)}_{t,2}$
$K_1 = 2$
1
500
1
500
g_{11}
g_{21}
$\mathbf{y}_t = (g_{11}\ x^{(1)}_{1,t} + g_{21}\ x^{(1)}_{2,t})\ x^{(2)}_{1,t}$
1
500
t

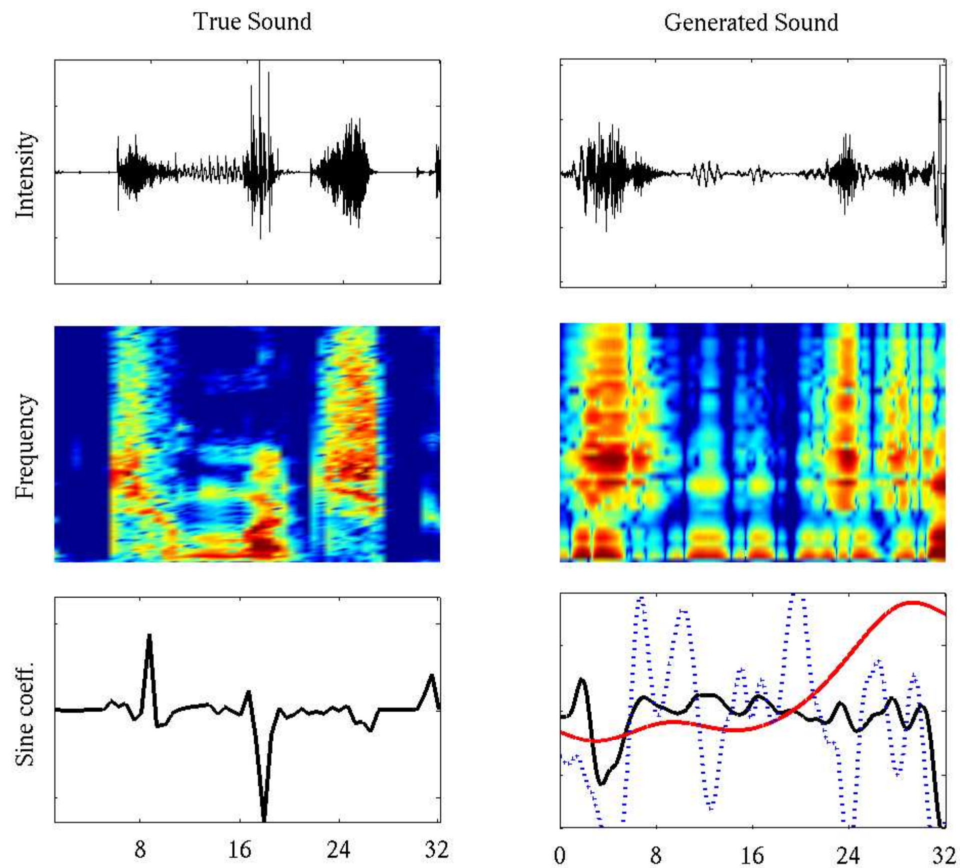


$x^{(2)}_{t,1}$
$K_2 = 1$
1
500
$x^{(1)}_{t,1}$
$x^{(1)}_{t,2}$
$K_1 = 2$
1
500
1
500
$\mathbf{g_{11}}$
+
$\mathbf{g_{21}}$
=
$\mathbf{y}_t = (\quad x^{(1)}_{1,t} + \quad x^{(1)}_{2,t}) \, x^{(2)}_{1,t}$
1
500
t



$$\mathbf{y}_t = x_{t,1}^{(3)} \left[x_{t,1}^{(2)} \left(x_{t,1}^{(1)}\ \mathbf{g}_{111} + x_{t,2}^{(1)}\ \mathbf{g}_{211} + x_{t,3}^{(1)}\ \mathbf{g}_{311}\right) + x_{t,2}^{(2)} \left(x_{t,1}^{(1)}\ \mathbf{g}_{121} + x_{t,2}^{(1)}\ \mathbf{g}_{221} + x_{t,3}^{(1)}\ \mathbf{g}_{321}\right) \right]$$

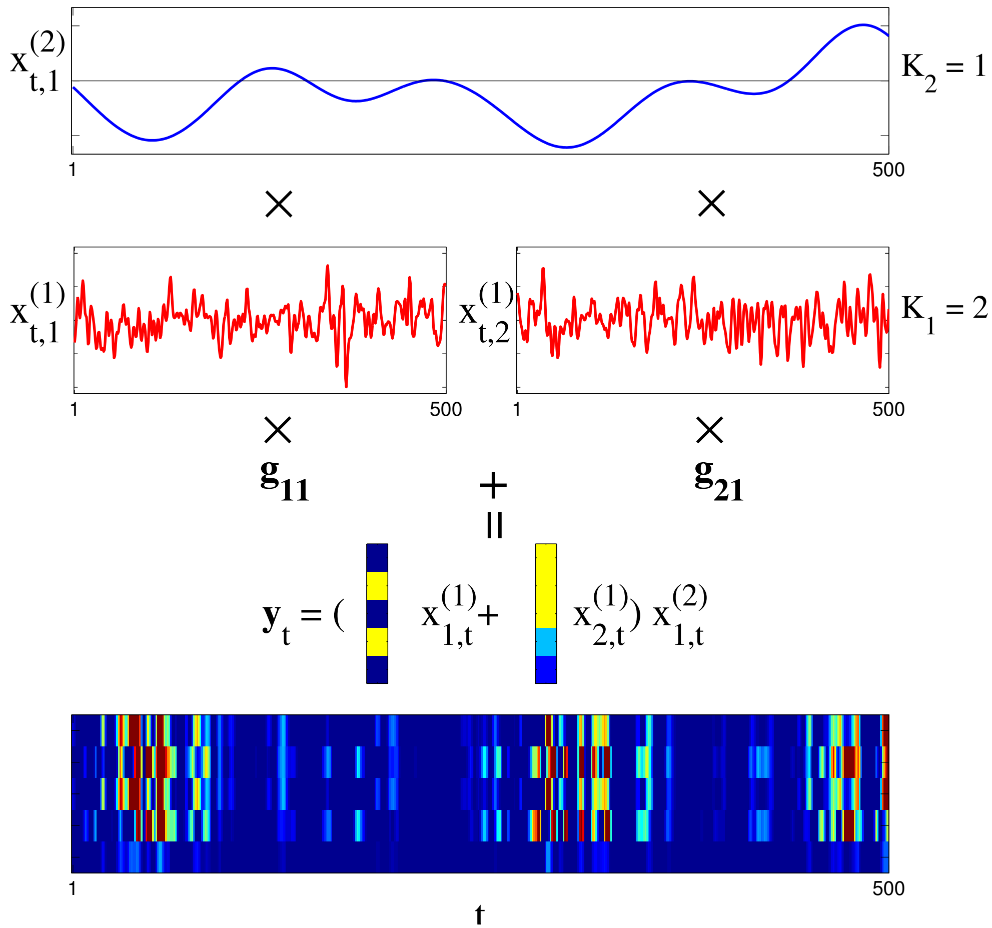
$$M = 4, K_1 = 4, K_2 = 2, K_3 = 2, K_4 = 1, D = 80$$



