Auditory Scene Analysis in Humans and Machines

Dan Ellis

Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

dpwe@ee.columbia.edu  http://labrosa.ee.columbia.edu/

1. The ASA Problem
2. Human ASA
3. Machine Source Separation
4. Systems & Examples
5. Concluding Remarks
Auditory Scene Analysis

- Sounds rarely occurs in isolation
  - ... but recognizing sources in mixtures is a problem
  - ... for humans and machines

[Image of spectrograms and waveforms]
Sound Mixture Organization

- **Goal:** recover individual sources from scenes
  - .. duplicating the perceptual effect

- **Problems:** competing sources, channel effects

- **Dimensionality loss**
  - need additional constraints
The Problem of Mixtures

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)

- Received waveform is a mixture
  - 2 sensors, N sources - underconstrained
- Undoing mixtures: hearing’s primary goal?
  - .. by any means available
Source Separation Scenarios

- Interactive voice systems
  - human-level understanding is expected
- Speech prostheses
  - crowds: #1 complaint of hearing aid users
- Archive analysis
  - identifying and isolating sound events

- Unmixing/remixing/enhancement...
How Can We Separate?

• By **between-sensor differences** (spatial cues)
  - ‘steer a null’ onto a compact interfering source

• By finding a ‘**separable representation**’
  - spectral? sources are broadband but sparse
  - periodicity? maybe – for pitched sounds
  - something more signal-specific...

• By **inference** (based on knowledge/models)
  - acoustic sources are **redundant**
    - → use part to guess the remainder
Outline

1. The ASA Problem
2. Human ASA
   - scene analysis
   - separation by location
   - separation by source characteristics
3. Machine Source Separation
4. Systems & Examples
5. Concluding Remarks
Auditory Scene Analysis

- Listeners organize sound mixtures into discrete perceived sources based on within-signal cues (audio + ...)
  - common onset
  - continuity
  - harmonicity
  - spatial, modulation, ...
  - learned “schema”

Reynolds-McAdams oboe

Lab ROSA

Auditory Scene Analysis - Dan Ellis
Perceiving Sources

- **Harmonics** distinct in ear, but perceived as one source ("fused"): depends on common onset, depends on harmonics.

- **Experimental techniques**
  - ask subjects "how many"
  - match attributes e.g. pitch, vowel identity
  - brain recordings (EEG "mismatch negativity")
Auditory Scene Analysis

• How do people analyze sound mixtures?
  - break mixture into small elements (in time-freq)
  - elements are grouped into sources using cues
  - sources have aggregate attributes

• Grouping rules (Darwin, Carlyon, ...):
  - cues: common onset/offset/modulation, harmonicity, spatial location, ...

![Diagram of Auditory Scene Analysis](after Darwin 1996)
Streaming

- Sound event sequences are organized into streams
  - i.e. distinct perceived sources
  - difficult to make comparisons between streams
- Two-tone streaming experiments:
  - ecological relevance?
Illusions & Restoration

- Illusion = hearing more than is "there"
  - e.g. "pulsation threshold"
    - example - tone is masked
  - "old-plus-new" heuristic:
    - existing sources continue

- Need to infer most likely real-world events
  - observation equally good match to either case
  - prior likelihood of continuity much higher
Human Performance: Spatial Separation

- **Task:** Coordinate Response Measure
  - “Ready Baron go to green eight now”
  - 256 variants, 16 speakers
  - correct = color and number for “Baron”

- **Accuracy as a function of spatial separation:**

  ![Graph](crm-11737+16515.wav)

- A, B same speaker
- Range effect
Separation by Vocal Differences

- CRM varying the level and voice character
  - (same spatial location)

- energetic vs. informational masking

Brungart et al.'01
Varying the Number of Voices

- Two voices OK;
  More than two voices harder
  (same spatial origin)

- mix of $N$ voices tends to speech-shaped noise...
Outline

1. The ASA Problem
2. Human ASA
3. Machine Source Separation
   - Independent Component Analysis
   - Computational Auditory Scene Analysis
   - Model-Based Separation
4. Systems & Examples
5. Concluding Remarks
Scene Analysis Systems

• “Scene Analysis”
  ○ not necessarily separation, recognition, ...
  ○ scene = overlapping objects, ambiguity

• General Framework:
  ○ distinguish input and output representations
  ○ distinguish engine (algorithm) and control (constraints, “computational model”)
Human and Machine Scene Analysis

- CASA (e.g. Brown’92):
  - **Input**: Periodicity, continuity, onset “maps”
  - **Output**: Waveform (or mask)
  - **Engine**: Time-frequency masking
  - **Control**: “Grouping cues” from input
    - or: spatial features (Roman, ...)

