Computational Perception
15-485/785

Auditory Scene Analysis
A framework for auditory scene analysis

- Auditory scene analysis involves low and high level cues
- Low level acoustic cues are often result in spontaneous grouping
- High level cues can be under attentional control
Cues for auditory grouping

- temporal separation
- spectral separation
- harmonicity
- timbre
- temporal onsets and offsets
- temporal/spectral modulations
- spatial separation
Bregman demo 1.
Perceptual streams can be masked

Bregman demo 2. “Auditory camouflage”
Stream segregation and rhythmic information

Bregman demo 3. The perception of rhythmic information depends on the segmentation of the auditory streams.
Streaming in African xylophone music

Bregman demo 7.
Effect of pitch range

Bregman demo 8.
Effect of timbre

Bregman demo 9.
Stream segregation based on spectral peak position

Bregman demo 10.
Effect of connectedness

Bregman demo 12.
Grouping based on common fundamental frequency

Bregman demo 18.

- Do the set of frequencies come from the same source?
- Those of a common fundamental a grouped together.
Fusion based on common frequency modulation

Bregman demo 19.
Fusion based on common frequency modulation

Bregman demo 20.
Onset rate affects segregation

A FOUR-TONE CLUSTER

Tone 1 = 800 hz
Tone 2 = 700 hz
Tone 3 = 900 hz
Tone 4 = 800 hz

tone 1 envelope

Amplitude (arbitrary units)

time in seconds

Bregman demo 21.
Rhythmic masking release

- This is an example of temporal grouping with overlapping frequency ranges.

- Frequencies outside the “critical band” of a target influence the ability to hear it.

Bregman demo 22.
Micro modulation in voice perception

Bregman demo 24.

- Normal speech vowels contain small, random fluctuations
- The pure tone plus harmonic has weak vowel quality
- Addition of micro-modulation enhances the perception of the sound as a singing vowel
Blind source separation
Modeling the cocktail party problem

Suppose we have two speakers (sources), \( s_1(t) \) and \( s_2(t) \) and two microphones (mixtures), \( x_1(t) \) and \( x_2(t) \):

The general problem is called *blind source separation*. We only observe the mixtures and are "blind" to the sources.

How do we model this in general?
A general formulation

Suppose we have $M$ sources

$$s_1(t), \ldots, s_M(t)$$

and $M$ mixtures

$$x_1(t), \ldots, x_2(t)$$

This can be represented diagrammatically as:

```
s_1(t) \rightarrow A \rightarrow x_1(t)

s_2(t) \rightarrow x_2(t)

\vdots

s_M(t) \rightarrow x_M(t)
```
A general formulation

Suppose we have $M$ sources

$$s_1(t), \ldots, s_M(t)$$

and $M$ mixtures

$$x_1(t), \ldots, x_M(t)$$

This can represented diagramatically as:

What’s the simplest mathematical model?

Assume linear, instantaneous mixing:

$$x(t) = As(t)$$
A general formulation

Suppose we have $M$ sources

$$s_1(t), \ldots, s_M(t)$$

and $M$ mixtures

$$x_1(t), \ldots, x_M(t)$$

This can represented diagramatically as:

What’s the simplest mathematical model?

Assume linear, instantaneous mixing:

$$x(t) = As(t)$$

$A$ is called the *mixing matrix*.

What does this model ignore?
A general formulation

Suppose we have $M$ sources

$$s_1(t), \ldots, s_M(t)$$

and $M$ mixtures

$$x_1(t), \ldots, x_M(t)$$

This can be represented diagrammatically as:

What’s the simplest mathematical model?

Assume linear, instantaneous mixing:

$$x(t) = As(t)$$

$A$ is called the *mixing matrix.*

What does this model ignore?