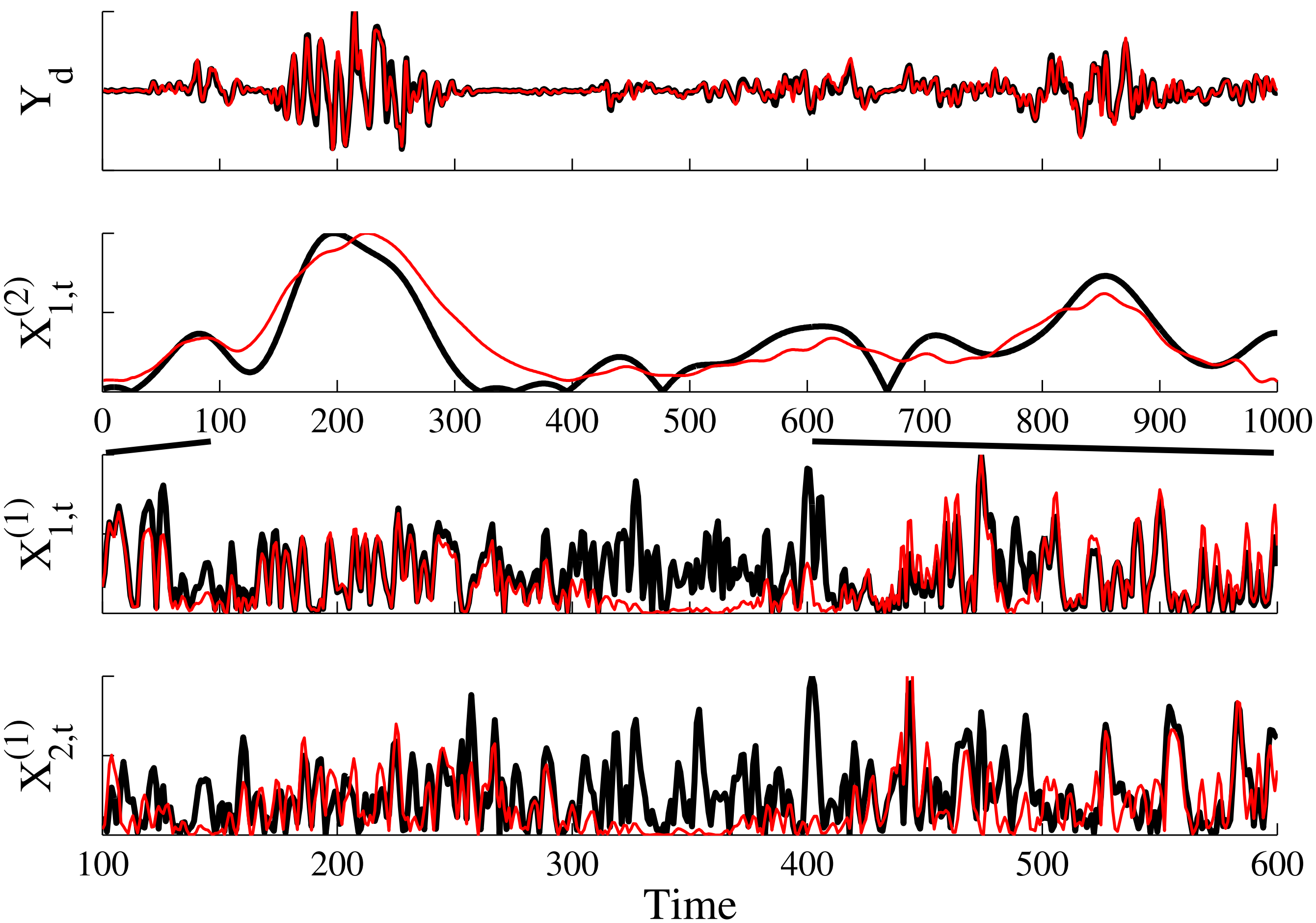
Inference and learning

- Inference and learning by EM has an **intractable E-Step**.
- But due to the structure of the non-linearity **variational EM** can be used.
- **Key point**: If we freeze all the latent time-series bar one, the distribution over the unfrozen chain is Gaussian.
- Leads to a **family of efficient variational approximations**: $p(X|Y) \approx \prod_m q(X^{(m)}_{1:T,1:K_m})$, where each approximating distribution is a Gaussian.

Inference and Learning: Proof of concept



Inference and Learning: Proof of concept



Current work and Future directions

- Apply to **natural sounds** (require good initialisation)

- **Generalise the model** to have

  - **non-local features**: to capture sweeps
  - **correlations in the prior**: to capture the mutual-exclusivity of voiced and unvoiced sections of speech

- **Representation**: real sounds live on a hyper-plane in STFT space: can we project our model onto this manifold?