Multiple sources → sound → Input representation → Separation engine → Output → distinct descriptions

Evaluation/control
Human and Machine Scene Analysis

- CASA (e.g. Brown’92):

- ICA (Bell & Sejnowski et seq.):
  - Input: waveform (or STFT)
  - Output: waveform (or STFT)
  - Engine: cancellation
  - Control: statistical independence of outputs
    - or energy minimization for beamforming
Human and Machine Scene Analysis

• CASA (e.g. Brown’92):
• ICA (Bell & Sejnowski et seq.):
• Human Listeners:
  ▪ Input: excitation patterns ...
  ▪ Output: percepts ...
  ▪ Engine: ?
  ▪ Control: find a plausible explanation
Machine Separation

- Problem: **Features of combinations are not combinations of features**

- Voice is easy to characterize when in isolation
- Redundancy needed for real-world communication

---

**Solo Voice**

**M+F Voice Mix**

**Original**

- crm-11-070307
- crm-11-070307+16-050105

**MFCC Noise resynth**

- crm-11-070307-noise
- crm-11-070307+16-050105-noise

---

Auditory Scene Analysis - Dan Ellis

2006-05-20 - 21/54
Separation Approaches

ICA
- Multi-channel
- Fixed filtering
- Perfect separation – maybe!

CASA / Model-based
- Single-channel
- Time-varying filtering
- Approximate Separation

• Very different approaches!
**Independent Component Analysis**

- Central idea:
  
  Search **unmixing space**
  
  to maximize independence of outputs

  \[
  \begin{bmatrix}
  a_{11} & a_{12} \\
  a_{21} & a_{22}
  \end{bmatrix}
  \times
  \begin{bmatrix}
  s_1 \\
  s_2
  \end{bmatrix}
  \rightarrow
  \begin{bmatrix}
  x_1 \\
  x_2
  \end{bmatrix}
  \]

  - simple mixing
  
  \[\text{→ a good solution (usually) exists}\]
Mixtures, Scatters, Kurtosis

- **Mixtures** of sources become more Gaussian
  - can measure e.g. via ‘kurtosis’ (4th moment)

\[
p(s_1) \text{ kurtosis} = 27.90 \\
p(s_2) \text{ kurtosis} = 53.85 \\
p(x_1) \text{ kurtosis} = 18.50 \\
p(x_2) \text{ kurtosis} = 27.90
\]
ICA Limitations

- **Cancellation** is very finicky
  - hard to get more than ~ 10 dB rejection

- The world is not instantaneous, fixed, linear
  - subband models for reverberation
  - continuous adaptation

- Needs **spatially-compact** interfering sources

---

Mixture Scatter

Kurtosis vs. $\theta$

from Parra & Spence’00

lee_ss_in_1.wav  lee_ss_para_1.wav
Central idea:

Segment time-frequency into sources based on perceptual grouping cues

... principal cue is harmonicity
CASA Preprocessing

- **Correlogram**: a 3rd “periodicity” axis
  - envelope of wideband channels follows pitch

- c/w Modulation Filtering [Schimmel & Atlas ’05]

---

Auditory Scene Analysis - Dan Ellis

2006-05-20 - 27/54
“Weft” Periodic Elements

- Represent harmonics without grouping?

- hard to separate multiple pitch tracks

Ellis ’96
Time-Frequency (T-F) Masking

- "Local Dominance" assumption

- Oracle masks are remarkably effective!
  \[ |mix - \max(male, female)| < 3\text{dB} \text{ for } \sim 80\% \text{ of cells} \]
Combining Spatial + T-F Masking

- **T-F masks** based on inter-channel properties
  - [Roman et al. ’02], [Yilmaz & Rickard ’04]
  - multiple channels make CASA-like masks better

- **T-F masking after ICA**
  - [Blin et al. ’04]
  - cancellation can remove energy within T-F cells
CASA limitations

- **Driven by local features**
  - problems with masking, aperiodic sources...

