

Probabilistic auditory scene analysis from natural statistics

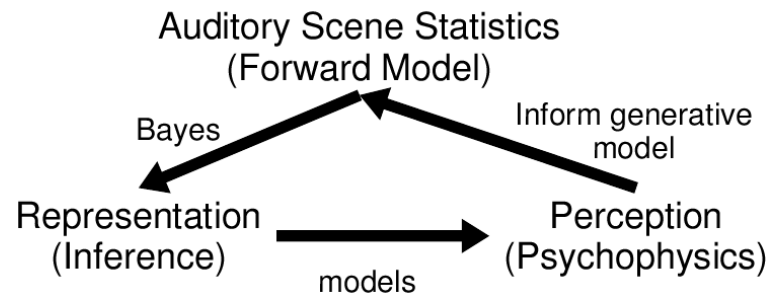
Richard Turner (turner@gatsby.ucl.ac.uk)

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

Gatsby Computational Neuroscience Unit, UCL, London

Introduction

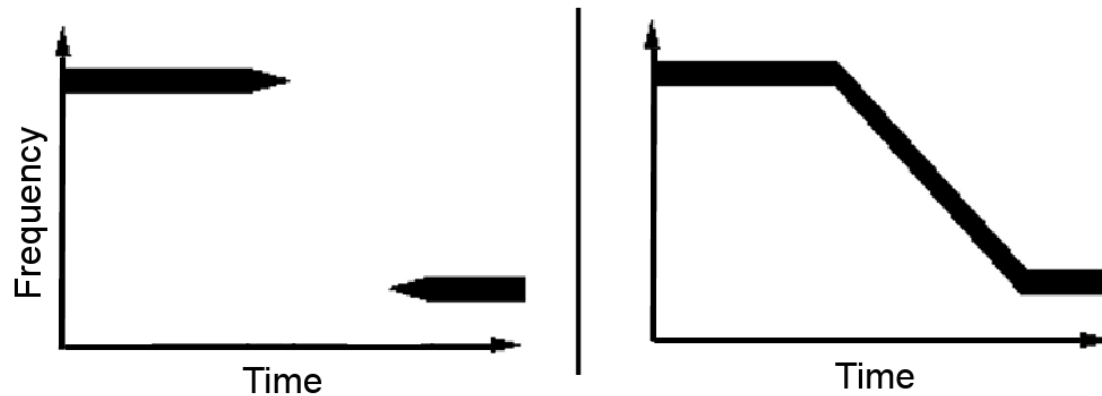
- The proverbial cocktail party problem: **How are auditory features grouped into auditory objects ?**
- **Gestalt Psychology** proposes a list of rules
- **Computational Auditory Scene Analysis** uses these rules as heuristics for source segregation
- **GOAL:** **Develop a generative model of auditory scene statistics, in which inference reproduces the Gestalt rules**



Grouping principle 1: Good continuation

Psychophysics

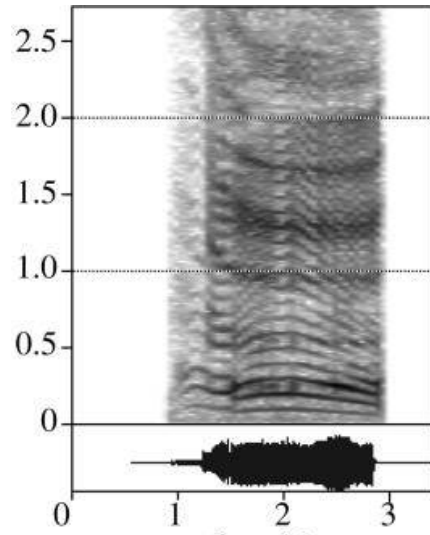
- Alternating tone pips are heard as two separate objects
- Joining the pips by frequency modulated tones binds the tones to one object



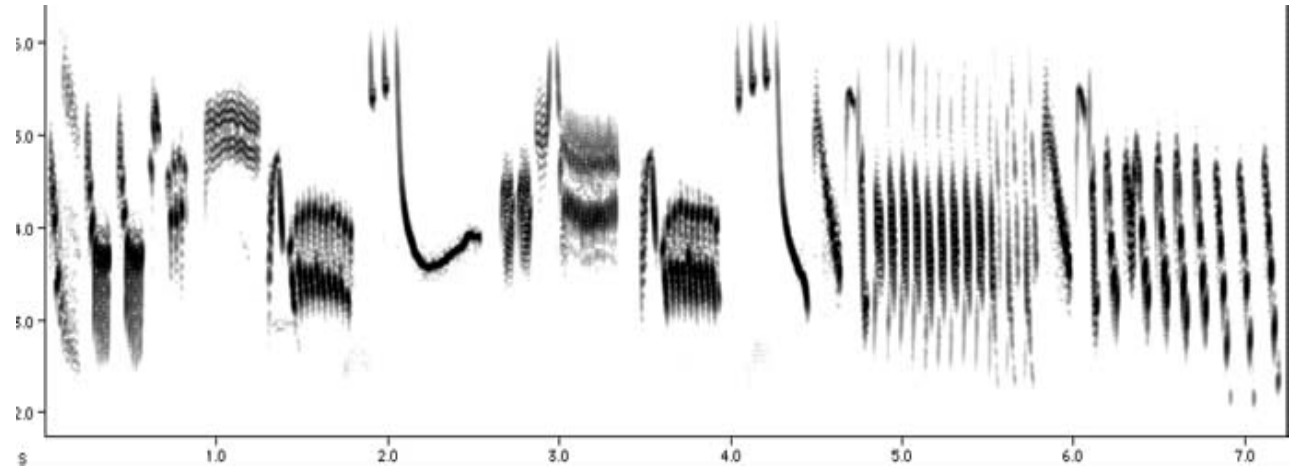
Gestalt

- **Good continuation:** smoothly varying features are bound to a single source.

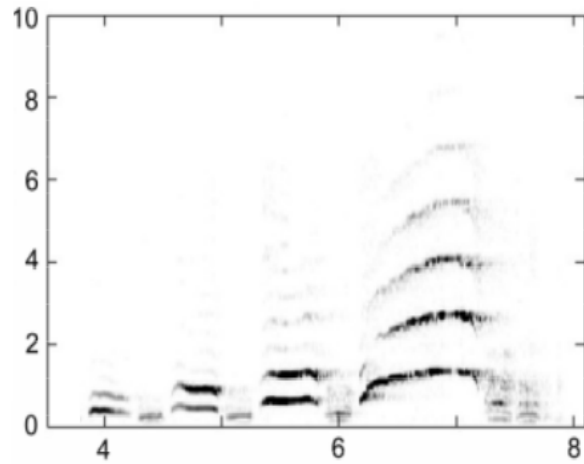
Red Deer Roar



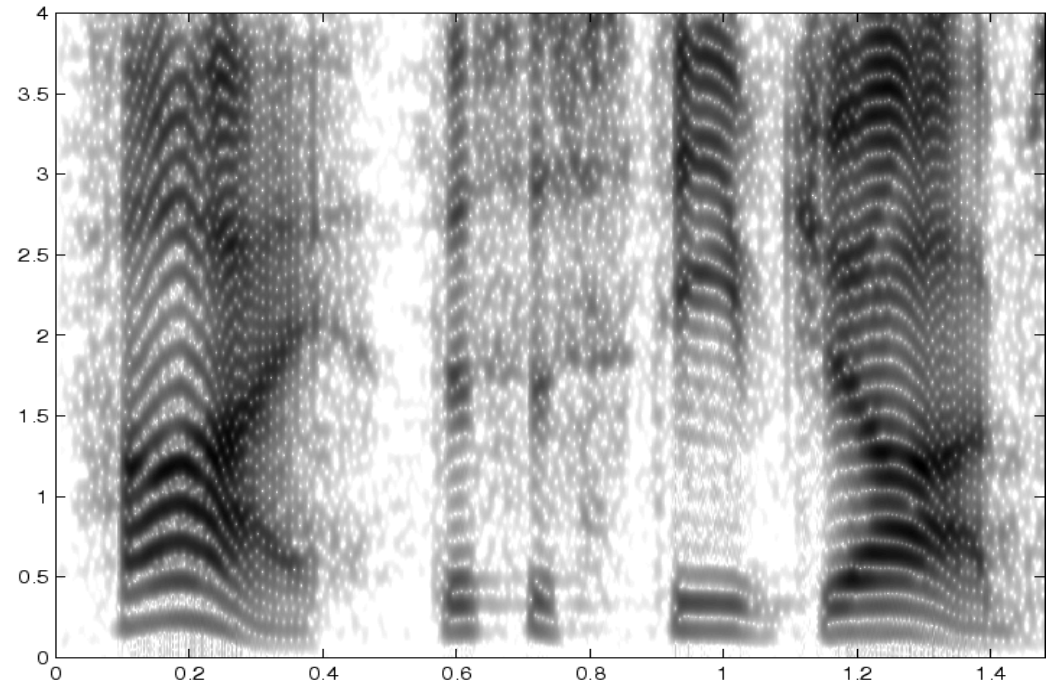
Sky Lark



Chimp Hoot



Speech – “Hello, this is”



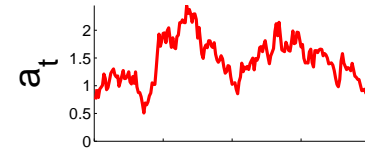
Statistical analogue

- **Sinusoidal components** of a single auditory object have **smoothly varying amplitudes and frequencies**
- Suggests a mixture of sinusoids model with smooth, **Amplitude Modulation (AM)** and **Frequency Modulation (FM)**

Statistical analogue

- **Sinusoidal components** of a single auditory object have **smoothly varying amplitudes and frequencies**
- Suggests a mixture of sinusoids model with smooth, **Amplitude Modulation (AM)** and **Frequency Modulation (FM)**

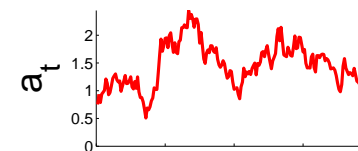
$$p(\mathbf{a}_{k,t} | \mathbf{a}_{k,t-1}) = \text{Slow}, a \geq 0$$



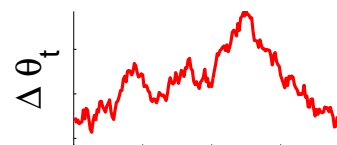
Statistical analogue

- **Sinusoidal components** of a single auditory object have **smoothly varying amplitudes and frequencies**
- Suggests a mixture of sinusoids model with smooth, **Amplitude Modulation (AM)** and **Frequency Modulation (FM)**

$$p(\mathbf{a}_{k,t} | \mathbf{a}_{k,t-1}) = \text{Slow}, \mathbf{a} \geq 0$$



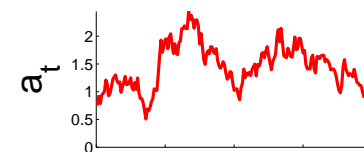
$$p(\Delta\theta_{k,t} | \Delta\theta_{k,t-1}) = \text{Slow}, \Delta\theta > 0$$