- How do we **map posteriors to spike-trains**: $p(\text{spike trains}\,|\mathbf{y}) = f[Q(\mathbf{x}|\mathbf{y})]$?

Extra slides...

What's the point in a generative model for sounds?

Theoretical Neuroscience

- **Understand neural receptive fields**: if a latent variable model provides a computationally effective representation of sounds, neurons might encode p(latent|sound)

- **Psychophysics**: Play sounds to subjects (possibly drawn from the forward model), and compare their inferences with inference in the model.

Machine Learning

- **Fill in missing data**: e.g. to fill in an intermittent recording or remove artifacts

- **Denoise sounds, Stream segregation, Compression**

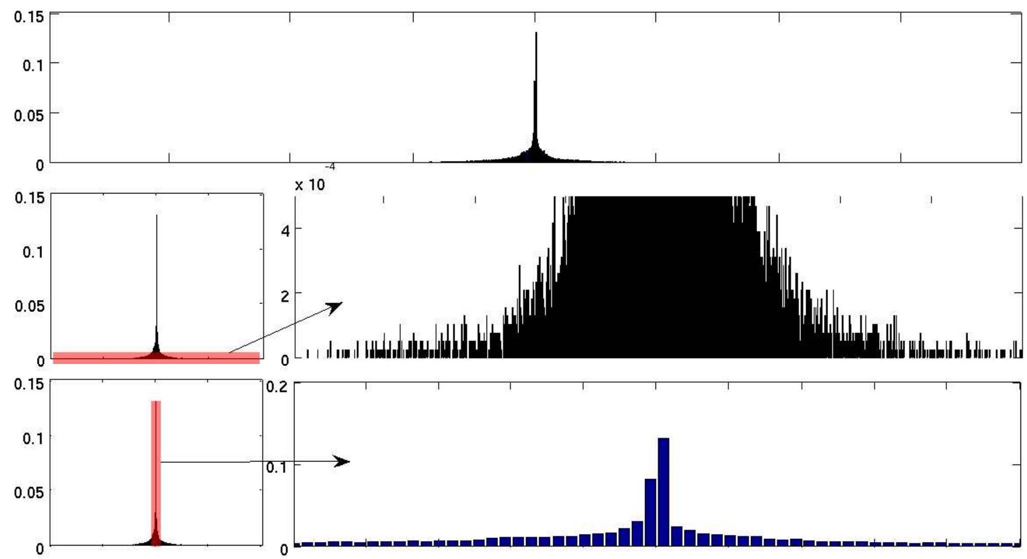
Outline

- **Natural auditory scene statistics** - Amplitude modulation is a key component
- **Amplitude modulation in the auditory system** - a candidate organising principle (the auditory system is short on such things)
- **Previous statistical models of natural scenes** - Gaussian Scale Mixture Models and AM
- **A new model for sounds: The Gaussian Modulation Cascade Process**
- On going issues...

Natural Auditory Scene statistics: Acoustic ecology

Marginal distribution

- very sparse (more so than in vision)



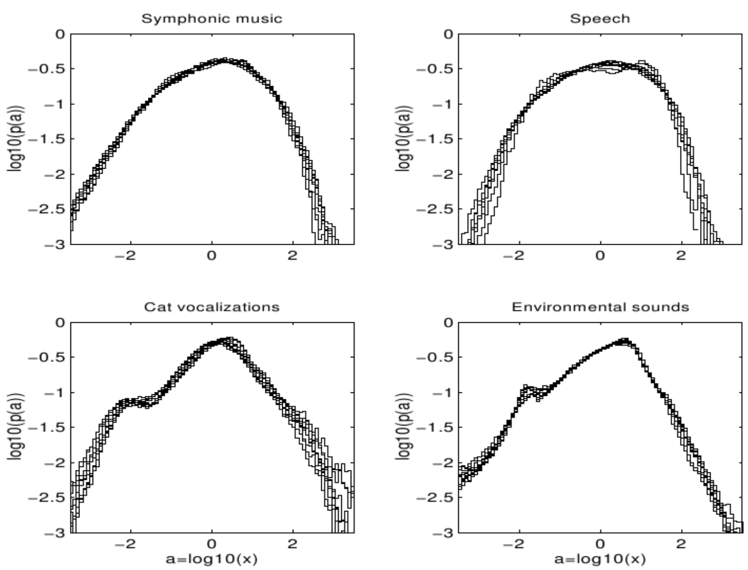
- WHY? Lots of soft sounds e.g. the pauses between utterances in speech sounds and rare highly structured localised events that carry substantial parts of the stimulus energy

Statistics of Amplitude Modulation (Attias & Schreiner)

Methods: sound $\rightarrow$ filterbank $\rightarrow$ Hilbert transform $\rightarrow$ find envelope $a(t,\omega)$ $\rightarrow$ take logs and transform to zero-mean and unit variance: $\hat{a}(t,\omega)$

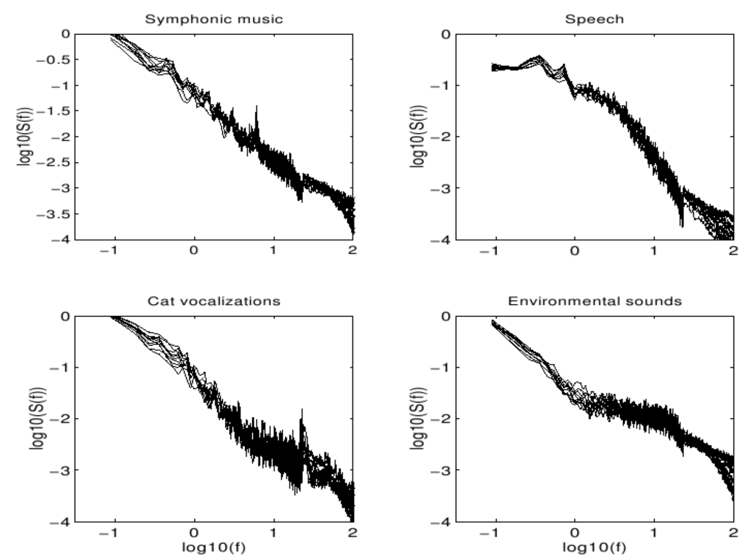
Marginal statistics of $\hat{a}(t,\omega)$

$$p[\hat{a}(t,\omega)] \propto \frac{\exp[-\gamma\hat{a}(t,\omega)]}{(\beta^2+\hat{a}(t,\omega)^2)^{\alpha/2}}$$



