Learning and Scene Analysis

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA
dpwe@ee.columbia.edu

1. Scene Analysis systems
2. Disambiguation
3. Learning
I. 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 (computational model)
Human and Machine Scene Analysis

- CASA (Brown’92 et seq.):
  - **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
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: cancelation
  - 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
2. Disambiguation

• **Scene** ⇒ multiple possible explanations
  **Analysis** ⇒ choose most reasonable one

• **Most reasonable** means...
  o consistent with grouping cues (CASA)
  o independent sources (ICA)
  o consistent with experience ... (human)
  o \( \max P(\{S_i\}|X) \propto P(X|\{S_i\}) P(\{S_i\}) \)

  \textit{combination physics source models}

• i.e. some kind of \textbf{constraints} to disambiguate
  o **Learning** as the source of this disambiguation knowledge
3. Learning

- **“Reasonable”** = like what we’ve seen before?
  - i.e. infer source models $P(\{S_i\})$ from observations

- **Ways to learn**
  - “memorize” instances
  - generalize to a subspace
    - linear or parametric

- **Learning and Recognition**
  - Recognition is *classification*: set of possible labels
  - learning properties (distinctions) as best approach
Disambiguating with Knowledge

• Use strength of match to models as reasonableness measure for control
• e.g. MAXVQ (Roweis’03)
  ○ learn dictionary of spectrogram slices
  ○ find the ones that ‘fit’
    - or a combination
  ○ ... then filter out excess energy

Noise-corrupt speech                  Matching templates

Matching templates from Sam Roweis’s Montreal 2003 presentation
Recognition for Separation

- Speech recognizers embody knowledge
  - trained on 100s of hours of speech
  - use them as a ‘reasonableness’ measure
- e.g. Seltzer, Raj, Reyes:
  - speech recognizer’s best-match provides optimization target
Learning Elsewhere

• **Control**: learn what is “reasonable”
• **Input**: discriminant features
  - learned subspaces
• **Engine**: clustering parameters
• **Output**: restoration...

Speech Separation: Learning - Dan Ellis
Obliteration and Outputs

• **Perfect separation is rarely possible**
  - e.g. no cancelation after psychoacoustic coding
  - strong interference will *obliterate* part of target

• **What should the output be?**
  - can *fill-in* missing-data holes using source models
    - ‘pretend’ we observed the full signal
    - but: *hides* observed/inferred distinction
  - output internal *model state* instead?
    - e.g. ASR output
    - depends on eventual use...
Conclusions

• Framework for scene analysis
  ○ Input, Output, Engine, Control

• Scene analysis as Disambiguation
  ○ finding the additional constraints

• Learning to spot a reasonable solution

• Various implementations
  ○ direct dictionary fit
  ○ compare output to recognizer’s state

• Learned states as the output?