Modulation cascades, sound textures and mid-level audition

Richard E. Turner (ret26@cam.ac.uk)
Computational and Biological Learning Lab, University of Cambridge

Maneesh Sahani (maneesh@gatsby.ucl.ac.uk)
Gatsby Computational Neuroscience Unit, UCL, London
Motivation

- object
  - object parts
    - structural primitives
      - sensory input
Motivation

- object
  - object parts
    - structural primitives
      - sensory input
  - trees
    - bark
      - oriented edges
        - image
  - birdsong
    - motifs
      - AM tones & noise
        - sound waveform
Auditory Scene Analysis

- object
  - object parts
    - structural primitives
      - sensory input
- trees
  - bark
    - oriented edges
      - image
- birdsong
  - motifs
    - AM tones & noise
      - sound waveform

- super schema-based grouping
- schema-based grouping
- primitive grouping
Statistics of sounds

\[ p(\text{object}) \]
\[ \downarrow \]
\[ p(\text{parts} | \text{object}) \]
\[ \downarrow \]
\[ p(\text{primitives} | \text{part}) \]
\[ \downarrow \]
\[ p(\text{sound} | \text{primitive}) \]

\text{birdsong}
\[ \downarrow \]
\text{motifs}
\[ \downarrow \]
\text{AM tones & noise}
\[ \downarrow \]
\text{sound waveform}

\text{super}
\[ \uparrow \]
\text{schema-based}
\text{grouping}

\text{schema-based}
\text{grouping}

\text{primitive}
\text{grouping}
Auditory scene analysis as inference

\[
p(\text{object} | \text{sound}) \\
p(\text{parts} | \text{sound}) \\
p(\text{primitives} | \text{sound}) \\
p(\text{sound} | \text{primitive})
\]

- birdsong
- motifs
- AM tones & noise
- sound waveform

- p(object)
- p(parts | object)
- p(primitives | part)
- p(sound | primitive)

super schema-based grouping
schema-based grouping
primitive grouping
Auditory scene analysis as inference

\[ p(\text{object}) \]
\[ p(\text{parts} | \text{object}) \]
\[ p(\text{primitives} | \text{part}) \]
\[ p(\text{sound} | \text{primitive}) \]
\[ p(\text{sound}) \]
\[ p(\text{primitives} | \text{sound}) \]
\[ p(\text{motifs}) \]
\[ p(\text{AM tones & noise}) \]
\[ p(\text{sound}) \]
\[ p(\text{sound} | \text{primitive}) \]

Bayes' Theorem

\[ p(\text{primitive} | \text{sound}) = \frac{p(\text{sound} | \text{primitive})p(\text{primitive})}{p(\text{sound})} \]
Probabilistic primitive auditory scene analysis

Bayes' Theorem

$$p(\text{primitive}|\text{sound}) = \frac{p(\text{sound}|\text{primitive})p(\text{primitive})}{p(\text{sound})}$$
Part 1: Statistical model: primitive auditory scene synthesis

Part 2: Inference: primitive auditory scene analysis

Provocative computational theory: Auditory grouping rules arise from inferences based on the statistics of natural sounds.
Primitive Probabilistic Auditory Scene Analysis

Part 1: Statistical model: primitive auditory scene synthesis

Part 2: Inference: primitive auditory scene analysis

Provocative computational theory: Auditory grouping rules arise from inferences based on the statistics of natural sounds.
Primitive Probabilistic Auditory Scene Analysis

Part 1: Statistical model: primitive auditory scene synthesis

What are the important low-level statistics of natural sounds?

Part 2: Inference: primitive auditory scene analysis

Provocative computational theory: Auditory grouping rules arise from inferences based on the statistics of natural sounds.
Heuristic Analysis: Fire sound
Heuristic Analysis: Fire sound
Heuristic Analysis: Fire sound

Heuristics:

- Filter
- Demodulate

Frequency Analysis:
- 4.1 KHz
- 2.4 KHz
- 1 KHz

Time /s

Y

0.5 0.52 0.54 0.56 0.58 0.6 0.62
Heuristic Analysis: Fire sound

![Diagram of sound analysis with various frequency components and filters.]
Heuristic Analysis: Fire sound
Heuristic Analysis: Fire sound

![Diagram showing the analysis of fire sound with frequency and time axes, demodulation and filtering processes highlighted, and PCA features indicated.]
Heuristic Analysis: Fire sound

