Auditory scene analysis and the statistics of natural sounds

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
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

Up arrows:
- super schema-based grouping
- schema-based grouping
- primitive grouping
Statistics of sounds

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

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

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

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

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

\[ \text{sound} \]
\[ \text{scope} \]
\[ \text{super schema-based grouping} \]
\[ \text{schema-based grouping} \]
\[ \text{primitive grouping} \]

\[ \text{birdsong} \]
\[ \text{motifs} \]
\[ \text{AM tones & noise} \]
\[ \text{sound waveform} \]
Auditory scene analysis as inference

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

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

Bayes' Theorem

\[
p(\text{primitive}|\text{sound}) = \frac{p(\text{sound}|\text{primitive})p(\text{primitive})}{p(\text{sound})}
\]

super schema-based grouping

schema-based grouping

primitive grouping
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.
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.
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

- 4.1 KHz
- 2.4 KHz
- 1 KHz

filter

Time /s

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

![Graph showing fire sound analysis at different frequencies]

- 4.1 KHz
- 2.4 KHz
- 1 KHz

**Demodulate**

**Filter**

*Time (s)*

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

[Diagram showing waveforms at different frequencies for modulators and carriers, with filtering and demodulation processes indicated.]
Heuristic Analysis: Fire sound

modulators

carriers

demodulate
filter
Heuristic Analysis: Fire sound
Heuristic Analysis: Fire sound

[Diagram showing time and frequency analysis with demodulate and filter labels, and PCA features]

PC
features
Heuristic Analysis: Fire sound

Demodulate
Filter

Time / s

Frequency / kHz

PCA features

demodulate
filter
Heuristic Analysis: Water

![Image of a graph showing the analysis of water with indications of time, frequency, PCA features, and demodulation and filtering processes.]

- Time in seconds (x-axis)
- Frequency in kHz (y-axis)
- PCA features
- Demodulation and filtering

The graph illustrates the analysis process with various time points and frequency values, indicating how water data is processed through demodulation and filtering techniques.
Heuristic Analysis: Speech

PCA features

0.5 1 1.5 2

y
time /s
0.2
0.5
1.1
2.6
6

frequency /kHz
1 2 3
demodulate
filter

demodulate
filter

Y

0.5 1 1.5 2
time /s
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)
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:
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

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

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

envelopes

carriers

signal
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
Generation: adding comodulation

- Frequency /KHz: 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
- Envelope modulation patterns

Diagram showing spectrogram with frequency and time axes and a waveform for Y over time.
Generation: adding comodulation

Envelope modulation patterns
Generation: adding comodulation

[Diagram showing envelope modulation patterns, frequency (KHz) vs. time (s), and a waveform graph.]
Generation: Decreasing time-scale

![Diagram showing 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

frequency /KHz
0.1 0.3 0.7 1.5 3.3

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

envelope modulation patterns

frequency /KHz

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: Decreasing time-scale

envelope modulation patterns

frequency /KHz

0.1
0.3
0.7
1.5
3.3

time /s

0.1
0.3
0.7
1.5
3.3

Y

0 0.05 0.1 0.15 0.2 0.25 0.3

1 14 27
Generation: Decreasing time-scale

envelope modulation patterns

frequency /KHz

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
Generation: Decreasing time-scale

[Diagram showing a spectrogram with frequency on the y-axis and time on the x-axis, and an envelope modulation pattern on the right side.]
Generation: Increasing sparsity

envelope modulation patterns

0 0.05 0.1 0.15 0.2 0.25 0.3
y
time /s
0.1
0.3
0.7
1.5
3.3

frequency /KHz
1 14 27
Generation: Increasing sparsity

envelope modulation patterns

frequency /KHz

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
Generation: Increasing sparsity
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 Synthesis

The image illustrates the process of auditory scene synthesis. It shows the interaction between different components such as envelopes, carriers, and signals over time (time /ms) and frequency (freq /kHz). The diagram represents how these elements combine to create a complex auditory 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

\[ y = a_1, t + a_2, t \]

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

0 50 100

0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

0 80 120 160

0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

time /s

frequency /Hz

y = a_1, t + a_2, t
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

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

Signal 1

Signal 2

Frequency / Hz

Time / s
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
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

- \( p(\text{source}) \)
- \( p(\text{parts}|\text{source}) \)
- \( p(\text{primitive}|\text{part}) \)
- \( p(\text{sound}|\text{primitives}) \)
- \( \text{bird song} \)
- \( \text{motif} \)
- \( \text{AM tones and noise} \)
- \( \text{auditory filter activities} \)
- \( \text{trees} \)
- \( \text{bark} \)
- \( \text{edges} \)
- \( \text{photoreceptor activities} \)
Recognition model: Probabilistic Auditory Scene Analysis

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

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

\[ p(\text{source}|\text{sound}) \]
\[ p(\text{part}|\text{sound}) \]
\[ p(\text{primitive}|\text{sound}) \]

bird song
motif
AM tones and noise
auditory filter activities
sound waveform

super schema–based grouping
schema–based grouping
primitive grouping
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

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

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

\[
\begin{align*}
p(c_{d,t}|c_{d,t-1:t-2}, \theta) &= \text{Norm} \left( c_{d,t}; \sum_{t'=1}^{2} \lambda_{d,t'} c_{d,t-t'}, \sigma_d^2 \right) \\
y_t &= \sum_{d=1}^{D} c_{d,t} a_{d,t} + \sigma_y \epsilon_t
\end{align*}
\]
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 $+/-$

$$p(x_{e,1:T}|\Gamma_{e,1:T,1:T}) = \text{Norm}(x_{e,1:T}; 0, \Gamma_{e,1:T,1:T})$$

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

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

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

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

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

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

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

\( \sigma_{d}^{2} \) control the power in each sub-band

\( \tau_{e} \) controls time-scale of modulators

\( \mu_d \) and \( \sigma_{e}^{2} \) control the modulation depth, skew and the sparsity

\( g_{d,e} \) control the patterns of modulation across sub-bands