- room acoustics, reverberation, echoes
- filtering, noise
- might have more than two sounds
- sounds might not come from a single source
- sound sources could change location
The distribution of sample points

- The histograms show the amplitude distribution of each source.
- The scatter plot shows the 2D distribution of $x_1(t)$ vs $x_2(t)$.
- The parameter $\beta$ characterizes the sharpness of the data using the distribution

$$P(s) \propto e^{-|s|^2/(1+\beta)/2}$$

- $\beta = 0$ corresponds to a Gaussian distribution.
The distribution of sample points

- Sources can have a wide range of amplitude distributions.
- Different directions correspond to different mixing matrices.
- A mixture of non-Gaussian sources create a distribution whose axes can be uniquely determined (within a sign change).
The distribution of sample points

The axes of Sub-Gaussian sources, i.e. $\beta < 0$ or negative kurtosis, can still be determined.
A Gaussian distribution of sample points

The principal axes of two Gaussian sources are ambiguous.

- Why?

- Because the product of two Gaussians is still a Gaussian, so there are an infinite number of directions that fit the 2D distribution.
Inferring the (un)mixing matrix

How do we determine the axes from just the data?
Modeling non-Gaussian distributions

Learning objective: model statistical density of sources:
\[ \Rightarrow \text{maximize } P(x|A) \text{ over } A. \]

Probability of pattern ensemble is:
\[
P(x_1, x_2, \ldots, x_N|A) = \prod_k P(x_k|A)
\]

To obtain \( P(x|A) \) marginalize over \( s \):
\[
P(x|A) = \int ds \, P(x|A, s)P(s) = \frac{P(s)}{|\det A|}
\]

Learning rule (ICA):
\[
\Delta A \propto AA^T \frac{\partial}{\partial A} \log P(x|A)
= -A(zs^T - I),
\]

where \( z = (\log P(s))' \). Use \( P(s_i) \sim \text{ExPwr}(s_i|\mu, \sigma, \beta_i) \).

This is identical to the procedure that was used to learn efficient codes, i.e. independent component analysis.
Separating mixtures of real sources
Separating mixtures of real sources
Separating mixtures of real sources

Source 1
Mixture 2
Mixture 4
Source 2
Source 3
Mixture 3
Mixture 1
Source 4
Separating mixtures of real sources

Source 1

Mixture 2

Source 2

Mixture 4

Source 3

Mixture 3

Source 4
Separating mixtures of real sources
Separating mixtures of real sources
Separating mixtures of real sources

Source 1

Mixture 1

Mixture 2

Source 2

Mixture 3

Source 3

Source 4

Mixture 4
Separating mixtures of real sources
Separating mixtures of real sources
Separating mixtures of synthetic sources: 4 sources, 4 mixtures

<table>
<thead>
<tr>
<th>source</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>avg. SNR (dB)</td>
<td>40.48</td>
<td>4.84</td>
<td>4.71</td>
<td>17.29</td>
<td>35.03</td>
<td>43.07</td>
<td>45.85</td>
<td>44.10</td>
</tr>
<tr>
<td>std. dev.</td>
<td>1.35</td>
<td>0.39</td>
<td>0.42</td>
<td>2.08</td>
<td>1.49</td>
<td>0.50</td>
<td>1.75</td>
<td>2.00</td>
</tr>
</tbody>
</table>

- Experiment: recover synthetic sources from a random mixing matrix. Repeat 5 times.
- SNR reflects the accuracy of inferring the mixing matrix.
- The near-Gaussian sources cannot be accurately recovered.
Computational auditory scene analysis

How do we incorporate these ideas in a computational algorithm?

