Computational Auditory Scene Analysis

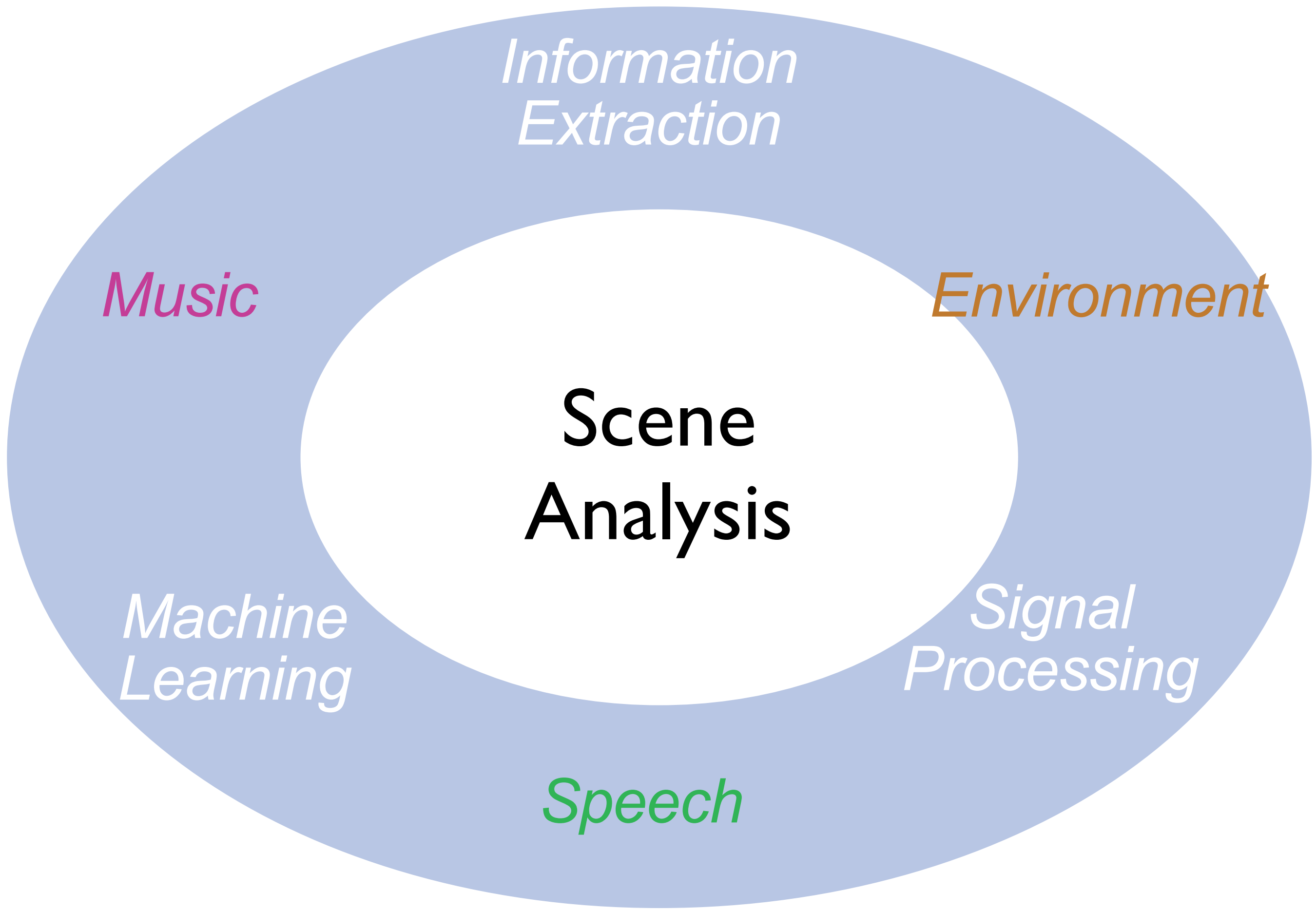
Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

