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
- **birds**
  - motifs
    - AM tones & noise
      - sound waveform

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

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

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

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

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

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

birdsong \ \rightarrow \ \text{motifs} \ \rightarrow \ \text{sound waveform}

\[ \text{AM tones & noise} \]

\[ \text{primitive grouping} \]

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

\[ \text{schema-based grouping} \]
Auditory scene analysis as inference

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

Bayes' Theorem

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

- primitive grouping
- schema-based grouping
- super schema-based grouping
Probabilistic primitive auditory scene analysis

Bayes' Theorem
\[ p(primitive|sound) = \frac{p(sound|primitive)p(primitive)}{p(sound)} \]
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

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.
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

- 4.1 KHz
- 2.4 KHz
- 1 KHz

Filter
Heuristic Analysis: Fire sound

Frequency bands and time analysis of sound signals.
Heuristic Analysis: Fire sound

The diagram illustrates the analysis of fire sound signals. The top graph shows the modulators with frequency bands at 4.1 KHz, 2.4 KHz, and 1 KHz. The middle graph depicts the carriers, while the bottom graph represents the time-domain signal with labeled 'y' and 'time /s' axes. The process includes filtering and demodulation steps, indicated by the labels in the diagram.
Heuristic Analysis: Fire sound

modulators

X

 carriers

filter
demodulate

y

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

demodulate
filter
0.5 0.52 0.54 0.56 0.58 0.6 0.62

time /s
0.2
0.6
1.9
5.8
18

frequency /kHz
Heuristic Analysis: Fire sound

PCA features

Demodulate
Filter
Heuristic Analysis: Fire sound

PCA features

demodulate filter

0.5 0.52 0.54 0.56 0.58 0.6 0.62

time /s

0.2
0.6
1.9
5.8
18

frequency /kHz

1 2 3

0.5 0.52 0.54 0.56 0.58 0.6 0.62

time /s

0.2
0.6
1.9
5.8
18

frequency /kHz

1 2 3

PCA features
Heuristic Analysis: Rain

PCA features

Demodulate
Filter
Heuristic Analysis: Water

PCA features

demodulate
filter

0.1 0.15 0.2 0.25 0.3 0.35 0.4

0.2
0.5
1.3
3.2
8

frequency /kHz

time /s

0.1 0.15 0.2 0.25 0.3 0.35 0.4

y

1 6 10
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
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 $\rightarrow$ 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 $\rightarrow$ 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:

sound $\leftarrow$ modulate carriers $\leftarrow$ modulators $\leftarrow$ 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

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

where \( c_d(t) \) is bandpass Gaussian noise.
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

- Frequency / kHz

- Time / s
Generation: adding comodulation

- Frequency (kHz):
  - 0.1
  - 0.3
  - 1
  - 14
  - 27

- Time (s):
  - 0
  - 0.05
  - 0.1
  - 0.15
  - 0.2
  - 0.25
  - 0.3

- Envelope modulation patterns:
  - Linear increase from 1 to 27 over time
Generation: adding comodulation

envelope modulation patterns

<table>
<thead>
<tr>
<th>frequency /KHz</th>
<th>0.1</th>
<th>0.3</th>
<th>0.7</th>
<th>1.5</th>
<th>3.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14</td>
<td>27</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>time /s</th>
<th>0</th>
<th>0.05</th>
<th>0.1</th>
<th>0.15</th>
<th>0.2</th>
<th>0.25</th>
<th>0.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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

Envelop modulation patterns

- Frequency / KHz: 1, 14, 27
Generation: adding comodulation

- Frequency / KHz: 0.1, 0.3, 0.7, 1.5, 3.3
- Time / s: 0.1, 0.3, 0.7, 1.5, 3.3
- Envelope modulation patterns: 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3

Diagram shows the effect of adding comodulation on a signal with time-domain and frequency-domain representations.
Generation: Decreasing time-scale

envelope modulation patterns

1 14 27
Generation: Decreasing time-scale

![Spectrogram and waveform for decreasing time-scale generation.](image)
Generation: Decreasing time-scale