PCA features

demodulate
filter

0.5 0.52 0.54 0.56 0.58 0.6 0.62

y
time /s
0.2
0.6
1.9
5.8
18

frequency /kHz
1 2 3

1 2 3
Heuristic Analysis: Rain

- PCA features: 0.7, 0.75, 0.8, 0.85
- y vs time /s: 0.2, 0.6, 1.6, 4.6, 13
- Frequency /kHz: 1, 2, 3
- Demodulate, Filter

Diagram showing oscillation patterns over time and frequency, with a color scale indicating PCA features.
Heuristic Analysis: Water

PCA features

demodulate
filter

Y

0.1 0.15 0.2 0.25 0.3 0.35 0.4

time /s

0.2 0.5 1.3 3.2 8

frequency /kHz

1 6 10

Heuristic Analysis: Speech
Summary

sound → filter bank → demodulate → envelope patterns

• **Important statistics include**
  
  – energy in sub-bands (power-spectrum)
  – patterns of co-modulation
  – time-scale of the modulation
  – depth of the modulation (sparsity)
Summary

sound $\rightarrow$ filter bank $\rightarrow$ demodulate $\rightarrow$ envelope patterns

- **Important statistics include**
  - energy in sub-bands (power-spectrum)
  - patterns of co-modulation
  - time-scale of the modulation
  - depth of the modulation (sparsity)

- Formulate a probabilistic model to capture these statistics:
Summary

sound → filter bank → demodulate → envelope patterns

- Important statistics include
  - energy in sub-bands (power-spectrum)
  - patterns of co-modulation
  - time-scale of the modulation
  - depth of the modulation (sparsity)

- Formulate a probabilistic model to capture these statistics:

sound ← modulate carriers ← modulators ← envelope patterns

Structural Primitives = co-modulated narrow-band processes
Statistical Model

\[ y(t) = \sum_{d=1}^{D} c_d(t)a_d(t) \]
Statistical Model

\[ y(t) = \sum_{d=1}^{D} c_d(t) a_d(t) \]

c\_d(t) = bandpass Gaussian noise

signal

carriers

envelopes
Statistical Model

\( x_k(t) = \text{lowpass Gaussian noise} \)

\( a_d(t) = g_+ \left( \sum_{k=1}^{K} w_{d,k} x_k(t) \right) \)

\( c_d(t) = \text{bandpass Gaussian noise} \)

\( y(t) = \sum_{d=1}^{D} c_d(t) a_d(t) \)
Intuitions for role of model parameters: Generation

envelope modulation patterns
Generation: adding comodulation

envelope modulation patterns

Y

0 0.05 0.1 0.15 0.2 0.25 0.3

time /s

0.1

0.3

0.7

1.5

3.3

frequency /KHz

1 14 27
Generation: adding comodulation

Envelope modulation patterns
Generation: adding comodulation

![Graph and diagram showing frequency and time relationships with envelope modulation patterns.]
Generation: adding comodulation

envelope modulation patterns

frequency /KHz

1 14 27

Y

0 0.05 0.1 0.15 0.2 0.25 0.3
time /s
Generation: Decreasing time-scale

envelope modulation patterns
Generation: Decreasing time-scale

![Diagram showing decreasing time-scale with frequency and time axes.](image)
Generation: Decreasing time-scale

envelope modulation patterns
Generation: Decreasing time-scale

Envelope modulation patterns

Y

0 0.05 0.1 0.15 0.2 0.25 0.3
time /s

0.1 0.3 0.7 1.5 3.3
frequency /KHz

1 14 27
Generation: Decreasing time-scale
Generation: Decreasing time-scale

![Graph showing envelope modulation patterns and frequency over time](image)
Generation: Decreasing time-scale

![Waveform and spectrogram with time and frequency axes labeled.](image)
Generation: Increasing sparsity

Envelope modulation patterns
Generation: Increasing sparsity

envelope modulation patterns

y
time /s
0.1
0.3
0.7
1.5
3.3
frequency /KHz
1 14 27

0 0.05 0.1 0.15 0.2 0.25 0.3

0 0.05 0.1 0.15 0.2 0.25 0.3

time /s
Generation: Increasing sparsity

envelope modulation patterns

frequency /Hz

0.1
0.3
0.7
1.5
3.3
time /s

0 0.05 0.1 0.15 0.2 0.25 0.3

y

Generation: Increasing sparsity

![Waveform and frequency spectrum](image)

- **Frequency (KHz)**: 1, 14, 27
- **Time (s)**: 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3
- **Envelope modulation patterns**
Sound Generation Demo

fire stream wind rain
tent-zip foot step Turner, 2010

time /s
frequency /kHz
Part 1: Statistical model: primitive auditory scene synthesis