Blind source separation (ICA) only solves special case

- non-Gaussian sources
- linear, stationary mixtures
- equal number of sources and mixtures
- need at least two mixtures

What about approaches that use auditory grouping cues?
Cooke (1993)

“Modeling Auditory Processing and Organisation”

- uses gammatone filter bank to model auditory periphery
- auditory “objects” are represented by “synchrony strands”
- synchrony strands are formed according to auditory grouping principles

Want an auditory representation that

- describes the auditory scene in terms of time-frequency objects
- characterizes onsets, offsets, and movement of spectral components
- allows for further parameters to be calculated
Overview of stages in synchrony strand formation

- Gammatone filter output
- Extraction of dominant frequency components in each channel
- Dominant frequencies in neighbouring channels
- Cross-channel grouping
- Place-groups
- Temporal correspondence of place-groups
- Synchrony strands
Auditory periphery model

- gammatone filter
- static nonlinearity
- hair cell model

Impulse responses

(normalised response)

(output (arbitrary units))

time (samples/20 kHz)

(input dB)
Hair cell adaptation

**rate-intensity functions (model)**

![Graph showing rate-intensity functions](image)

- **normalised rate**
- **intensity (dB)**
  - --- onset
  - --- adapted

**incremental/decremental responses (model)**

![Graph showing incremental/decremental responses](image)

- **increase in rate**
  - --- 6 dB increment
  - --- 6 dB decrement

Hair cell model is based on adaptive non-linearities of cochlear responses
Next stage: extraction of dominant frequency
Dominant frequency estimation

Want a pure frequency representation in order to apply grouping rules.

a) waveform for the vowel /ae/

b) auditory filter output for 1kHz center frequency

c) instantaneous frequency
   - from Hilbert transform of gammatone filter

d) an alternative method of estimating instantaneous frequency

  e) smoothed instantaneous frequency
     - median filtering over 10ms, plus linear smoothing
     - also allows for data reduction: samples are collapsed over 1ms window

Figure 3.2 Dominant frequency estimation: a: waveform for the vowel /ae/; b: auditory filter output (CF: 1 kHz); c: \( \nu(t) \) by analytic signal method; d: instantaneous frequency by linear prediction analysis; e: \( \hat{\nu}(t) \).
Next stage: grouping frequency channels

- Gammatone filter output
- Extraction of dominant frequency components in each channel
- Dominant frequencies in neighbouring channels
- Cross-channel grouping
- Place-groups
- Temporal correspondence of place-groups
- Synchrony strands
Calculation of place groups

- What frequencies belong to the same sound?
- Goal of this stage is to locate and characterize intervals along filterbank with synchronous activity.
- (a) Notice that center frequency is not distributed evenly due to the median filtering in the frequency estimation stage.
- (b) The instantaneous frequency varies around the estimated frequency.
- Idea is to group all channels centered around the “dominant frequency”, i.e. grouping by spectral location.
- (c) $S(f)$ is a smoothed frequency derivative estimate. Channels at the minimum are grouped together.
Plot of place groups for a speech waveform

Figure 3.4 Place-groups for the utterance whose waveform is shown. Frequency axis is linear in Hz.
Waveform resynthesized from place groups

Resynthesize waveform by summing sines in place group

Figure 3.5  Top: Original waveform for the utterance “I’ll willingly marry Marilyn”. Bottom: Resynthesised waveform.

Reconstruction quality is good

- best is indistinguishable from original
- worst is still clearly intelligible
- best for harmonic sounds
More synchrony strand displays: female speech

Figure 3.6 Female speech (1.83 s duration).
Synchrony strand display of male speech

Figure 3.7 Male speech (2.5 s duration).
Synchrony strand display of noise burst sequence

**Figure 3.8** Noise burst sequence (1.76 s duration).
Synchrony strand display of new wave music

Figure 3.9  New wave music (2 s duration).
Modeling auditory scene exploration

- The synchrony strand display has done local grouping of frequencies, but this alone does not provide a way to group sounds from the same source.
- The next part of the model develops ways to construct higher order groups

Figure 3.7 Male speech (2.5 s duration).
A hierarchical view of auditory scene analysis

**Top level:** the stable interpretation which results when all organization has been discovered and all competitions resolved.