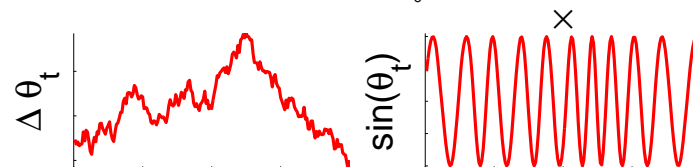
Statistical analogue

- **Sinusoidal components** of a single auditory object have **smoothly varying amplitudes and frequencies**
- Suggests a mixture of sinusoids model with smooth, **Amplitude Modulation (AM)** and **Frequency Modulation (FM)**

$$p(\mathbf{a}_{k,t} | \mathbf{a}_{k,t-1}) = \text{Slow}, \mathbf{a} \geq 0$$



$$p(\Delta\theta_{k,t} | \Delta\theta_{k,t-1}) = \text{Slow}, \Delta\theta > 0$$



$$\sin(\theta_{k,t})$$

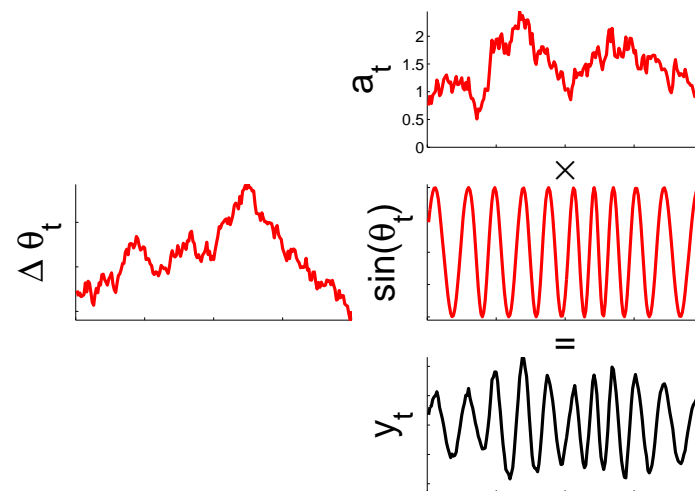
Statistical analogue

- **Sinusoidal components** of a single auditory object have **smoothly varying amplitudes and frequencies**
- Suggests a mixture of sinusoids model with smooth, **Amplitude Modulation (AM)** and **Frequency Modulation (FM)**

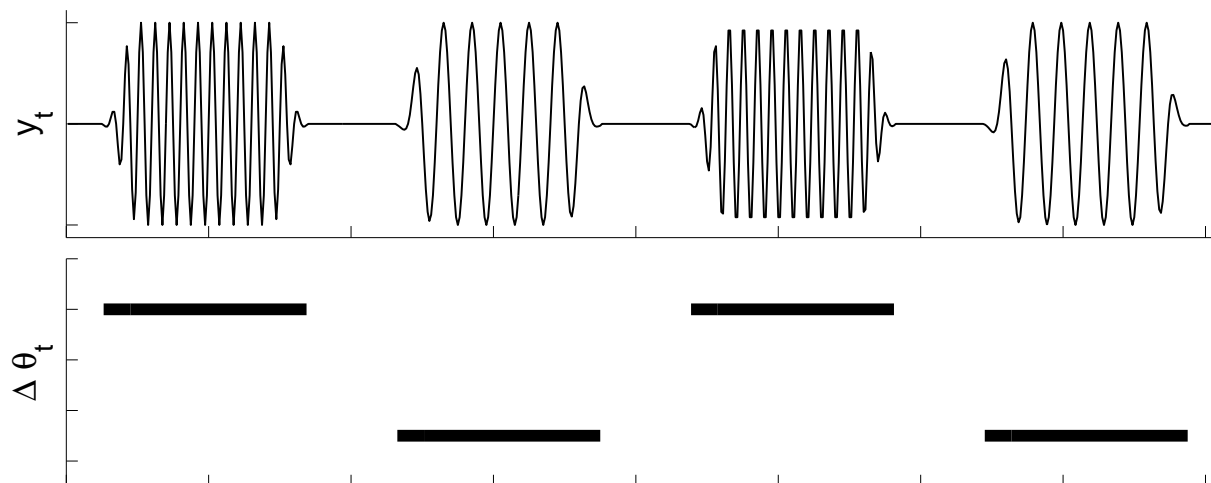
$$p(\mathbf{a}_{k,t} | \mathbf{a}_{k,t-1}) = \text{Slow}, \mathbf{a} \geq \mathbf{0}$$

$$p(\Delta\theta_{k,t} | \Delta\theta_{k,t-1}) = \text{Slow}, \Delta\theta > \mathbf{0}$$

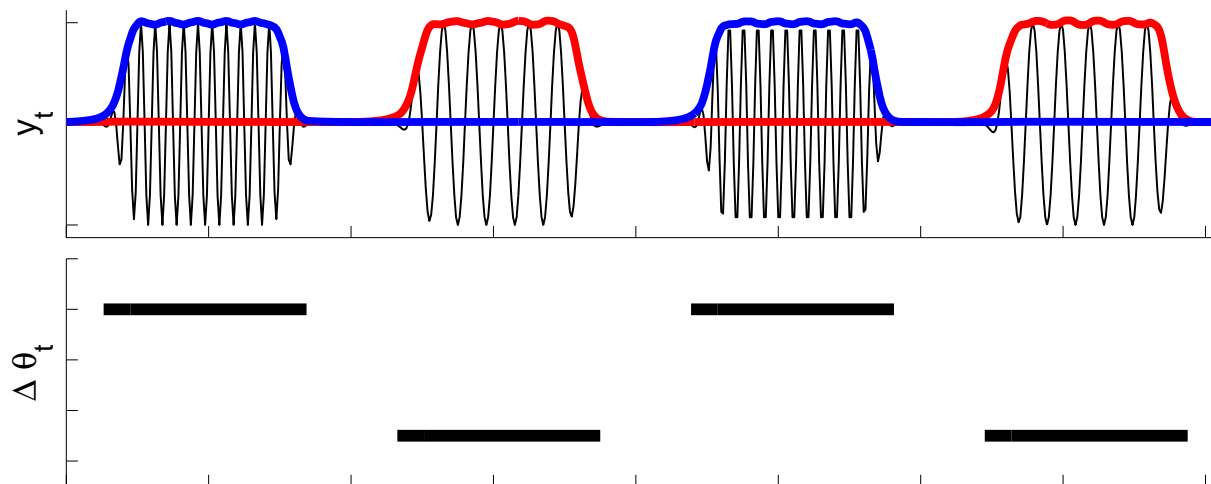
$$\mathbf{y}_t = \sum_{k=1}^K \mathbf{a}_{k,t} \sin(\theta_{k,t})$$