Part 2: Inference: primitive auditory scene analysis

Provocative computational theory: Auditory grouping rules arise from inferences based on the statistics of natural sounds.
Primitive Probabilistic Auditory Scene Synthesis

envelope patterns

tenvelopes

carriers

signal
Primitive Probabilistic Auditory Scene Analysis

- Envelope patterns
- Envelopes
- Carriers
- Signal
- Modulation pattern
- IC/auditory cortex
- Demodulation
- Auditory nerve
- Auditory filter bank (with gain control)
- Inner ear
Continuity Illusion

\[ y(t) = a_1(t)c_1(t) + a_2(t)c_2(t) \]

Signal

Frequency / Hz
Continuity Illusion

\[ y_t = a_{1,t}c_{1,t} + a_{2,t}c_{2,t} \]
Continuity Illusion

\[ y_t = a_{1,t}c_{1,t} + a_{2,t}c_{2,t} \]
Common Amplitude Modulation

\[ y(t) = (a_{1,t} + a_{3,t})c_{1,t} + (a_{2,t} + a_{3,t})c_{2,t} \]
Common Amplitude Modulation

\[ y_t = (a_{1,t} + a_{3,t})c_{1,t} + (a_{2,t} + a_{3,t})c_{2,t} \]
Common Amplitude Modulation

\[ y_t = (a_{1,t} + a_{3,t})c_{1,t} + (a_{2,t} + a_{3,t})c_{2,t} \]
Good Continuation

signal 1

signal 2

frequency /Hz

time /s
Good Continuation

\[ y_t = a_{1,t} c_{1,t} + a_{2,t} c_{2,t} \]
Good Continuation

\[ y_t = a_{1,t}c_{1,t} + a_{2,t}c_{2,t} \]
Proximity

\[ y_{t} = a_{1,t}c_{1,t} + a_{2,t}c_{2,t} \]
Proximity

\[ y_t = a_{1,t} c_{1,t} + a_{2,t} c_{2,t} \]
Proximity

\[ y_t = a_{1,t}c_{1,t} + a_{2,t}c_{2,t} \]
Conclusions

- Developed a model for natural sounds comprising *quickly varying carriers* and *slowly varying modulators*
- Captures the statistics of simple *auditory textures*
- **Inference** replicates characteristics of *primitive auditory scene analysis*
Additional Slides
Motivation

- object
  - parts of objects
    - structural primitives
      - sensory input
  - trees
    - bark
    - edges
    - photoreceptor activities
Motivation

- object
  - parts of objects
    - structural primitives
      - sensory input
  - bird song
    - motif
      - AM tones and noise
        - auditory filter activities
  - trees
    - bark
      - edges
        - photoreceptor activities
Generative model: Probabilistic Auditory Scene Synthesis

- $p(\text{source})$
- $p(\text{parts}|\text{source})$
- $p(\text{primitive}|\text{part})$
- $p(\text{sound}|\text{primitives})$

- bird song
- motif
- AM tones and noise
- auditory filter activities

- trees
- bark
- edges
- photoreceptor activities

- super schema–based grouping
- schema–based grouping
- primitive grouping
Recognition model: Probabilistic Auditory Scene Analysis

\[
\begin{align*}
p(\text{source}) \quad \rightarrow \\
p(\text{parts}|\text{source}) \quad \rightarrow \\
p(\text{primitive}|\text{part}) \quad \rightarrow \\
p(\text{sound}|\text{primitives})
\end{align*}
\]

\[
\begin{align*}
\text{bird song} \quad \rightarrow \\
\text{motif} \quad \rightarrow \\
\text{AM tones and noise} \quad \rightarrow \\
\text{auditory filter activities}
\end{align*}
\]

\[
\begin{align*}
p(\text{source}|\text{sound}) \quad \leftarrow \\
p(\text{part}|\text{sound}) \quad \leftarrow \\
p(\text{primitive}|\text{sound}) \quad \leftarrow \\
p(\text{sound waveform})
\end{align*}
\]
Primitive Probabilistic Auditory Scene Analysis

- $p(\text{source})$
- $p(\text{parts}|\text{source})$
- $p(\text{primitive}|\text{part})$
- $p(\text{sound}|\text{primitives})$
- $p(\text{sound}|	ext{waveform})$
- $p(\text{source}|	ext{sound})$
- $p(\text{part}|	ext{sound})$
- $p(\text{primitive}|	ext{sound})$
- Super schema-based grouping
- Schema-based grouping
- Primitive grouping