- **Limitations of T-F masking**
  - need to identify single-source regions
  - cannot undo overlaps – leaves gaps

From Hu & Wang '04

---

Auditory Scene Analysis - Dan Ellis

2006-05-20 - 31/54
Auditory “Illusions”

- How do we explain illusions?
  - pulsation threshold
  - sinewave speech
  - phonemic restoration

- Something is providing the missing (illusory) pieces ... source models
Adding Top-Down Constraints

• Bottom-up CASA: **limited** to what’s “there”

  - Bottom-up CASA: limited to what’s “there”
  - Top-down predictions allow **illusions**

  - match observations to a “**world-model**”...
Separation vs. Inference

• **Ideal** separation is rarely possible
  - i.e. no projection can completely remove overlaps

• **Overlaps** ⇒ **Ambiguity**
  - scene analysis = find “most reasonable” explanation

• **Ambiguity** can be expressed **probabilistically**
  - i.e. posteriors of sources \( \{S_i\} \) given observations \( X \):
    
    \[
    P(\{S_i\} \mid X) \propto P(X \mid \{S_i\}) \cdot P(\{S_i\})
    \]

• **Better** source models → better inference
  - .. learn from examples?
Simple Source Separation

- **Given models** for sources, find “best” (most likely) states for spectra:

  \[ p(x|i_1, i_2) = \mathcal{N}(x; c_{i_1} + c_{i_2}, \Sigma) \]

  \[ \{i_1(t), i_2(t)\} = \arg\max_{i_1, i_2} p(x(t)|i_1, i_2) \]

  - can include **sequential constraints**...
  - different **domains** for combining \( c \) and defining \( \Sigma \)

- **E.g. stationary noise:**

  ![Original speech](original.png)  ![In speech-shaped noise (mel magsnr = 2.41 dB)](speech-shaped.png)  ![VQ inferred states (mel magsnr = 3.6 dB)](vq-inferred.png)
Can Models Do CASA?

• **Source models** can learn **harmonicity**, **onset**
  ... to **subsume** rules/representations of CASA

  ![VQ800 Codebook - Linear distortion measure](image)

• can capture **spatial** info too [Pearlmutter & Zador’04]

• **Can also capture** **sequential structure**
  e.g. consonants follow vowels
  ... like people do?

• **But:** need **source-specific models**
  ... for **every possible source**

• use model **adaptation**? [Ozerov et al. 2005]
Separation or Description?

- **Are isolated waveforms required?**
  - Clearly sufficient, but may not be necessary
  - Not part of perceptual source separation!
- **Integrate separation with application?**
  - E.g., speech recognition

- **Words output = abstract description of signal**

**Auditory Scene Analysis - Dan Ellis**

2006-05-20 - 37/54
Missing Data Recognition

- Speech models $p(x|M)$ are multidimensional...
  - need values for all dimensions to evaluate $p(\bullet)$
- But: can make inferences given just 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)

Cooke et al. '01
The Speech Fragment Decoder

- Match ‘uncorrupt’ spectrum to ASR models using missing data

- Joint search for model $M$ and segregation $S$ to maximize:

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

Isolated Source Model

Segregation Model

Observation $Y(f)$
Source $X(f)$
Using CASA cues

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

• **CASA can help search**
  - consider only segregations made from CASA chunks

• **CASA can rate segregation**
  - construct \( P(S|Y) \) to reward CASA qualities:

\[ \begin{align*}
\text{Frequency Proximity} & \quad \text{Common Onset} & \quad \text{Harmonicity}
\end{align*} \]
Outline

1. The ASA Problem
2. Human ASA
3. Machine Source Separation
4. Systems & Examples
   - Periodicity-based
   - Model-based
   - Music signals
5. Concluding Remarks
Current CASA

• State-of-the-art bottom-up separation
  - noise robust pitch track
  - label T-F cells by pitch
  - extensions to unvoiced transients by onset
Prediction-Driven CASA

- Identify objects in real-world scenes
  - using “sound elements”
Singing Voice Separation

- Pitch tracking + harmonic separation

![Singing Voice Separation Diagram](image)
Periodic/Aperiodic Separation

- Harmonic structure + repetition of drums

Virtanen'03
“Speech Separation Challenge”