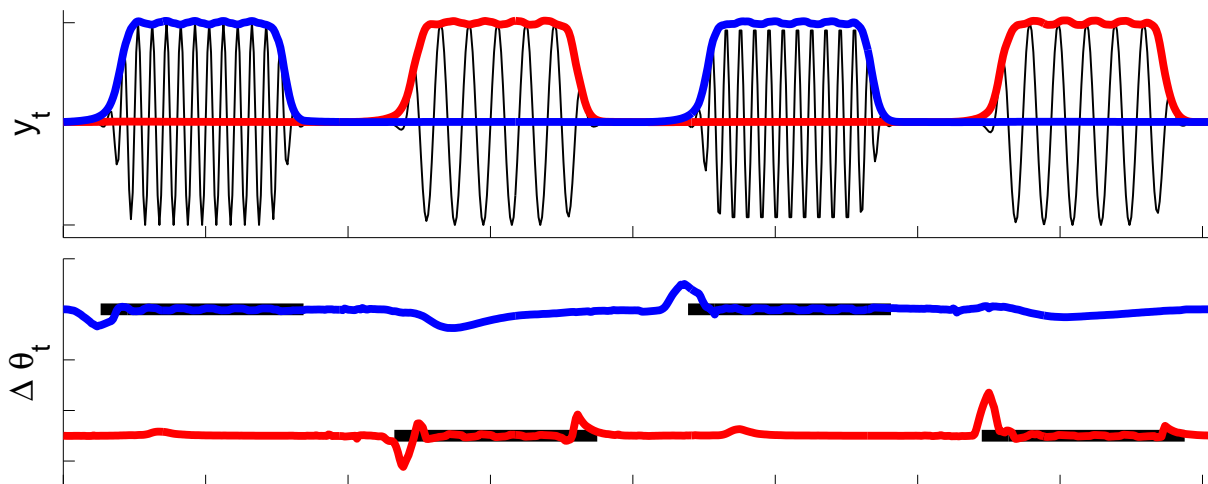
Simulation



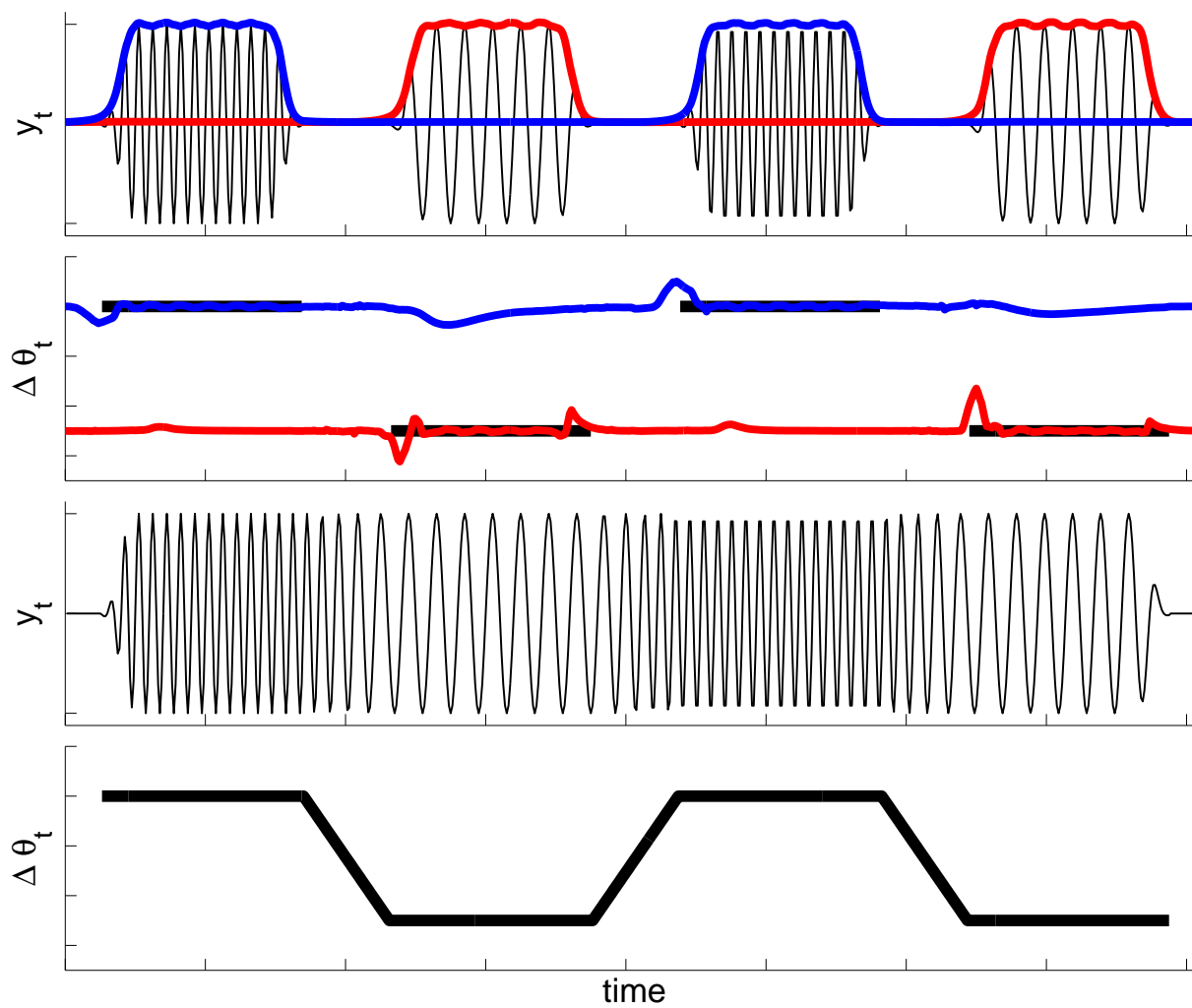
Simulation



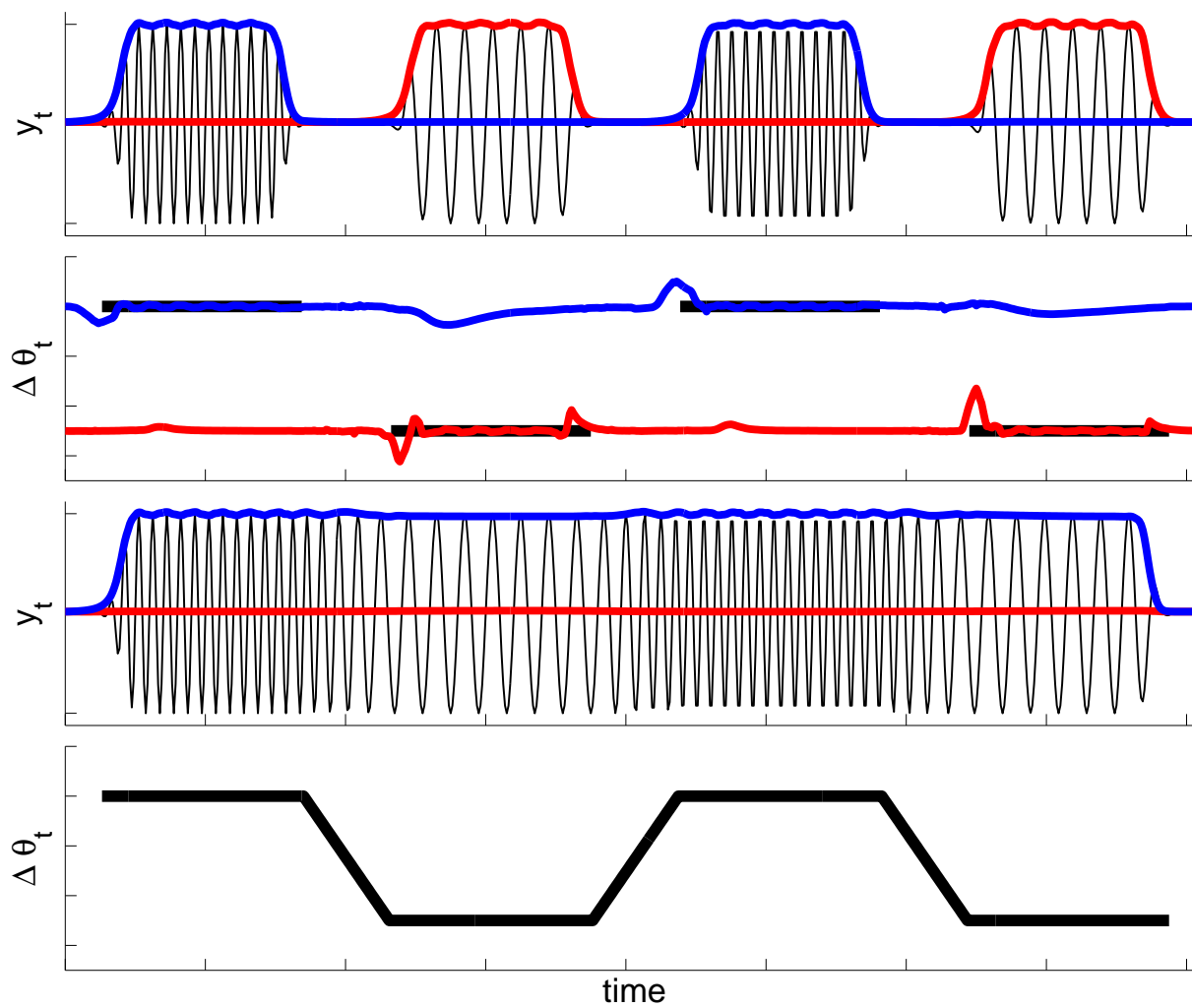
Simulation



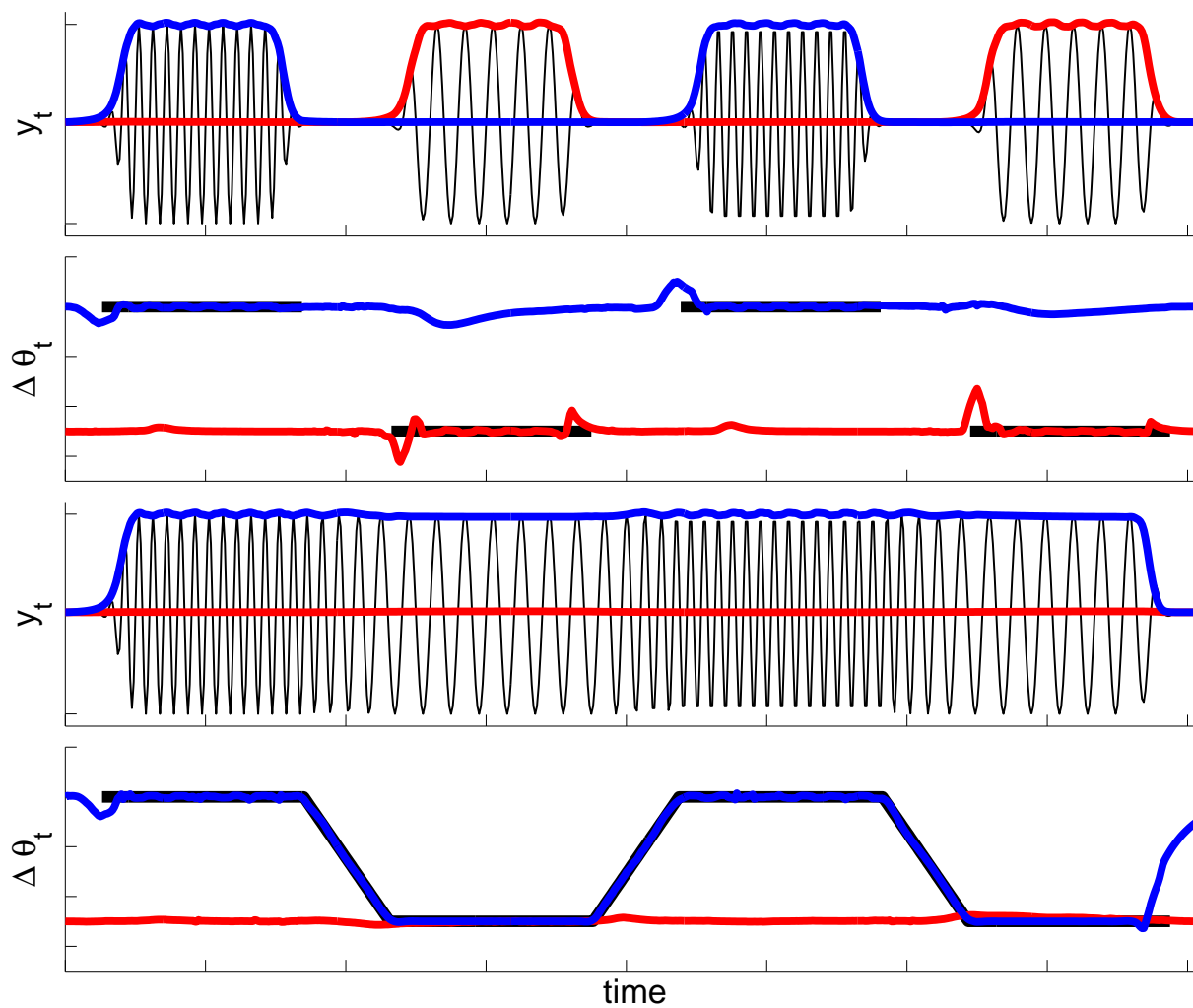
Simulation



Simulation



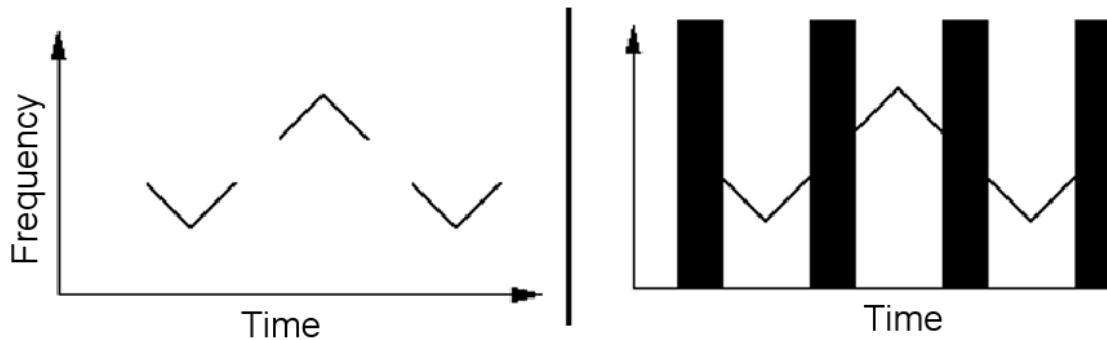
Simulation



Grouping principle 2: Closure

Psychophysics

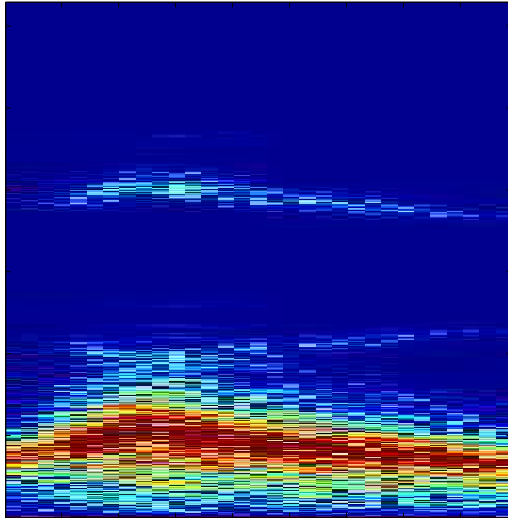
- A FM tone with un-natural gaps is grouped separately
- If the gaps are filled by noise sufficient to mask others components, the tone becomes continuous



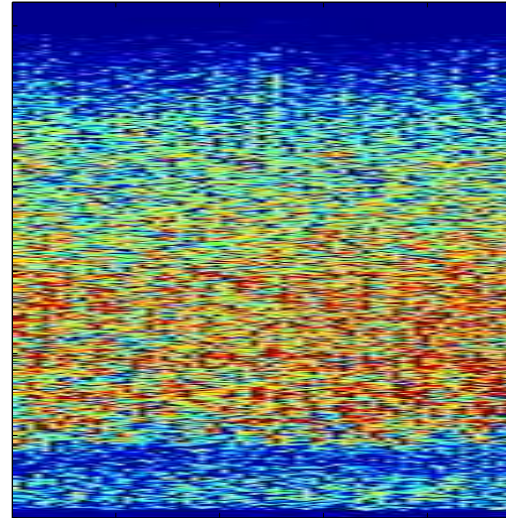
Gestalt

- **Closure:** Fragmentary features with “Good Gestalt” are completed when the fragments contain noise