- Bird song
- Motif
- AM tones and noise
- Auditory filter activities
- Sound waveform
Old Plus New Heuristic

\[ y(t) = a_1(t) + c_1(t) + c_2(t) + c_3(t) + a_2(t) + c_2(t) + c_4(t) + a_3(t) + c_2(t) + a_4(t) + c_1(t) + c_3(t) \]
Old Plus New Heuristic

\[ y_t = a_{1,t}(c_{1,t} + c_{2,t} + c_{3,t}) + a_{2,t}(c_{2,t} + c_{4,t}) + a_{3,t}c_{2,t} + a_{4,t}(c_{1,t} + c_{3,t}) \]
Old Plus New Heuristic

\[ y_t = a_{1,t}(c_{1,t} + c_{2,t} + c_{3,t}) + a_{2,t}(c_{2,t} + c_{4,t}) + a_{3,t}c_{2,t} + a_{4,t}(c_{1,t} + c_{3,t}) \]
Comodulation Masking Release

\[ y(t) = a_{1,t} c_{1,t} + a_{2,t} c_{2,t} \]

- Signal 1
- Signal 2
Comodulation Masking Release

\[ y_t = a_{1,t} c_{1,t} + a_{2,t} c_{2,t} \]
Comodulation Masking Release

\[ y_t = a_{1,t}c_{1,t} + a_{2,t}c_{2,t} \]
Inference with fixed amplitudes, $a_{d,t} = 1$
Probabilistic Model

\[ y_t = \sum_{d=1}^{D} c_{d,t} \alpha_{d,t} + \sigma_y \epsilon_t \]
Probabilistic Model

- **Envelopes**: slow and +/−
  \[ p(x_{e,1:T} | \Gamma_{e,1:T}) = \text{Norm}(x_{e,1:T}; 0, \Gamma_{e,1:T}) \]
  \[ \Gamma_{e,t} = \sigma^2_e \exp\left(\frac{-1}{2\tau^2_e}(t-t')^2\right) \]

- **Envelopes**: positive mixture of transformed envelopes with controllable sparsity
  \[ a_{d,t} = a(\sum e x_{e,t} + \mu_d), \text{e.g.} a(z) = \log(1 + \exp(z)) \]

- **Carriers**: band-limited Gaussian noise
  \[ p(c_{d,t} | c_{d,t-1:t-2}, \theta) = \text{Norm}(c_{d,t}; \sum_{t'=1}^{2} \lambda_{d,t'} c_{d,t-t'}, \sigma^2_d) \]
  \[ y_t = \sum_{d=1}^{D} c_{d,t} a_{d,t} + \sigma_y \epsilon_t \]
Probabilistic Model

- **Envelopes**: positive mixture of transformed envelopes with controllable sparsity

\[
a_{d,t} = a \left( \sum_{e=1}^{E} g_{d,e} x_{e,t} + \mu_d \right), \quad \text{e.g.} \quad a(z) = \log(1 + \exp(z)).
\]

- **Carriers**: band-limited Gaussian noise

\[
p(c_{d,t} | c_{d,t-1:t-2}, \theta) = \text{Norm} \left( c_{d,t}; \sum_{t' = 1}^{2} \lambda_{d,t'} c_{d,t-t'}, \sigma_d^2 \right)
\]

\[
y_t = \sum_{d=1}^{D} c_{d,t} a_{d,t} + \sigma_y \epsilon_t
\]
Probabilistic Model

- **envelope patterns**: slow and +/−

\[ p(x_{e,1:T}|\Gamma_{e,1:T,1:T}) = \text{Norm}\left(x_{e,1:T}; 0, \Gamma_{e,1:T,1:T}\right) \]

\[ \Gamma_{e,t-t'} = \sigma^2_e \exp\left(-\frac{1}{2\tau^2_e}(t - t')^2\right) \]

- **Envelopes**: positive mixture of transformed envelopes with controllable sparsity

\[ a_{d,t} = a\left(\sum_{e=1}^{E} g_{d,e} x_{e,t} + \mu_d\right), \text{ e.g. } a(z) = \log(1 + \exp(z)) \]

- **Carriers**: band-limited Gaussian noise

\[ p(c_{d,t}|c_{d,t-1:t-2}, \theta) = \text{Norm}\left(c_{d,t}; \sum_{t'=1}^{2} \lambda_{d,t'} c_{d,t-t'}, \sigma^2_d\right) \]

\[ y_t = \sum_{d=1}^{D} c_{d,t} a_{d,t} + \sigma_y \epsilon_t \]
What do the parameters control?

$\lambda_d$ control the centre frequency and bandwidth of each sub-band

$\sigma^2_d$ control the power in each sub-band

$\tau_e$ controls time-scale of modulators

$\mu_d$ and $\sigma^2_e$ control the modulation depth, skew and the sparsity

$g_{d,e}$ control the patterns of modulation across sub-bands