Sound, Mixtures, and Learning: 
A Perspective on CASA

1 Constraints and Scene Analysis
2 Model-Based Organization
3 Evaluation

Dan Ellis  <dpwe@ee.columbia.edu>

Laboratory for Recognition and Organization of Speech and Audio
(LabROSA)
Columbia University, New York
http://labrosa.ee.columbia.edu/
Acoustic/Auditory Scene Analysis

- Scene analysis is sound **understanding**

![Frequency-Time Spectrum](image)

- **Analysis**

<table>
<thead>
<tr>
<th>Voice (evil)</th>
<th>Voice (pleasant)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slab</td>
<td>Choir</td>
</tr>
<tr>
<td>Rumble</td>
<td>Strings</td>
</tr>
</tbody>
</table>

- understanding = **abstraction**

- **Applications**
  - robust interfaces
  - robots
  - indexing/retrieval
  - prostheses
The Mixture Problem

“Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?” (after Bregman’90)

- **Objects (sources), not waveforms**
  - .. and only their attributes “of interest”

- **Seems highly underconstrained**

- **But: Hearing is ecologically grounded**
  - reflects natural scene properties = constraints
  - subjective, not absolute
The Signal Separation Perspective

- Search for a representation / parameterization in which sources become separate

- Inverse filter & cancel (ICA, beamforming)

- TF-mask: find distinct time-freq support

- Innate limitations with dense maskers

![Graph showing signal separation and parameterization](image-url)
The Pattern Recognition Perspective

- **Bayes Rule:**
  Event / Model $M$,
  Evidence / observation $x$:
  \[
  Pr(M|x) = \frac{p(x|M) \cdot Pr(M)}{p(x)}
  \]

- **Trained signal model** $p(x|M)$
  - fit to training examples of $x$ under $M$
  - uncertainty from observation noise / ignorance

- **Uncertainty in** $Pr(M|x)$
  - from unambiguous separation ...
  - ... to hopeful guess

- **Structure of** $p(x|M) \cdot Pr(M)$
  - the possibilities under consideration
  - constraints on solution
Separation vs. Recognition

- Final goal is scene **abstraction**: Do we need signal separation?
  - separate-then-recognize is a nice approach
    - if you can separate
  - classification is often still possible when separation is hopeless

- **Classification/Recognition**
  - can express ambiguous answers
  - still applicable when data is missing (based on ignorance)

- “**Perceiving is more than recognizing**”
  - identify class
    + extract **parameters** of instance
      .. for description of scene
Constraints in Scene Analysis

- **Learned constraints** are central to human speech recognition
  - click-language example
  - foreign-language cocktail party
  - ... not just for speech

- **Computational systems need similar ‘constraints’ on real-world sounds**
  - hand-specify rules?
  - or: learn from examples?
Outline

1 Constraints and Scene Analysis

2 Model-Based Organization
   - Missing-Data Recognition
   - Comparing Segregation Masks
   - Multi-Source Decoding

3 Evaluation
Model-based Organization: Sound Fragment Decoding
(Cooke et al. ’01; Barker, Cooke & Ellis)

• Signal separation is too hard!
  Instead:
  - segregate features into partially-observed sources
  - then classify

• Made possible by missing data recognition
  - integrate over uncertainty in observations

• Goal:
  Relate clean speech models \( P(X|M) \)
  to speech-plus-noise mixture observations
  - .. and make it tractable
Missing Data Recognition

- **Speech models** $p(x|m)$ are multidimensional...
  - i.e. means, variances for every freq. channel
  - need values for all dimensions to get $p(\bullet)$

- **But: can evaluate over a subset of dimensions** $x_k$
  
  $$p(x_k|m) = \int p(x_k, x_u|m)dx_u$$

- **Hence, missing data recognition:**
  - hard part is finding the mask (segregation)
  
  $$P(x \mid q) = P(x_1 \mid q) \cdot P(x_2 \mid q) \cdot P(x_3 \mid q) \cdots P(x_6 \mid q)$$
Comparing Segregation Masks

- **Standard classification chooses between models** $M$ **to match source features** $X$

$$ M^* = \arg\max_M P(M|X) = \arg\max_M P(X|M) \cdot \frac{P(M)}{P(X)} $$

- **Mixtures**: **observed features** $Y$, **segregation** $S$, all related by $P(X|Y,S)$:

$$ P(M^*, S|M, Y) = P(M) \int P(X|M) \cdot \frac{P(X|Y, S)}{P(X)} dX \cdot P(S|Y) $$