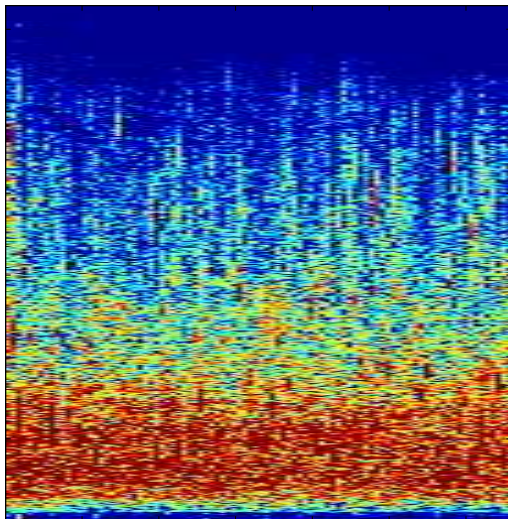
Wind



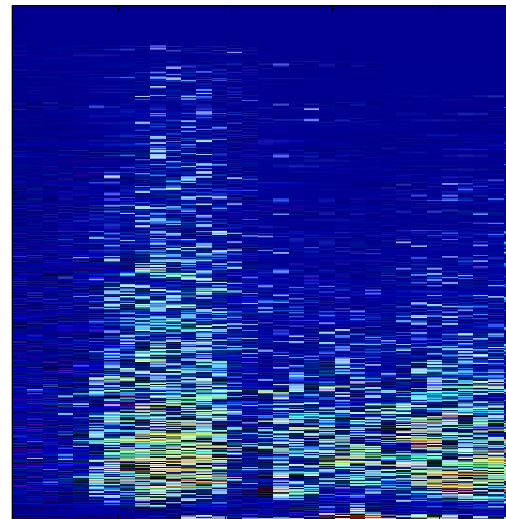
Rain



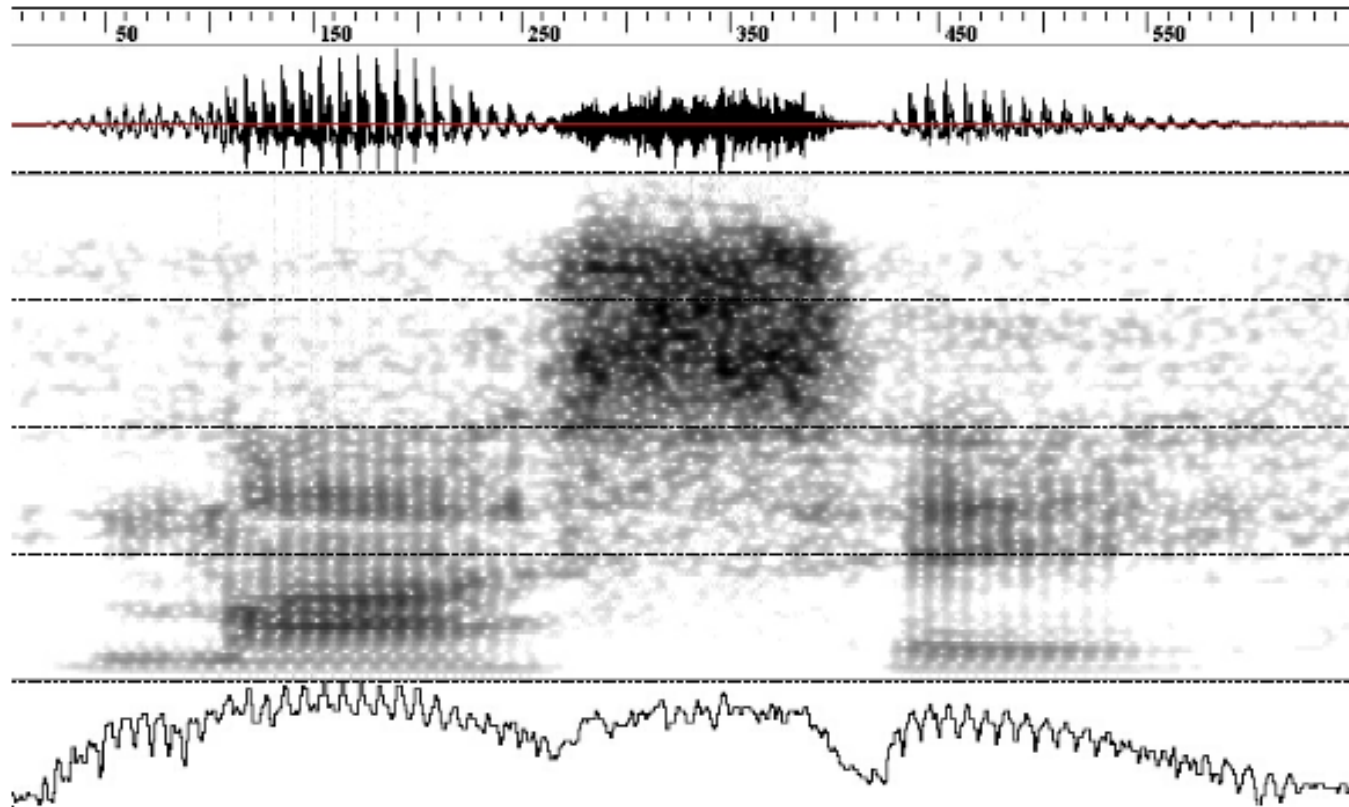
Stream



Waves



Fricatives



Statistical Analogue

- Auditory objects can include **time-varying noise processes**
- Amplitude of the noise is often smoothly varying
- Suggests augmenting the model with a non-stationary noise process

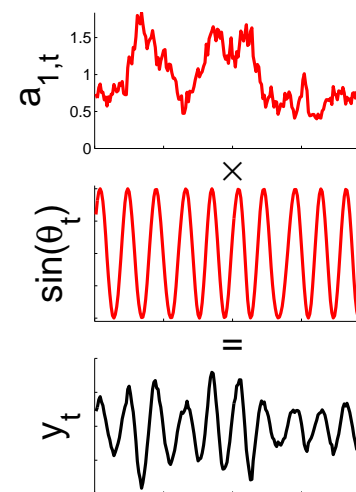
Statistical Analogue

- Auditory objects can include **time-varying noise processes**
- Amplitude of the noise is often smoothly varying
- Suggests augmenting the model with a non-stationary noise process

$$p(a_{k,t} | a_{k,t-1}) = \text{Slow}$$

$$p(\Delta\theta_t | \Delta\theta_{t-1}) = \text{Slow}$$

$$y_t = \sum_{\mathbf{k}} \mathbf{a}_{\mathbf{k},t} \sin(\theta_{\mathbf{k},t})$$

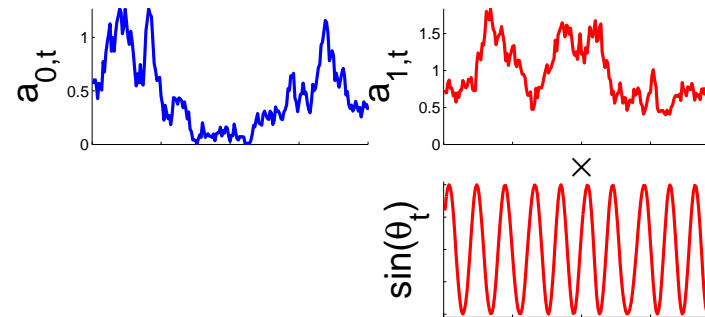


Statistical Analogue

- Auditory objects can include **time-varying noise processes**
- Amplitude of the noise is often smoothly varying
- Suggests augmenting the model with a non-stationary noise process

$$p(a_{k,t} | a_{k,t-1}) = \text{Slow}$$

$$p(\Delta\theta_t | \Delta\theta_{t-1}) = \text{Slow}$$



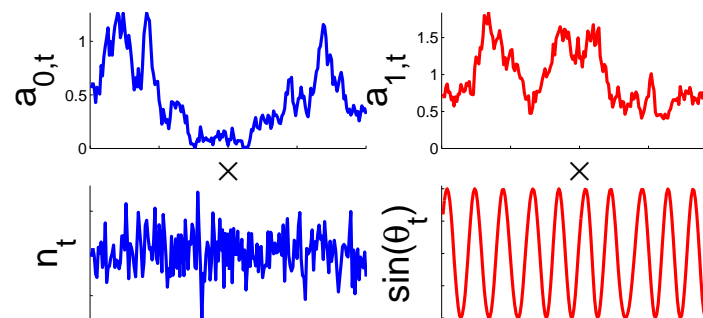
Statistical Analogue

- Auditory objects can include **time-varying noise processes**
- Amplitude of the noise is often smoothly varying
- Suggests augmenting the model with a non-stationary noise process

$$p(a_{k,t} | a_{k,t-1}) = \text{Slow}$$

$$\mathbf{p}(\mathbf{n}_t) = \text{Norm}(\mathbf{0}, \mathbf{1})$$

$$p(\Delta\theta_t | \Delta\theta_{t-1}) = \text{Slow}$$



Statistical Analogue

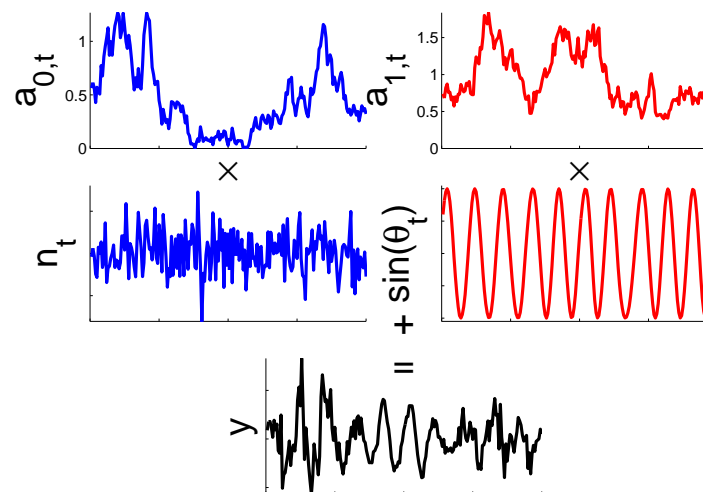
- Auditory objects can include **time-varying noise processes**
- Amplitude of the noise is often smoothly varying
- Suggests augmenting the model with a non-stationary noise process

$$p(a_{k,t} | a_{k,t-1}) = \text{Slow}$$

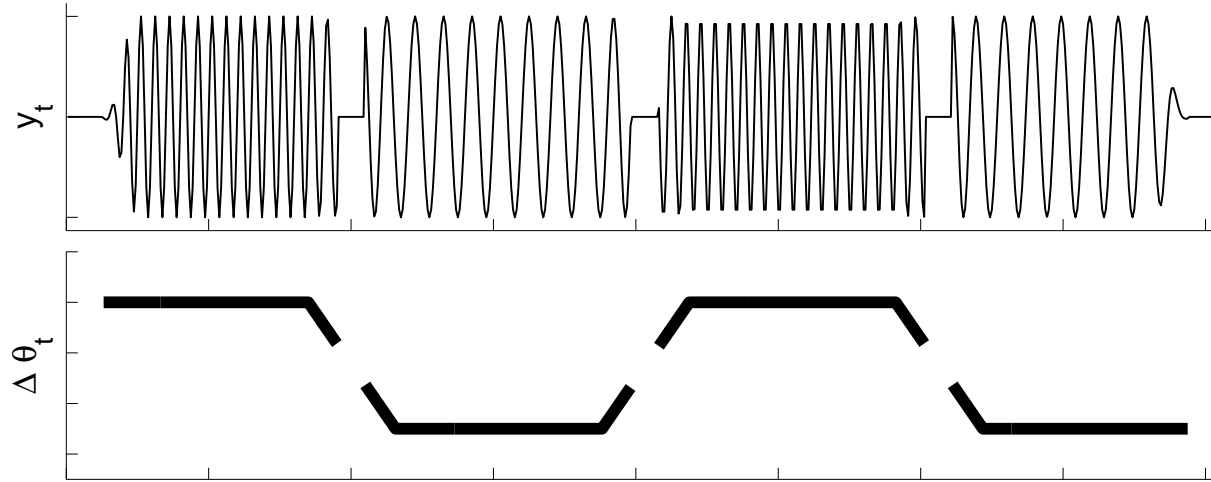
$$\mathbf{p}(\mathbf{n}_t) = \text{Norm}(\mathbf{0}, \mathbf{1})$$