**Intermediate level:** results from combining groups with similar derived properties such as pitch and contour.

**First level:** Results from the independent application of organizing principles such as harmonicity, common onset etc.

*But it’s not as simple as assigning strands to different sounds*
Duplex perception (Lieberman, 1982)

- Stimulus: a synthetic syllable with 3 formants
- Syllable minus the third formant transition is played to one ear
- Missing formant transition is played to the other ear

What do subjects hear?
- There hear the syllable as if all formants had been presented to one ear.
- But, they also hear an isolated chirp

This suggests that components are shared across groups
Cooke’s framework for auditory scene exploration

1. Seed selection (thick line)

2. Choose ‘simultaneous set’ i.e. those strands which overlap in time with the seed

3. Apply grouping constraint, e.g. suppose the black strands share a common rate of amplitude modulation with the seed

4. Sequential phase: choose strand to extend the temporal basis of the group

5. Back to simultaneous phase: consider for similarity strands that overlap with new seed

6. Seeds may be selected from any in the group which extended in time
Harmonic constraint propagation for a voiced utterance

Seed (highlighted) attracts several supporters to form a harmonic group.

A new focus is chosen, but recruits few new supporters to group.

New focus (f0 itself) successfully attracts virtually all the harmonically related strands in the utterance.

Process halts when no temporal extension to the group is possible.
Implementation of auditory grouping principles

Cooke’s algorithm implements the following grouping principles:

- harmonicity
- common amplitude modulation
- common frequency movement

Also need “subsumption” to form higher level groups.

- Idea is to remove groups that are contained within a larger group
- Groups expand until at some point they are assigned to a source

Also note that the algorithm involves many heuristics which are not covered here.
Harmonic grouping

Harmonic groups discovered by various synchrony strands

Scoring function used in assessing how well strands fit sieve slots

- h1(800)
- f0(400)
- h3(800)
- h2(600)
- h1(400)
- f0(200)
- h5(800)
- h2(400)
- h7(800)
- f0(100)
- h9(800)
- h8(720)
- h4(400)
Harmonic grouping of frequency modulated sounds

Figure 5.2 One of a series of temporally-extended harmonic sieves generated from a seed strand. Thin lines represent the +/-3% boundaries of sieve channels. Some strands fall wholly or partly into the sieve, whilst others do not. Strands may contribute to more than one sieve channel.
Harmonic grouping for synthetic 4-formant syllable

1^{st}, 3^{rd}, and 4^{th} formant are synthesized on a fundamental of 110Hz

2^{nd} formant synthesized on fundamental of 174Hz

Algorithm finds two groups
Grouping by common amplitude modulation
Grouping by common frequency movement

- common FM grouping
- harmonicity grouping
Subsumption: forming higher level groups

Remove groups that are contained within a larger group.
Grouping natural speech
Test sounds and noise sources

n0: 1 kHz tone (25 ms)

n1: random noise (25 ms)

n2: noise bursts (1.76 s)

n3: laboratory noise (2 s)

n4: music (2 s)

n5: siren (50 ms)

n6: telephone (1.83 s)

n7: female (2.37 s)

n8: male (2.5 s)

n9: female (1.83 s)
Synchrony strand representations of test sounds

- n0: 1 kHz tone
- n1: random noise
- n2: noise bursts
- n3: laboratory noise
- n4: music
- n5: siren
- n6: telephone
- n7: female
- n8: male
- n9: female
The mixture correspondence problem
Performance: Utterance characterization

no intrusion

percent

noise source type

grouped

random

79 76 67 78 67 78 72 78 75 76 67
42 19 41 21 21 30 37 40 43 32
Performance: crossover from intrusive source

How much is the intrusive source incorrectly grouped?

- Least intrusion when intrusive source is noise bursts (n2)
- Worst is when intrusive source is music (n4)
Grouping of speech from a mixture (cocktail party)
Grouping speech in presence of laboratory noise