($P(X)$ no longer constant)
Calculating fragment matches

\[
P(M, S|Y) = P(M) \int P(X|M) \cdot \frac{P(X|Y,S)}{P(X)} dX \cdot P(S|Y)
\]

- \( P(X|M) \) - the clean-signal feature model
- \( P(X|Y,S)/P(X) \) - is \( X \) ‘visible’ given segregation?
- Integration collapses some bands...
- \( P(S|Y) \) - segregation inferred from observation
  - just assume uniform, find \( S \) for most likely \( M \)
  - or: use extra information in \( Y \) to distinguish \( S \)’s...
- **Result:**
  - probabilistically-correct relation between clean-source models \( P(X|M) \) and inferred, recognized source + segregation \( P(M,S|Y) \)
Using CASA features

• $P(S|Y)$ links acoustic information to segregation
  - is this segregation worth considering?
  - how likely is it?

• Opening for CASA-style local features
  - periodicity/harmonicity:
    frequency bands belong together
  - onset/continuity:
    time-frequency region must be whole
Fragment decoding

- Limiting $S$ to whole fragments makes hypothesis search tractable:

- choice of fragments reflects $P(S|Y) \cdot P(X|M)$
  i.e. best combination of segregation and match to speech models

- Merging hypotheses limits space demands
  - .. but erases specific history
Speech fragment decoder results

- Simple $P(S|Y)$ model forces contiguous regions to stay together
  - big efficiency gain when searching $S$ space

- Clean-models-based recognition rivals trained-in-noise recognition
Multi-Source Decoding

- Match multiple models at once?

- disjoint subsets of cells for each source
- each model match $P(M_x|S_x,Y)$ is independent
- masks are mutually dependent: $P(S_1,S_2|Y)$
Model-Based Organization: Summary

- **Results constrained by source model** $P(X|M)$
  - single, ideal clean-signal model

- **Local signal cues introduced via** $P(S|Y)$
  - limited subset of segregations are considered
  - opening for bottom-up CASA cues

- **Output is classification** $M^*$
  - could do TF-mask filtering, but not the point
Outline

1. Constraints and Scene Analysis
2. Model-Based Organization
3. Evaluation
   - Tasks
   - Domains
Evaluation: Tasks

- **Evaluation standards** make research **fundable**
  - sponsors want tangible progress

- **The DARPA / ASR experience**
  - pro: able to judge relative merits
  - con: **extinction** of ‘2nd-best’ techniques
    - neglected aspects e.g. source separation

- **Minimize pathologies by:**
  - defining a ‘real’ task - get something useful
  - allowing ‘ecological niches’
Scene Analysis Task Example

-70 dB

200 400 1000 2000 4000 f/Hz

City

200 400 1000 2000 4000 f/Hz

Noise1

200 400 1000 2000 4000 f/Hz

Noise2, Click1

200 400 1000 2000 4000 f/Hz

Wefts1–4

200 400 1000 2000 4000 f/Hz

Weft5

200 400 1000 2000 4000 f/Hz

Wefts6,7

200 400 1000 2000 4000 f/Hz

Wefts8

200 400 1000 2000 4000 f/Hz

Wefts9–12

Horn1 (10/10)

Horn2 (5/10)

Horn3 (5/10)

Horn4 (8/10)

Horn5 (10/10)

Crash (10/10)

Noise1

Squeal (6/10)

Truck (7/10)
Domains: Personal Audio

- LifeLog / MyLifeBits / Remembrance Agent: Easy to record everything you hear

- Then what?
  - prohibitively time consuming to search
  - but .. applications if access easier

- Automatic content analysis / indexing...

![Graph showing frequency analysis over time with clock time and frequency in Hz and Bark scales.]
Domains: ICSI Meeting Recorder Corpus

- Real meetings, 16 channel recordings, 80 hrs
- released through NIST/LDC

- Lots of speaker overlap, noise, etc.

- Spkr A: speaker active
- Spkr B: speaker B cedes floor
- Spkr C: interruptions
- Spkr D: breath noise
- Spkr E: crosstalk
Summary

- Scene analysis is abstraction of objects
- Real-world constraints come from sound models
- Speech Fragment Decoding finds best model, best segregation
  - without too much search

- Field needs standardized, ‘real-world’ evaluation task