$$p(\Delta\theta_t | \Delta\theta_{t-1}) = \text{Slow}$$

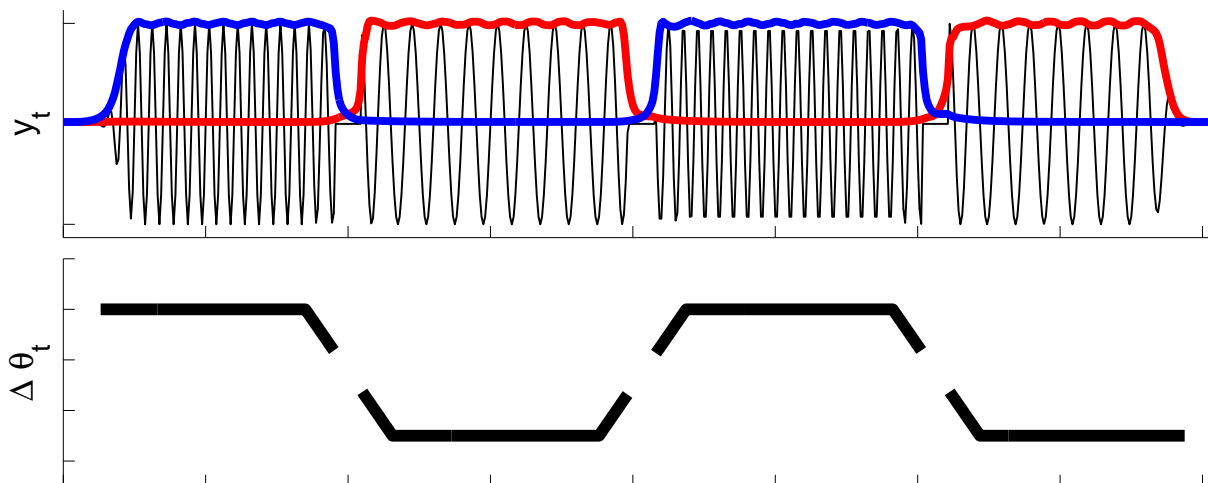
$$\mathbf{y}_t = \sum_{\mathbf{k}} \mathbf{a}_{\mathbf{k},t} \sin(\theta_{\mathbf{k},t}) + \mathbf{a}_{0,t} \mathbf{n}_t$$