Spectrum of $\hat{a}(t,\omega)$

$$|FT[\hat{a}(t,\omega)]| \propto \frac{1}{(\omega_0^2+\omega^2)^{\alpha/2}}$$



Results summary

- **Marginal distribution**
  - very sparse (finite prob of arb. soft sounds):
  - Independent of filter centre frequency, filter bandwidth, time resolution.
  - If AM stats were uncorrelated over time and frequency, CLT would predict increasing the filter bandwidth/time resolution would make the distribution more Gaussian.
- **Spectrum of the amplitude modulations**
  - Independent of filter centre frequency
  - Modified power law, indicating long temporal correlations (scale invariance)
  - Independent of filter bandwidth (cf. Voss and Clarke $1/f$)

Implication: A good generative model of sounds should capture...

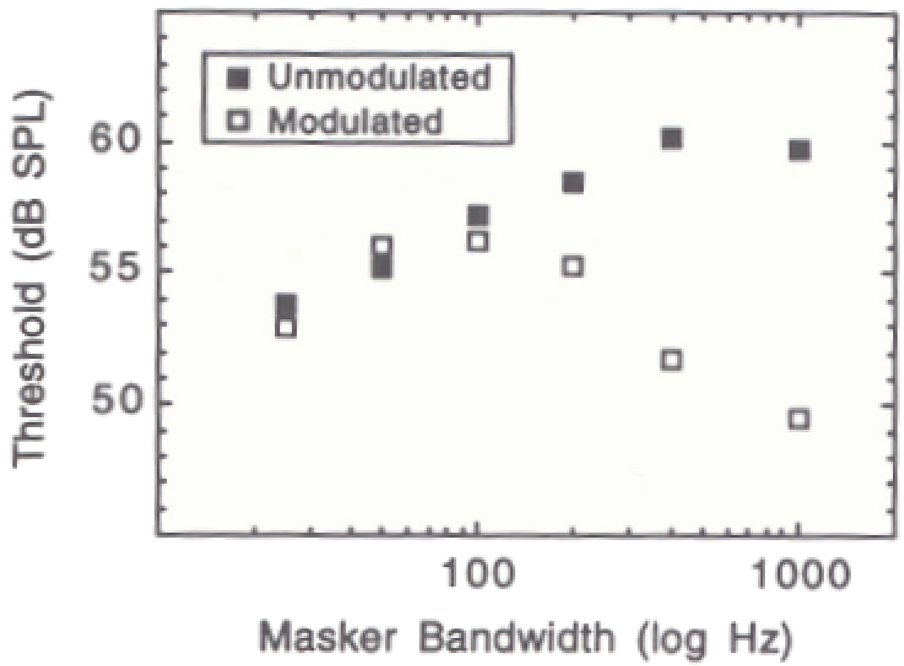
1. **long correlations in AM across frequency bands and time (> 100ms)**

2. **Each location on the cochlea sees the same AM stats**

AM in the Auditory System - Highlights from lots of work

Psychophysics: Comodulation masking release

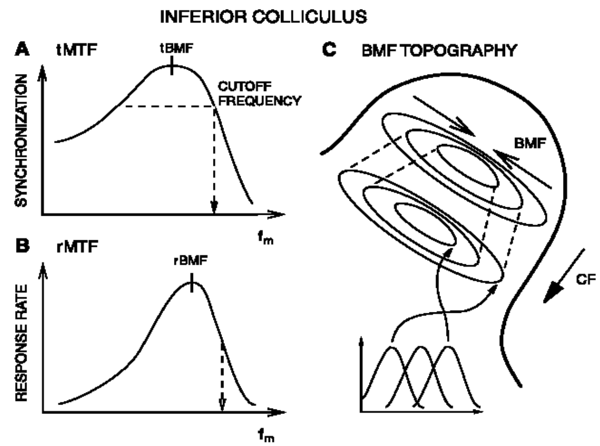
- task: detect a tone in noise.
- Alter the bandwidth of the noise and measure threshold
- Repeat but amplitude modulate the masker.



Electrophysiological data

- Type 1 AN fibres phase lock to AM (implies temporal code), as we move up the neuraxis the tuning moves from temporal to rate.

- Evidence in IC for topographic AM tuning (Schreiner and Langer, 1988)



- Cortex: AM processing seems $\sim$ filter independent (modulation filter bank?)

- **Jury still out but AM may be a fundamental organising principle**

What's wrong with statistical models for natural scenes?