• Mixed and Noisy Speech ASR task defined by Martin Cooke and Te-Won Lee
  ◦ short, grammatically-constrained utterances:

<command:4><color:4><preposition:4><letter:25><number:10><adverb:4>
  e.g. "bin white at M 5 soon"

• Results to be presented at Interspeech’06
  ◦ http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm

• See also “Statistical And Perceptual Audition” workshop
  ◦ http://www.sapa2006.org/
IBM’s “Superhuman” Separation

- Optimal inference on Mixed Spectra
  - model each speaker (512 mix GMM)

- Applied to Speech Separation Challenge:
  - Infer speakers and gain
  - Reconstruct speech
  - Recognize as normal...

- Use grammar constraints
Transcription as Separation

- **Transcribe** piano recordings by **classification**
  - train SVM detectors for every piano note
  - 88 separate detectors, independent smoothing
- **Trained on** player piano recordings

- **Sse transcription to resynthesize...**
Piano Transcription Results

• Significant improvement from classifier:
  • frame-level accuracy results:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Errs</th>
<th>False Pos</th>
<th>False Neg</th>
<th>$d'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>43.3%</td>
<td>27.9%</td>
<td>15.4%</td>
<td>3.44</td>
</tr>
<tr>
<td>Klapuri &amp; Ryynänen</td>
<td>66.6%</td>
<td>28.1%</td>
<td>38.5%</td>
<td>2.71</td>
</tr>
<tr>
<td>Marolt</td>
<td>84.6%</td>
<td>36.5%</td>
<td>48.1%</td>
<td>2.35</td>
</tr>
</tbody>
</table>

• Breakdown by frame type:

http://labrosa.ee.columbia.edu/projects/melody/
Outline

1. The ASA Problem
2. Human ASA
3. Machine Source Separation
4. Systems & Examples
5. Concluding Remarks
   ⊗ Evaluation
Evaluation

• How to measure separation performance?
  - depends what you are trying to do

• SNR?
  - energy (and distortions) are not created equal
  - different nonlinear components [Vincent et al. ’06]

• Intelligibility?
  - rare for nonlinear processing to improve intelligibility
  - listening tests expensive

• ASR performance?
  - separate-then-recognize too simplistic; ASR needs to accommodate separation

Transmission
errors

Net
effect

Increased
artefacts

Reduced
interference

optimum

Agressiveness
of processing

optimum

Net
effect

Increased
artefacts

Reduced
interference

optimum

Agressiveness
of processing

Transmission
errors
Evaluating Scene Analysis

- Need to establish **ground truth**
  - subjective sources in real sound mixtures?

![Diagram showing a user interface for evaluating scene analysis with names and marks for 'horn1', 'crash', 'squeal', and 'horn2'.]
More Realistic Evaluation

- **Real-world speech tasks**
  - crowded environments
  - applications: communication, command/control, transcription

- **Metric**
  - human intelligibility?
  - ‘diarization’ annotation (not transcription)

![Graph showing pitch track and speaker active ground truth](image-url)
Summary & Conclusions

• **Listeners** do well separating sound mixtures
  - using signal cues (location, periodicity)
  - using source-property variations

• **Machines** do less well
  - difficult to apply enough **constraints**
  - need to exploit signal **detail**

• **Models** capture constraints
  - learn from the real world
  - adapt to sources

• **Separation** feasible in certain domains
  - describing source properties is easier
Sources / See Also

- NSF/AFOSR Montreal Workshops ’03, ’04
  - [www.ebire.org/speechseparation/](http://www.ebire.org/speechseparation/)
  - as well as the resulting book...

- Hanse meeting:
  - [www.lifesci.sussex.ac.uk/home/Chris_Darwin/Hanse/](http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Hanse/)

- DeLiang Wang’s ICASSP’04 tutorial
  - [www.cse.ohio-state.edu/~dwang/presentation.html](http://www.cse.ohio-state.edu/~dwang/presentation.html)

- Martin Cooke’s NIPS’02 tutorial
  - [www.dcs.shef.ac.uk/~martin/nips.ppt](http://www.dcs.shef.ac.uk/~martin/nips.ppt)
References 1/2