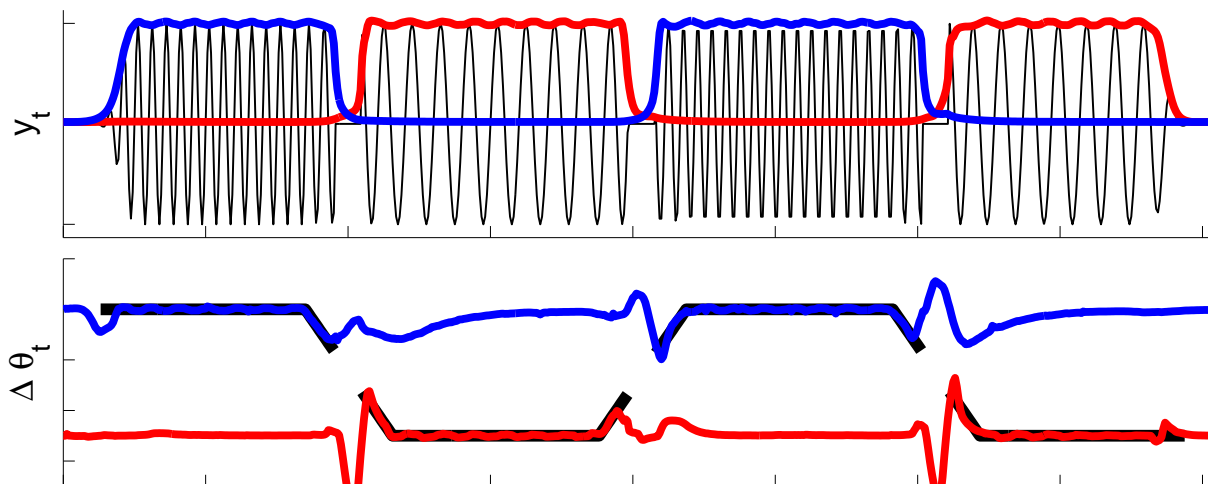
Simulation



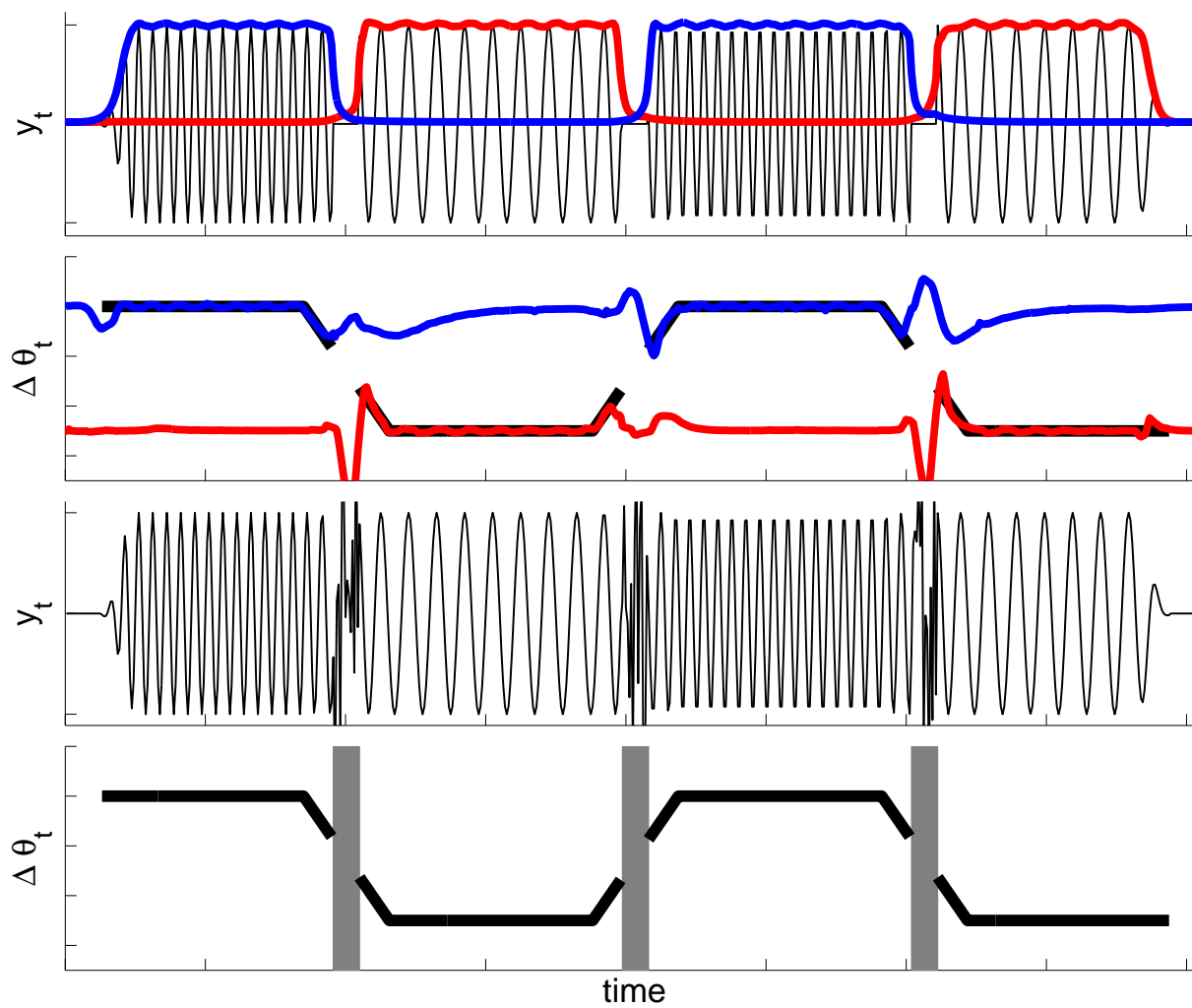
Simulation



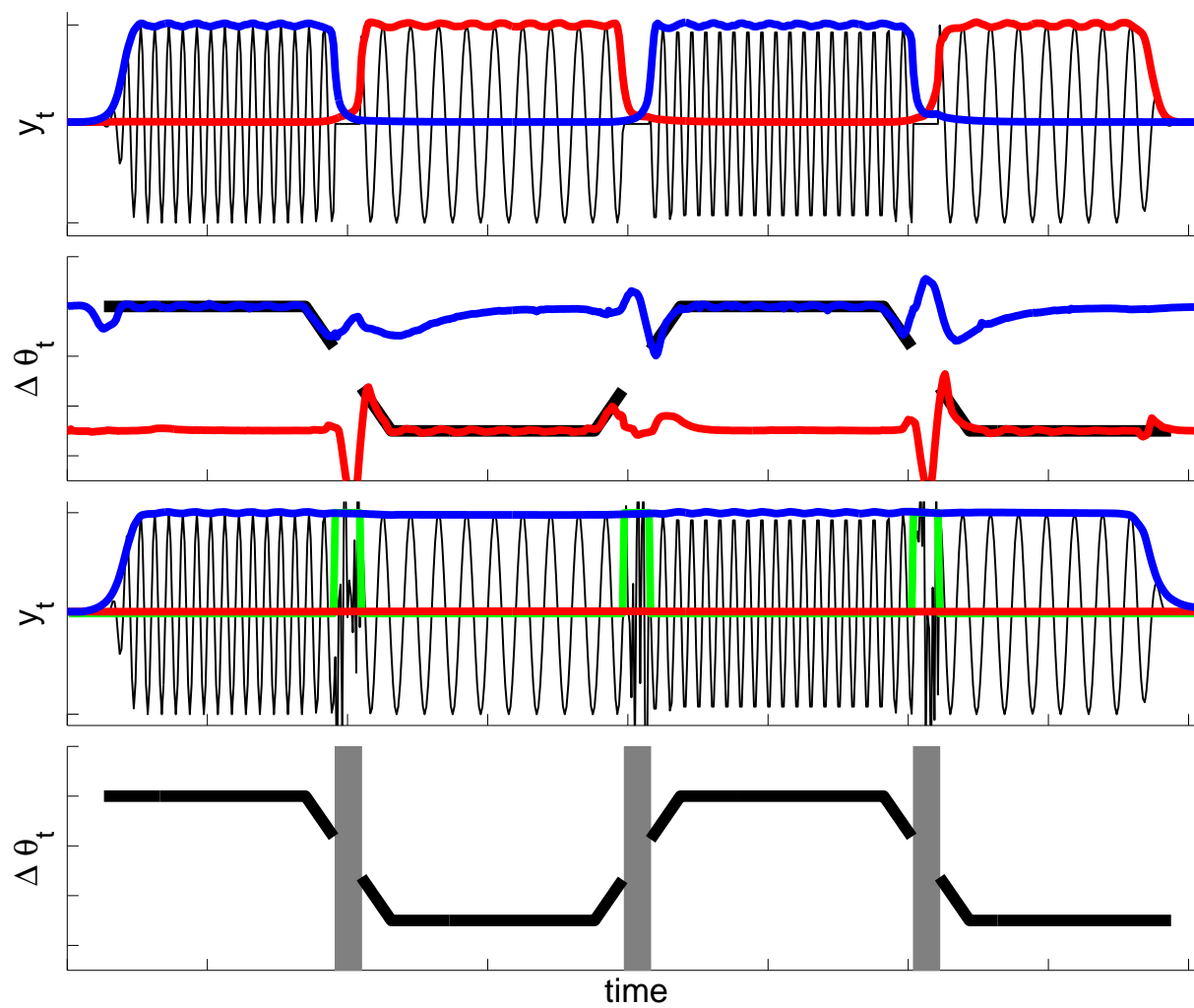
Simulation



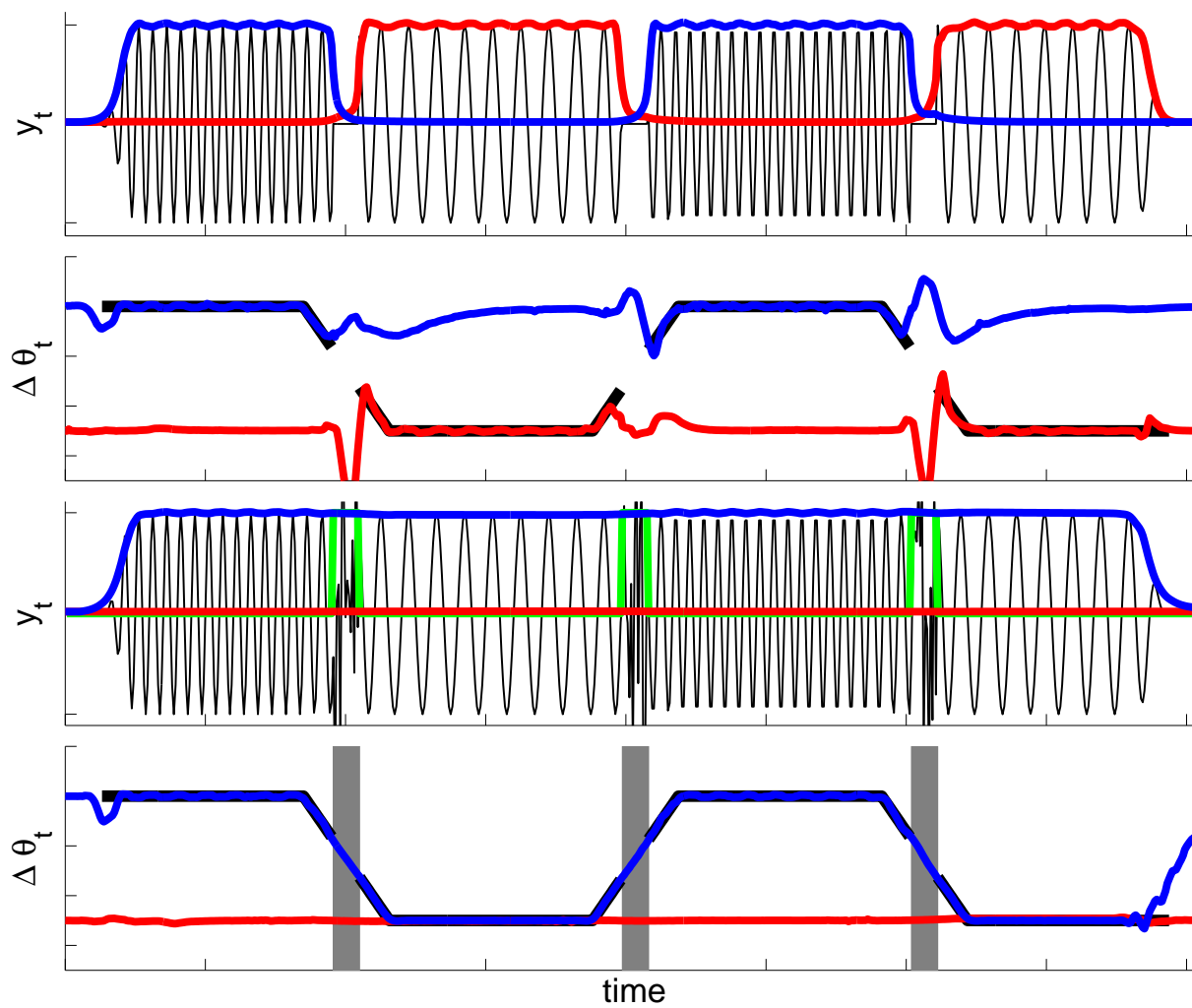
Simulation



Simulation



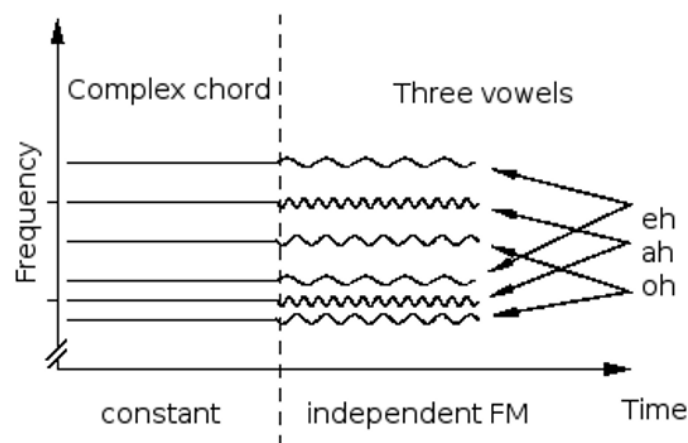
Simulation



Grouping principle 3: Common Fate

Psychophysics

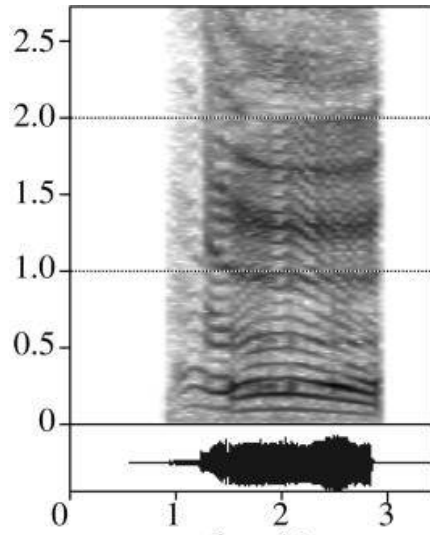
- Pure tones in a stack with common onset are bound to a single object
- Pairs of tones within the stack with common FM are bound separately



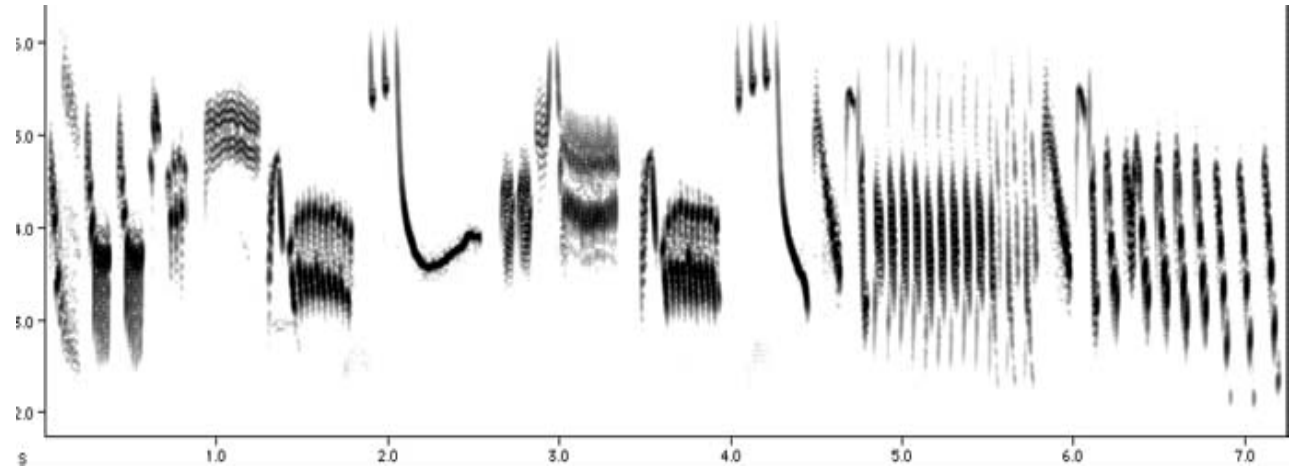
Gestalt

- **Common fate:** Frequency components are grouped when components undergo similar changes simultaneously.

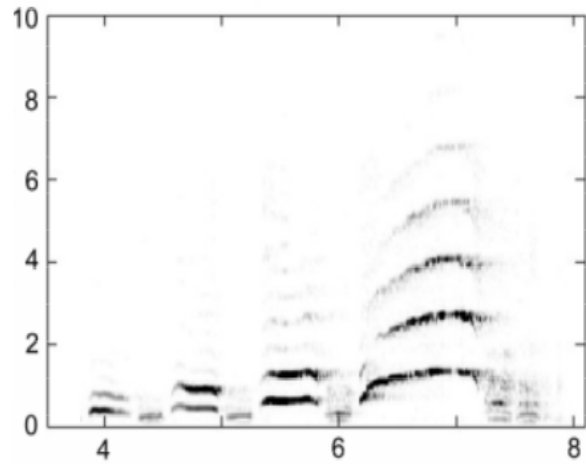
Red Deer Roar



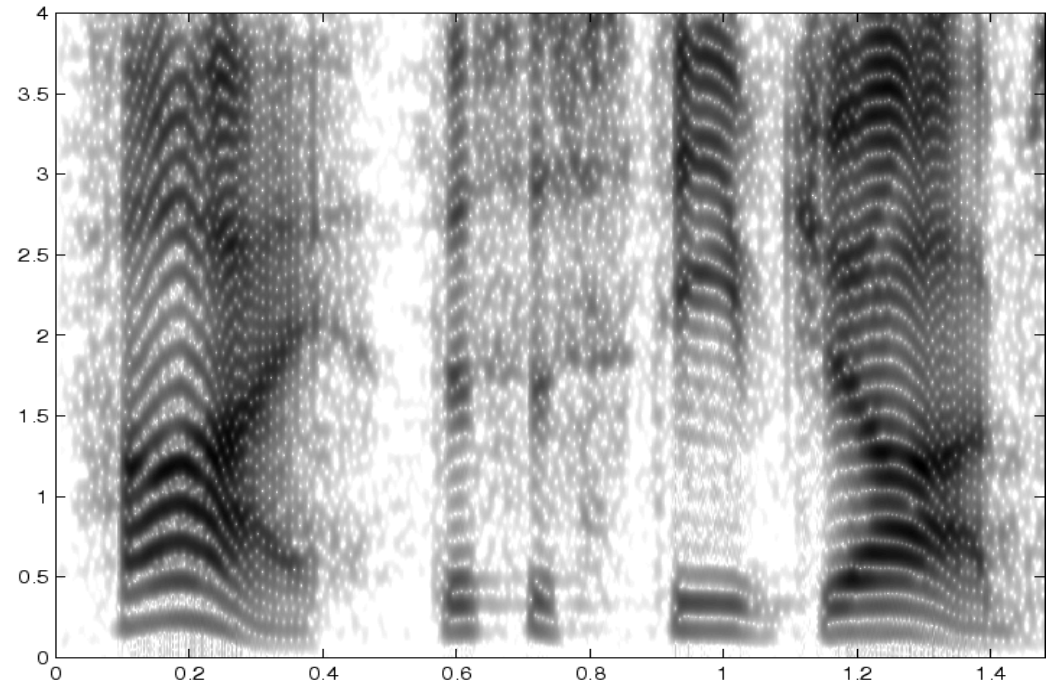
Sky Lark



Chimp Hoot



Speech – “Hello, this is”



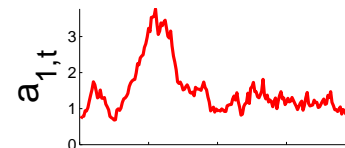
Statistical analogue

- Auditory objects containing **harmonic stacks**, have components with
 - Common onset and AM
 - Common FM
- Extend the model to include co-modulated stacks of tones