ICA and Sparse coding

$$p(\mathbf{x}) = \prod_{i=1}^{I} p(x_i) \quad (1)$$

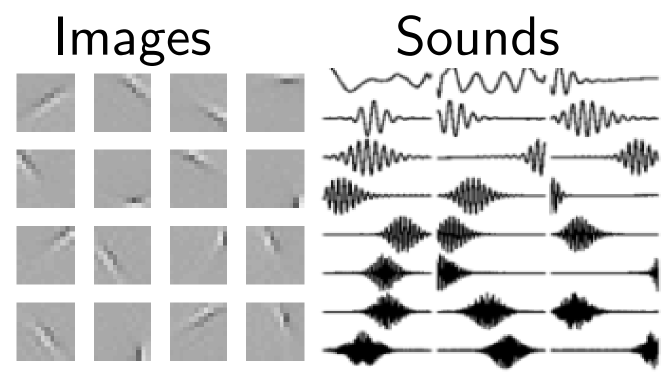
$$p(x_i) = \text{sparse} \quad (2)$$

$$p_{ICA}(\mathbf{y}|\mathbf{x}) = \delta(G\mathbf{x} - \mathbf{y}) \quad (3)$$

$$p_{SC}(\mathbf{y}|\mathbf{x}) = \text{Norm}(G\mathbf{x}, \sigma^2 I) \quad (4)$$

Recognition weights:

$$p_{ICA}(\mathbf{x}|\mathbf{y}) = \delta(\mathbf{x} - R\mathbf{y})$$

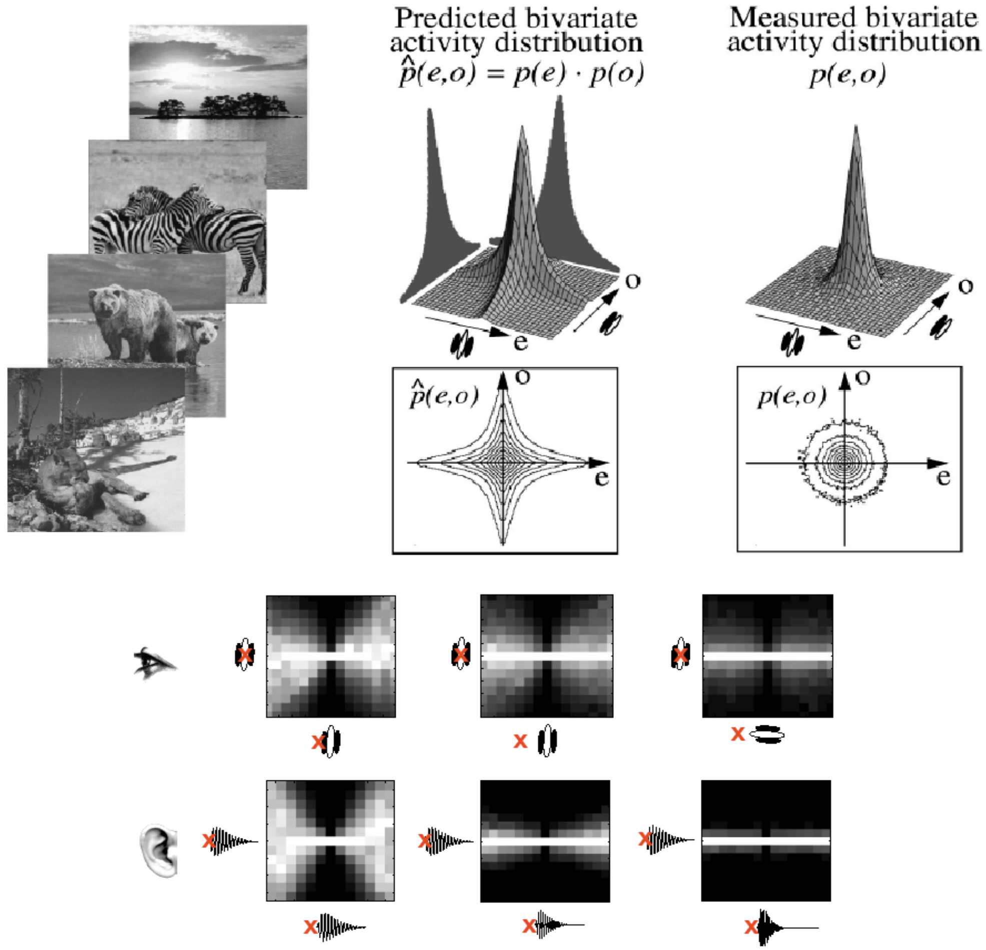


Problems:

1. **Extracted latents are not independent: correlations in their power.**

2. **No explicit temporal dimension - not a true generative model for sounds or movies**

1. Empirical distribution of the latents - power correlations



Explanation

- Caption for previous figure:
  - Expected joint distribution of latents was starry (top left).
  - Empirically the joint is found to be elliptical - many more instances of two latents being high than expected (top right).
  - Another way of seeing this is to look at the conditionals (bottom): if one latent has high power then near by latents also tend to have high power.

- **How do we improve the model?**

1. Fix the bottom level $p(\mathbf{y}|\mathbf{x}) = \delta(\mathbf{y} - G\mathbf{x})$
2. Fixes recognition distribution $p(\mathbf{x}|\mathbf{y}) = \delta(\mathbf{x} - R\mathbf{y})$
3. $p(\mathbf{x}) = \int d\mathbf{y} p(\mathbf{x}|\mathbf{y}) p(\mathbf{y}) \approx \frac{1}{N} \sum_n p(\mathbf{x}|\mathbf{y}_n)$
4. These are exactly the distributions plotted on the previous page
5. No matter what R is we get similar empirical distributions
6. So choose a new prior to match them ...

Gaussian Scale Mixtures (GSMs)

- $\mathbf{x} = \lambda \mathbf{u}$
  - $\lambda \geq 0$ a scalar random variable
  - $\mathbf{u} \sim G(0, Q)$
  - λ and $\mathbf{u}$ are independent

- density of these *semi-parametric* models can be expressed as an integral:

$$p(\mathbf{x}) = \int p(\mathbf{x}|\lambda)p(\lambda)d\lambda = \int |2\pi\lambda^2 Q|^{-1/2} \exp\left(-\frac{\mathbf{x}^T Q^{-1}\mathbf{x}}{2\lambda^2}\right) p(\lambda)d\lambda \qquad (5)$$

- e.g. $p(\lambda)$ discrete $\rightarrow$ MOG (components 0 mean), $p(\lambda) =$Gamma $\rightarrow$ student-T.