Envelope modulation patterns

frequency /KHz

time /s

y

0 0.05 0.1 0.15 0.2 0.25 0.3
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

1 14 27
Generation: Decreasing time-scale

---

**envelope modulation patterns**

---

**time /s**

- 0.1
- 0.3
- 0.7
- 1.5
- 3.3

**frequency /KHz**

- 0.1
- 0.3
- 1
- 14
- 27

---

**Y**

---

**time /s**

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

- Decreasing time-scale results in
- Changes in envelope modulation patterns
- Illustrated with a spectrogram and waveform
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: 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

y
Generation: Increasing sparsity

envelope modulation patterns

frequency /KHz

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

1 14 27
Generation: Increasing sparsity

- Increasing sparsity
- Envelope modulation patterns

- Frequency (kHz): 0.1 to 3.3
- Time (s): 0 to 0.3

Graph showing time-frequency analysis with increasing sparsity.
Generation: Increasing sparsity

- **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**:
- **Values**:
  - 1
  - 14
  - 27
Sound Generation Demo

- fire
- stream
- wind
- rain
- tent-zip
- foot step

Turner, 2010
Relationship to McDermott and Simoncelli

- compute statistics from a target sound
- iteratively shape white noise until it matches statistics of target
<table>
<thead>
<tr>
<th>McDermott et al.</th>
<th>Turner et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>model free</td>
<td>model-based</td>
</tr>
<tr>
<td>fast to “train”</td>
<td>slow to train</td>
</tr>
<tr>
<td>slow to generate</td>
<td>fast to generate</td>
</tr>
<tr>
<td>rich selection of statistics</td>
<td>narrower selection of statistics</td>
</tr>
<tr>
<td>other statistics uncontrolled</td>
<td>clear which statistics are imposed</td>
</tr>
<tr>
<td>cannot handle mixtures</td>
<td>can handle mixtures (ASA/denoising/source separation)</td>
</tr>
</tbody>
</table>

$\implies$ complementary approaches
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.
What’s missing from current descriptions of auditory scene analysis?

- where do the grouping rules come from? is there a compact description?
- can we predict how the plethora of rules trade-off against one another? can we make quantitative predictions for complex stimuli?
- how do we incorporate uncertainty into the description? (noisy mixtures)
Primitive Probabilistic Auditory Scene Synthesis

envelope patterns

envelopes

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

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

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

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

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

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

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

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

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

\[ y_t = a_{1,t}c_{1,t} + a_{2,t}c_{2,t} \]
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
Probabilistic signal processing \( \rightarrow \) Classical signal processing

\[ \text{Signal} \overset{\text{operations}}{=} \text{Signal} \]
Probabilistic signal processing

Classical signal processing

fixed Gaussian process
Probabilistic signal processing ? Classical signal processing

- Fixed Gaussian process
- STFT filter bank & Hilbert spectrogram
Probabilistic signal processing

Classical signal processing

robustness adaptation

fast methods important variables

fixed
Gaussian process

estimation

STFT filter bank & Hilbert spectrogram

fixed
Gaussian process
Denoising

SNR improvement /dB

SNR before /dB

PESQ improvement

PESQ before

SNR log–spec improvement /dB

SNR log–spec before /dB

$Y_t$

time /ms

$Y_t$

time /ms

NMF
tNMF
GTF
GTFtNMF
Old Plus New Heuristic

\[ y(t) = a_1(t) + c_1(t) + c_2(t) + a_2(t) + c_2(t) + c_3(t) + a_3(t) c_2(t) + a_4(t) c_1(t) + c_3(t) \]
$y_t = a_{1,t}(c_{1,t} + c_{2,t} + c_{3,t}) + a_{2,t}(c_{2,t} + c_{4,t}) + a_{3,t}c_{2,t} + a_{4,t}(c_{1,t} + c_{3,t})$
Old Plus New Heuristic

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