Statistical analogue

- Auditory objects containing **harmonic stacks**, have components with
 - Common onset and AM
 - Common FM
- Extend the model to include co-modulated stacks of tones

$$p(a_{k,t}|a_{k,t-1}) = \text{Slow}$$

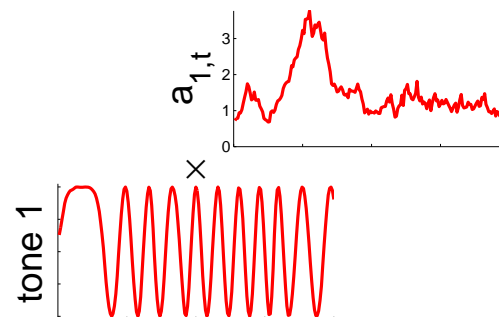


Statistical analogue

- Auditory objects containing **harmonic stacks**, have components with
 - Common onset and AM
 - Common FM
- Extend the model to include co-modulated stacks of tones

$$p(a_{k,t} | a_{k,t-1}) = \text{Slow}$$

$$p(\Delta \hat{\theta}_{\mathbf{k},t} | \Delta \hat{\theta}_{\mathbf{k},t-1}) = \text{Slow}$$



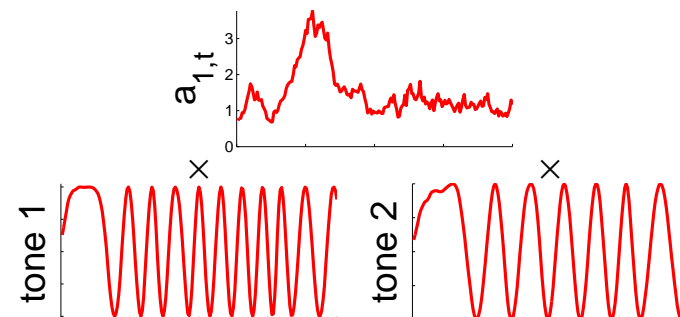
Statistical analogue

- Auditory objects containing **harmonic stacks**, have components with
 - Common onset and AM
 - Common FM
- Extend the model to include co-modulated stacks of tones

$$p(a_{k,t} | a_{k,t-1}) = \text{Slow}$$

$$p(\Delta \hat{\theta}_{\mathbf{k},t} | \Delta \hat{\theta}_{\mathbf{k},t-1}) = \text{Slow}$$

$$\sin(\omega_d \hat{\theta}_{\mathbf{k},t})$$



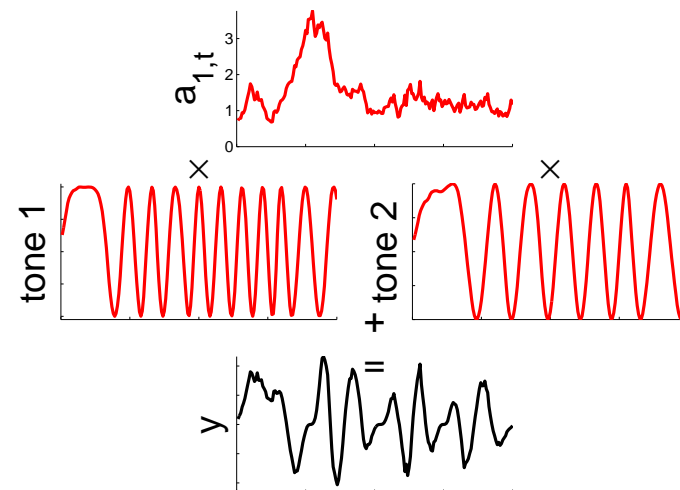
Statistical analogue

- Auditory objects containing **harmonic stacks**, have components with
 - Common onset and AM
 - Common FM
- Extend the model to include co-modulated stacks of tones

$$p(a_{k,t} | a_{k,t-1}) = \text{Slow}$$

$$p(\Delta \hat{\theta}_{\mathbf{k},t} | \Delta \hat{\theta}_{\mathbf{k},t-1}) = \text{Slow}$$

$$\mathbf{y}_t = \mathbf{a}_{\mathbf{k},t} \sum_{d=1}^D \mathbf{g}_{d,\mathbf{k}} \sin(\omega_d \hat{\theta}_{\mathbf{k},t})$$



Statistical analogue

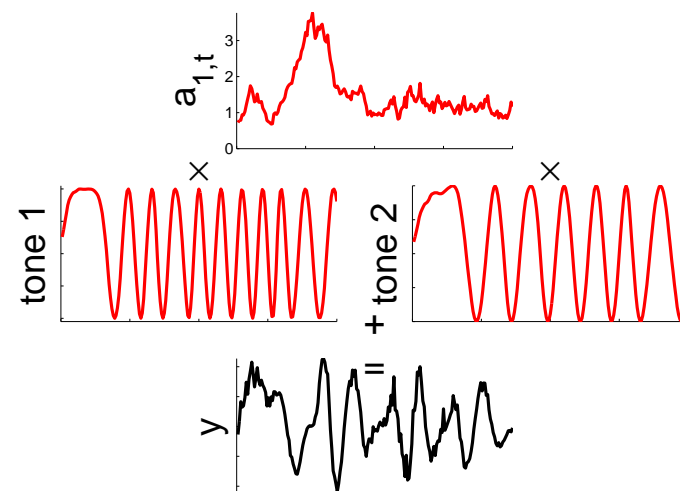
- Auditory objects containing **harmonic stacks**, have components with
 - Common onset and AM
 - Common FM
- Extend the model to include co-modulated stacks of tones

$$p(a_{k,t} | a_{k,t-1}) = \text{Slow}$$

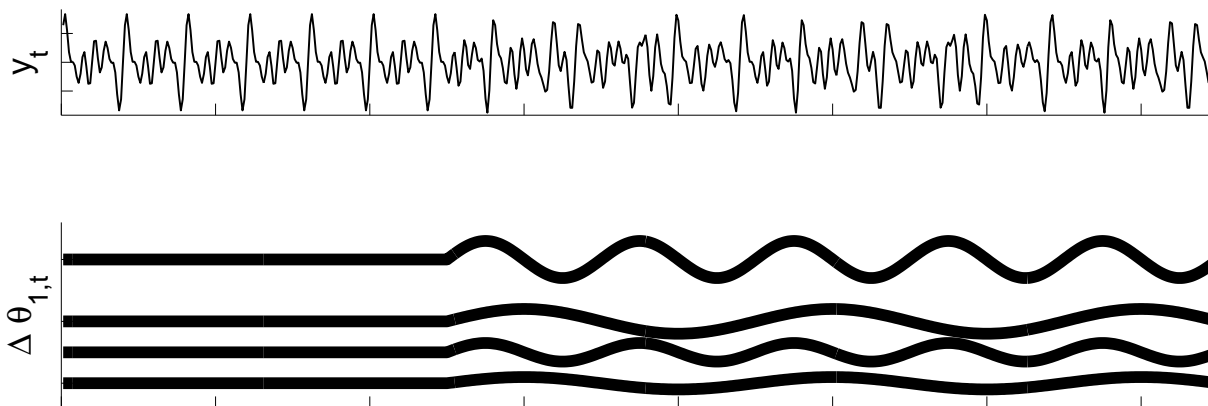
$$p(\Delta \hat{\theta}_{\mathbf{k},t} | \Delta \hat{\theta}_{\mathbf{k},t-1}) = \text{Slow}$$

$$p(\mathbf{n}_t) = \text{Norm}(0, 1)$$

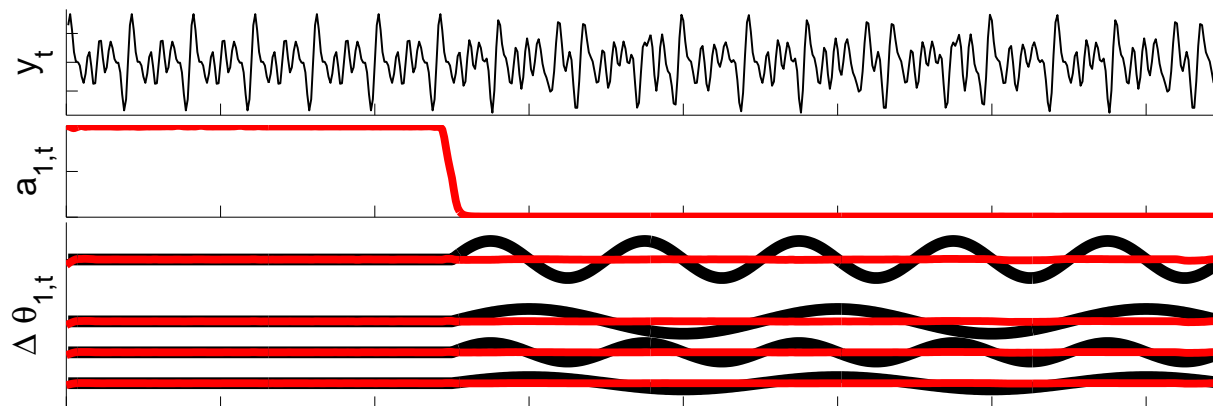
$$\mathbf{y}_t = \sum_{\mathbf{k}} \mathbf{a}_{\mathbf{k},t} \sum_{d=1}^D \mathbf{g}_{d,\mathbf{k}} \sin(\omega_d \hat{\theta}_{\mathbf{k},t}) + \mathbf{a}_{0,t} \mathbf{n}_t$$