Imagine generalising this to the temporal setting:

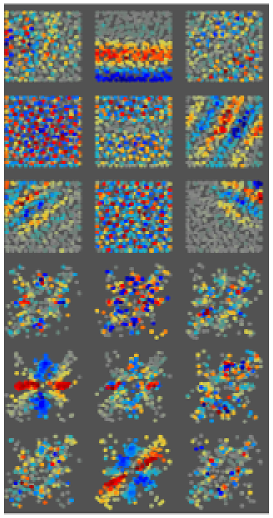
$$\mathbf{x}(t) = \lambda(t)\mathbf{u}(t) = \text{positive envelope} \times \text{carrier}$$

Are we seeing the hall marks of AM in a non-temporal setting?

Learning a neighbourhood (Karklin and Lewicki 2003, 05, 06)

- The statistical dependencies between filters depends on their separation (in space, scale, and orientation.)
- We'd like to learn these neighbourhoods of dependence
- Solution: Share multipliers in a linear fashion using a GSM with a generalised log-normal over the variances.

$$
\begin{aligned}
p(z_j) &= \mathrm{Norm}(0,1) && (6)\\
\lambda_i^2 &= \exp\left[\sum_j h_{ij} z_j\right] && (7)\\
p(x_i|\mathbf{z},B) &= \mathrm{Norm}(0,\lambda_i) && (8)\\
p(\mathbf{y}|G,\mathbf{x}) &= \mathrm{Norm}(G\mathbf{x},\sigma^2 I) && (9)
\end{aligned}
$$



A temporal GSMM: The proof of concept Bubbles model

$$
\begin{aligned}
p(z_{i,t}) &= \text{sparse point process} && (10)\\
\lambda^2_{i,t} &= f\left[\sum_j h_{ij}\Phi(t)\otimes z_j(t)\right] && (11)\\
p(x_{i,t}|\lambda_{i,t}) &= \mathrm{Norm}(0,\lambda^2_{i,t}) && (12)\\
p(\mathbf{y}_t|G,\mathbf{x}_t) &= \delta(G\mathbf{x}_t-\mathbf{y}_t) && (13)
\end{aligned}
$$

- **Temporal correlations** between the multipliers are captured by the moving average $\Phi(t)\otimes u_j(t)$ (creates a **SLOW ENVELOPE**)
- $\Rightarrow$ **Bubbles** of activity in latent space (both in space and time)
- Columns of $h_{i,j}$ fixed and change smoothly to induce **topographic structure** - **computationally useful**

Learning

- Common to learn the parameters using zero-temperature EM:

$$q(X) = \delta(X - X_{MAP}) \approx p(X|Y) \tag{14}$$

- uncertainty and correlational information is lost.

1. Effects learning
2. To compare to neural data need to specify a mapping: $p(\text{spike trains}\,|\mathbf{y}) = f[q(\mathbf{x}|\mathbf{y})] = f(X_{MAP})$ for this approximation.
3. BUT we believe neural populations will **represent uncertainty and correlations** in latent variables.

We'd like to retain variance and correlational information, both for learning and for comparison to biology.

Motivations for the GMCP

1. **Sparse outputs**.
2. **Explicit temporal dimension**, smooth latent variables.
3. **Hierarchical prior** that captures the AM statistics of sounds at different time scales: cascade of modulatory processes, with slowly varying processes at the top modulating more rapidly varying signals towards the bottom.
4. **Learnable**; and we would like to preserve information about the uncertainty, and possibly correlations, in our inferences.

The Gaussian Modulation Cascade Process

Dynamics:

$$p(x_{k,t}^{(m)}|x_{k,t-1:t-\tau_m}^{(m)}, \lambda_{k,1:\tau_m}^{m}, \sigma_{m,k}^2) = \mathrm{Norm}\left(\sum_{t'=1}^{\tau_m} \lambda_{k,t'}^{(m)} x_{k,t-t'}^{(m)}, \sigma_{m,k}^2\right)$$

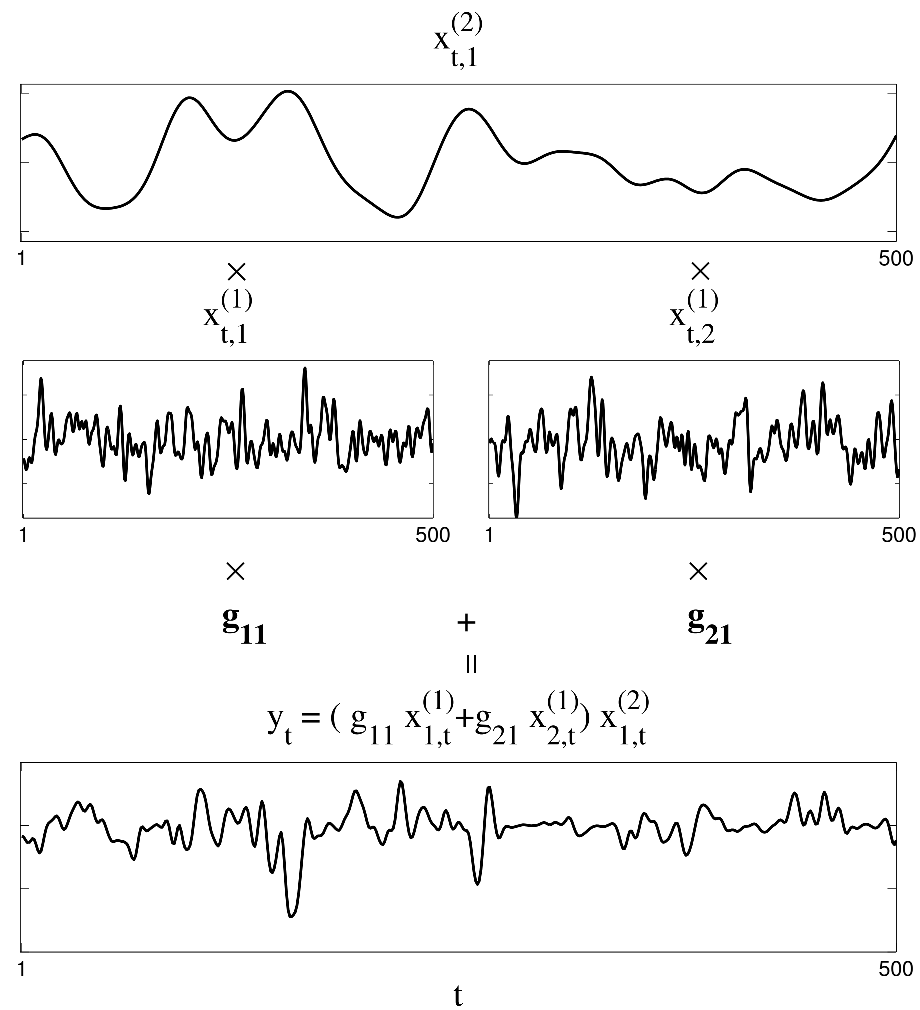
Emission Distribution:

$$p(\mathbf{y}_t|\mathbf{x}_t^{(1:M)}, \mathbf{g}_{k_1:k_M}, \sigma_y^2) = \mathrm{Norm}\left(\sum_{k_1:k_M} \mathbf{g}_{k_1:k_M} \prod_{m=1}^{M} x_{k_m,t}^{(m)}, \sigma_y^2 I\right)$$

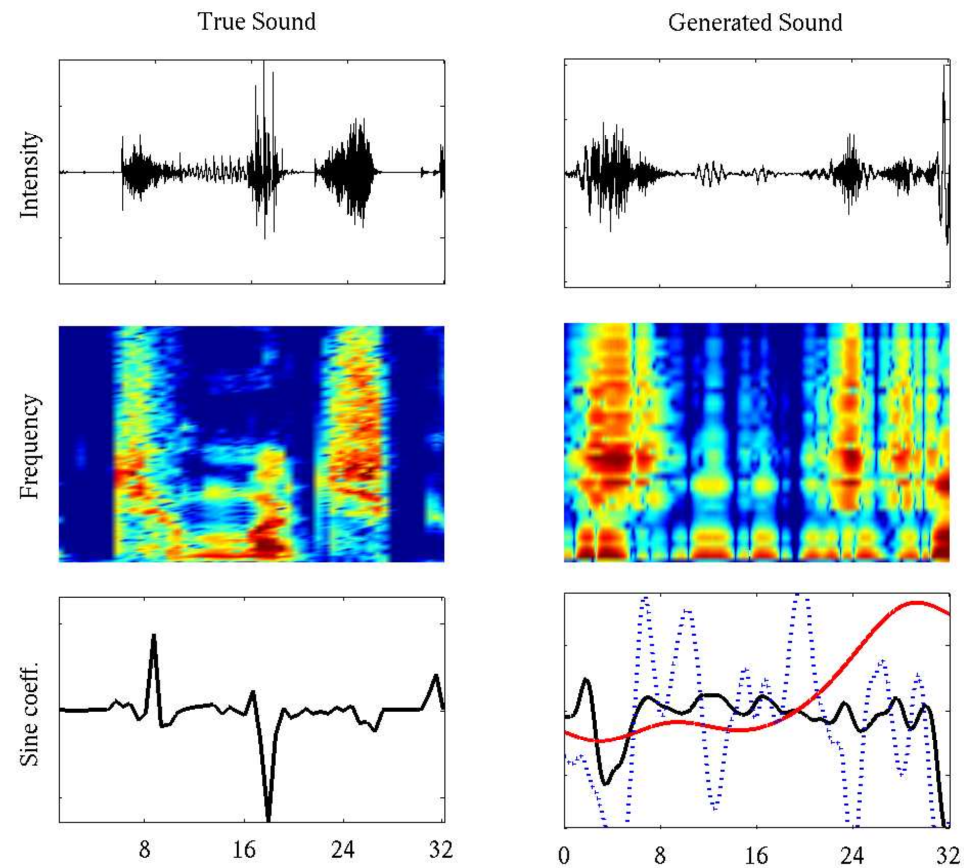
Time-frequency representation:

$\mathbf{y}_t =$ filter bank outputs

e.g. $M = 2, K_1 = 2, K_2 = 1, D = 1$



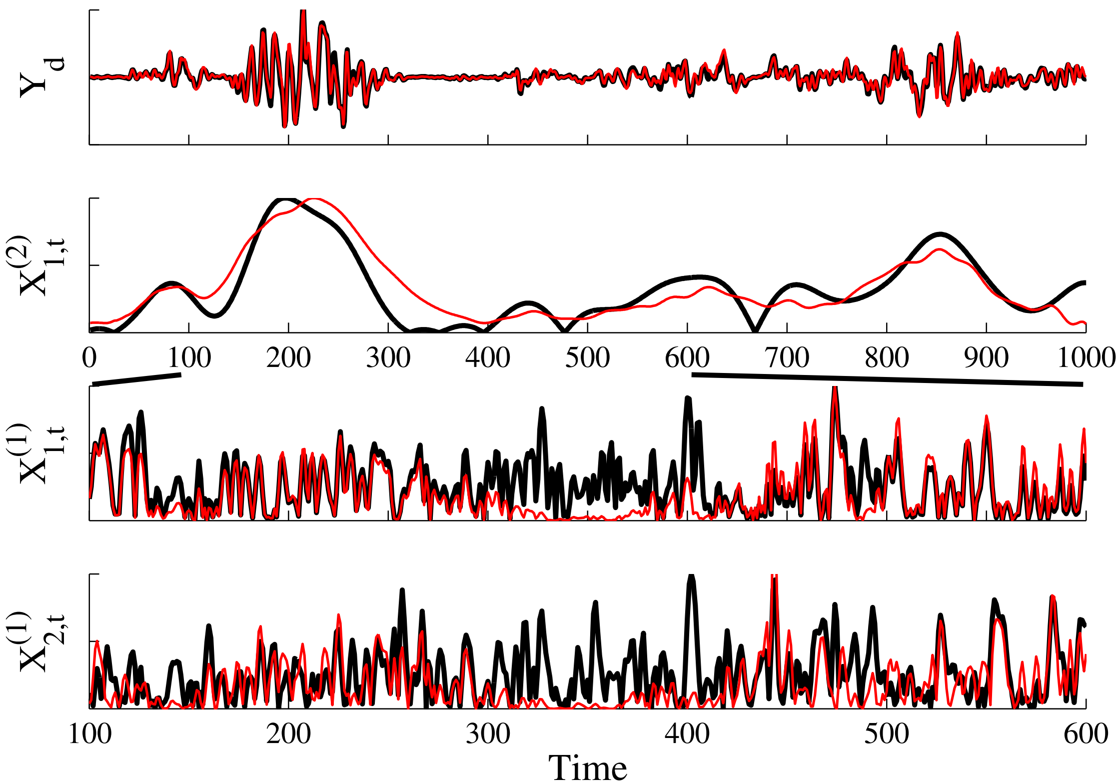
e.g. $M = 4, K_1 = 4, K_2 = 2, K_3 = 2, K_4 = 1, D = 80$