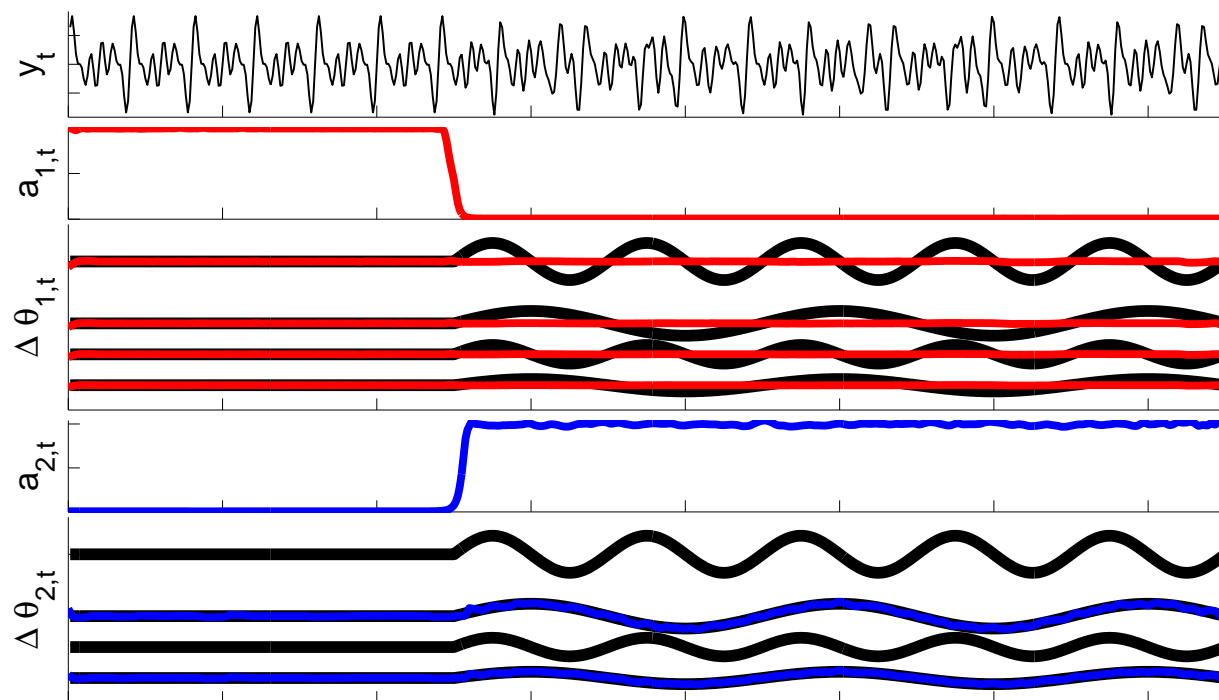
Simulation



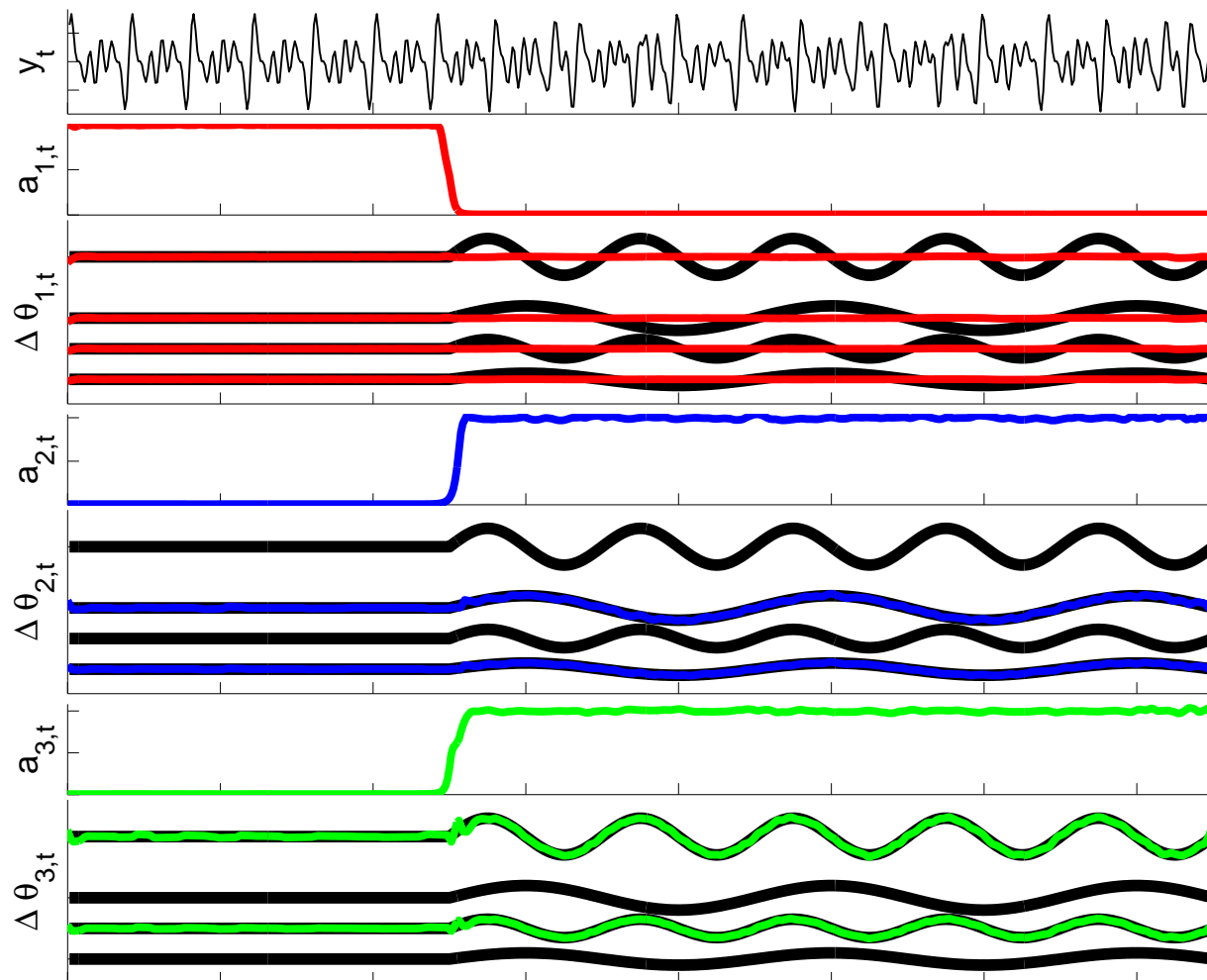
Simulation



Simulation



Simulation

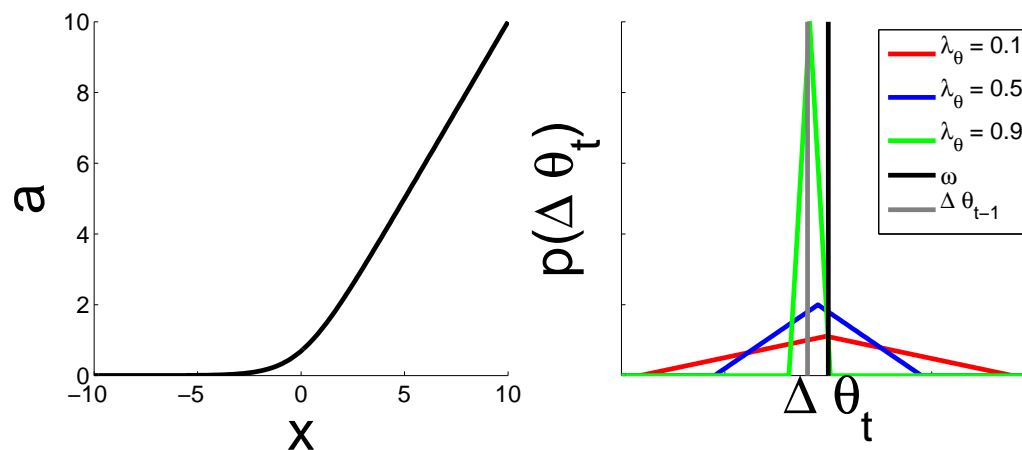


Technical slide on the priors

$$p(x_t|x_{t-1}) = \text{Norm}(\lambda_x x_{t-1}, \sigma^2(1 - \lambda_x^2)) \text{ (Slow)}$$

$$a_t = \log(1 + \exp(x_t)) \text{ (Positive and Sparse)}$$

$$\Delta\theta_t = \text{tri}(\lambda_\theta \Delta\theta_{t-1} + (1 + \lambda_\theta)\omega) \text{ (Limited)}$$



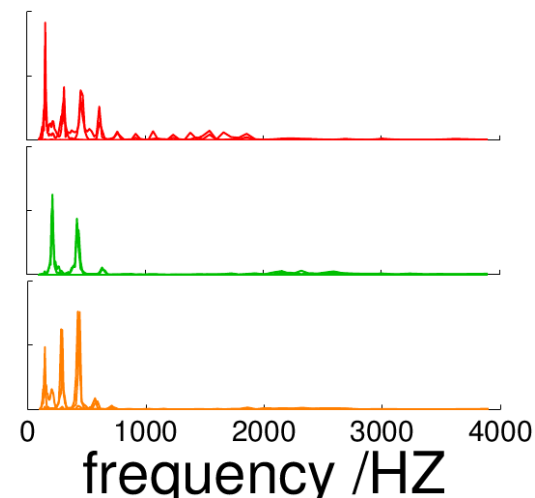
- Learning proceeds via **MAP**: $\arg \max_{X, \Theta} \log p(Y, X, \Theta)$
- Optimising phases far superior to optimising frequencies, but **still difficult**
- **Online inference** helps prevent the phase carpet getting rucked-up

Towards Computational Auditory Scene Analysis

- Previous model for toy Gestalt results, and many computational scene analysis algorithms, have **hand set parameters**
- BUT: We can **learn the parameters of the model** from auditory scenes

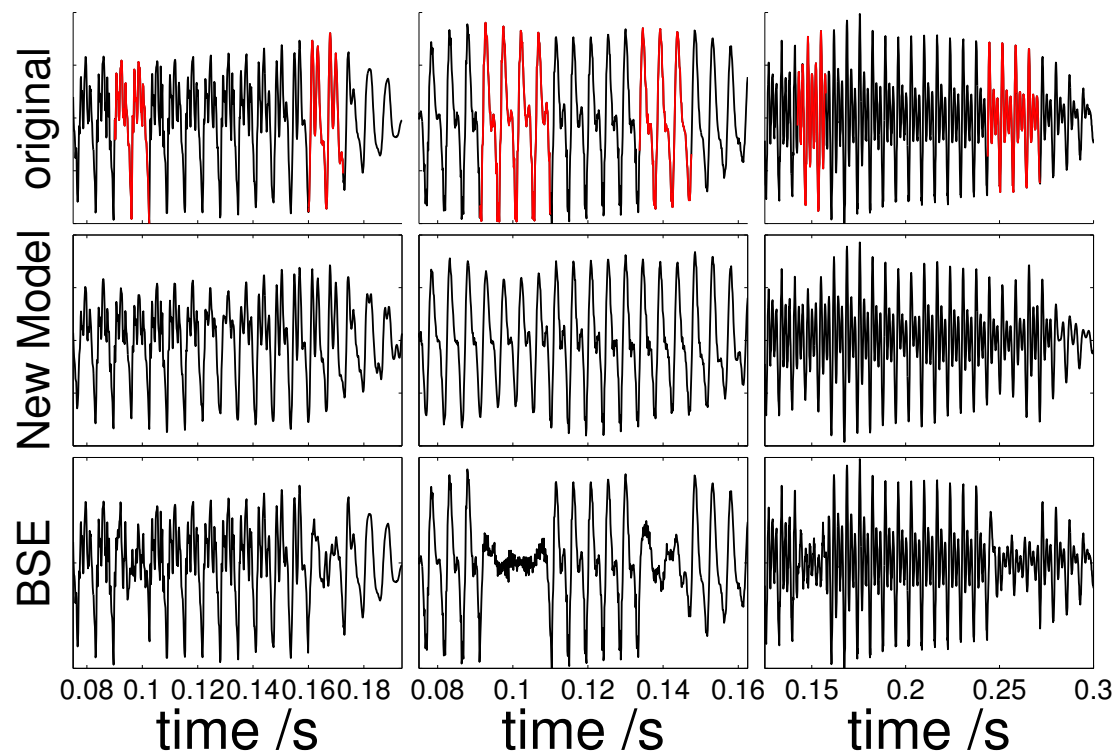
$$X, \Theta, G = \arg \max_{X, \Theta, G} \log p(Y, X, \Theta, G)$$

- We then have a **generative model for natural scenes** which we can use for
 - denoising, filling in missing data, CASA, resynthesis and manipulation etc.



Preliminary Results for Filling In Missing Data

Trained on different sentences from the same speaker, fixed frequencies.



Relationship to signal processing representations

- **Hilbert transform**: $y_t = a_t \Re(\exp(i\theta_t))$ s.t. FT $(a_t \exp(i\theta_t))$ is one sided
 - Envelope produced is often poor, probabilistic algorithms are superior
 - Decomposition is fundamentally **ill-posed** so probabilistic technique is natural
- **Time-frequency representations**: e.g. short-time FT, wavelet or spectrogram
 - New method **side-steps uncertainty** issues: **automatic window selection**
 - Resynthesis and filling in missing data simple: **waveform forward model**
- **Harmonic plus noise analysis**: Like a heuristic version of the above
 - New methods can **learn** the many parameters in this representation

Viewing classical signal processing methods as ill-posed inference problems in a dynamical model has many advantages, but inference is slow.

Future Work

- **Psychophysics**

- Quantitative modeling of results **missing fundamental**, Proximity, and prediction of new phenomena
- Use samples from the model as stimuli

- **Computational auditory scene analysis**

- Use mixtures of these models for auditory scene analysis
- Run psychophysical tests on model trained on natural stimuli

- **Technical**

- Improve optimisation over the phases
- Improve the prior distribution over the frequencies: Slow at high amplitudes, Fast at low amplitudes.