Comments on the model

- A GSM: latent $=$ rectified Gaussian $\times$ Gaussian
- Emission distribution related to Grimes & Rao, Tenenbaum & Freeman (for M=2) and Vasilescu & Terzopoulos for general M., but the temporal priors allow us to be fully unsupervised.
- As $p(x^m_{1:K,1:T}|x^{1:M\neq m}_{1:K,1:T}, \mathbf{y}_{1:T}, \theta)$ is Gaussian, there are a family of variational EM algorithms we can use to learn the model.

Inference and Learning: Proof of concept



Future directions and current issues

DIRECTIONS

- **correlated prior**: speech is usually periodic (voiced) or unvoiced - mutually exclusive, easier to represent sweeps.

- **non-local features**: make G temporal to capture sweeps

ISSUES

- **Representation**: real sounds live on a hyper-plane in filter-bank space - can we project our model onto this manifold?

- **Initialisation**: The free energy has lots of local minima: use Slow modulations analysis to initialise $\arg\min_{\mathbf{g}_n} \left\langle \left[\frac{da(\mathbf{g}_n^T \mathbf{y})}{dt}\right]^2 \right\rangle$ such that $\langle a(\mathbf{g}_n^T \mathbf{y}) a(\mathbf{g}_m^T \mathbf{y}) \rangle = \delta_{mn}$

- How do we **map posteriors to spike-trains**: $p(\text{spike trains}\,|\mathbf{y}) = f[Q(\mathbf{x}|\mathbf{y})]$?

What's up with current generative models?

Model		$\mathbf{p(x^{(1)})}$	$\mathbf{p(y \mid x^{(1)})}$
ICA		sparse	$\delta(y - Gx^{(1)})$
SC		sparse	$\mathrm{Norm}(Gx^{(1)}, \sigma_y^2 I)$

Assumption: Latent variables are sparse.

What's up with current generative models?

Model	$\mathbf{p(x^{(2)})}$	$\mathbf{p(x^{(1)}\|x^{(2)})}$	$\mathbf{p(y\|x^{(1)})}$
ICA	N/A	sparse	$\delta(y - Gx^{(1)})$
SC	N/A	sparse	$\mathrm{Norm}(Gx^{(1)}, \sigma_y^2 I)$
GSM	$\mathrm{Norm}(0, I)$	$\mathrm{Norm}(0, \lambda_i^2)$ $\lambda_i^2 = \exp(h_i^T x^{(2)})$	$\mathrm{Norm}(Gx^{(1)}, \sigma_y^2 I)$

Assumption: Latents are sparse, and share power.

- $\mathbf{x}^{(1)} = \lambda \mathbf{u}$
  - $\lambda \geq 0$ a positive scalar random variable, $\mathbf{u} \sim G(0, Q)$

What's up with current generative models?

Model	$\mathbf{p}(\mathbf{x}^{(2)})$	$\mathbf{p}(\mathbf{x}^{(1)}\vert\mathbf{x}^{(2)})$	$\mathbf{p}(\mathbf{y}\vert\mathbf{x}^{(1)})$
ICA	N/A	sparse	$\delta(y - Gx^{(1)})$
SC	N/A	sparse	$\mathrm{Norm}(Gx^{(1)}, \sigma_y^2 I)$
GSM	$\mathrm{Norm}(0, I)$	$\mathrm{Norm}(0, \lambda_i^2)$ $\lambda_i^2 = \exp(h_i^T x^{(2)})$	$\mathrm{Norm}(Gx^{(1)}, \sigma_y^2 I)$
SFA	N/A	$\mathrm{Norm}(\gamma x_{t-1}, \sigma^2)$	$\delta(y - Gx^{(1)})$

Assumption: Latents are slow.

What's up with current generative models?

Model	$\mathbf{p(x^{(2)})}$	$\mathbf{p(x^{(1)}\vert x^{(2)})}$	$\mathbf{p(y\vert x^{(1)})}$
ICA	N/A	sparse	$\delta(y - Gx^{(1)})$
SC	N/A	sparse	$\mathrm{Norm}(Gx^{(1)}, \sigma_y^2 I)$
GSM	$\mathrm{Norm}(0, I)$	$\mathrm{Norm}(0, \lambda_i^2)$ $\lambda_i^2 = \exp(h_i^T x^{(2)})$	$\mathrm{Norm}(Gx^{(1)}, \sigma_y^2 I)$
SFA	N/A	$\mathrm{Norm}(\gamma x_{t-1}, \sigma^2)$	$\delta(y - Gx^{(1)})$
Bubbles	point-process	$\mathrm{Norm}(0, \lambda_{i,t}^2)$ $\lambda_{i,t}^2 = f(h_i^T x_t^{(2)} \otimes \phi_t)$	$\delta(y - Gx^{(1)})$

Assumption: Latents are sparse, slow (and share power).