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

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

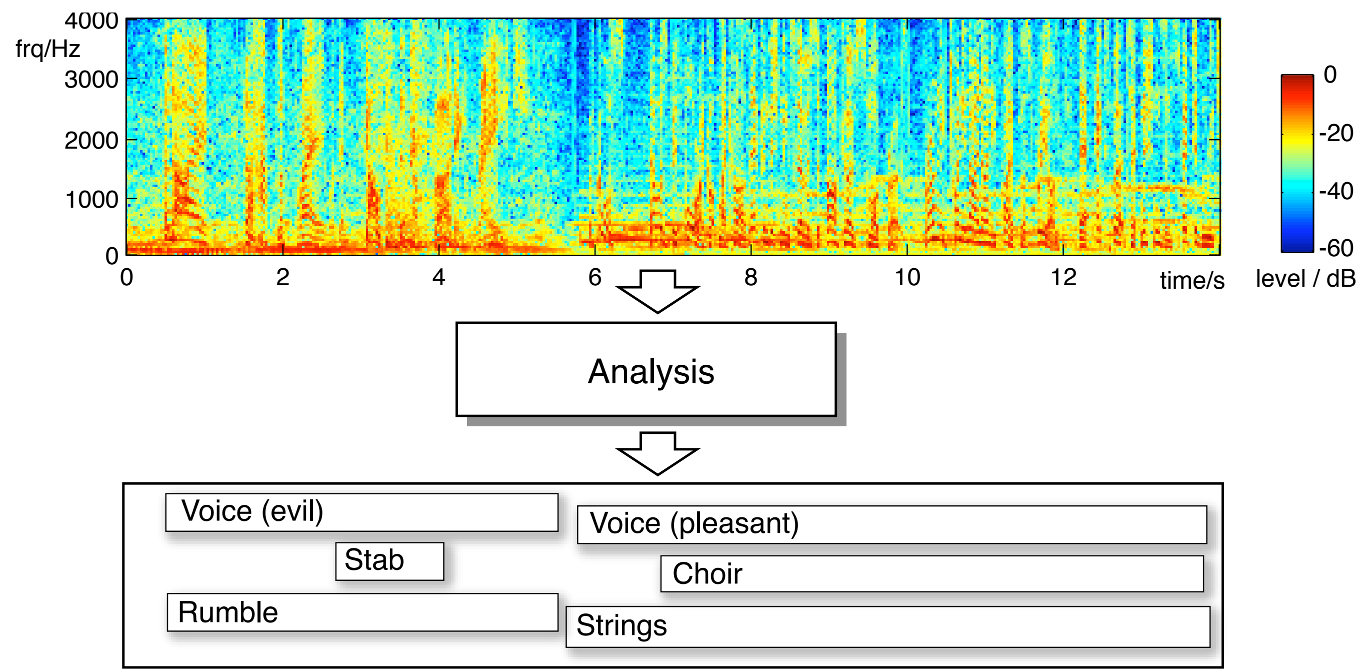
Columbia University
in the city of New York

LabROSA Overview



1. Scene Analysis

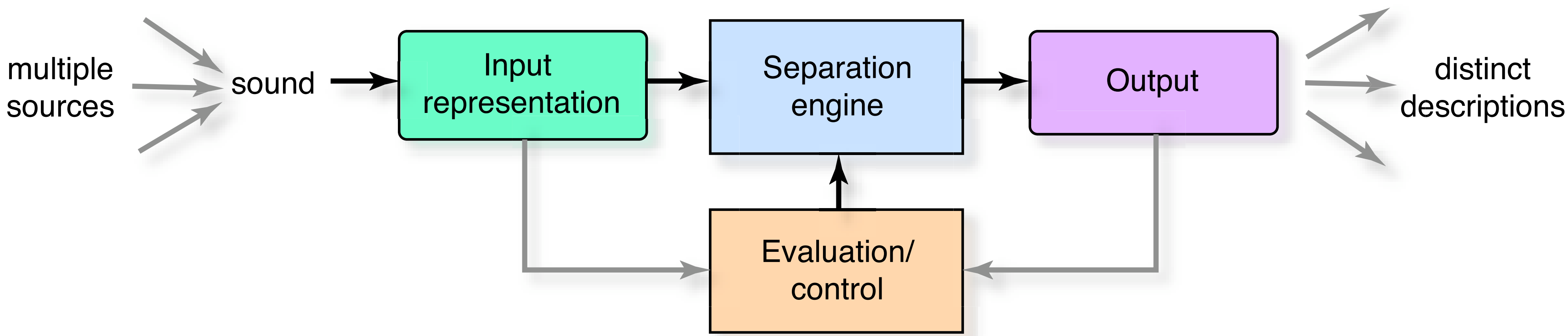
- Recover individual sources from scenes
  - .. duplicate the perceptual effect



- Problems competing sources, channel effects
- Dimensionality loss
  - need additional constraints

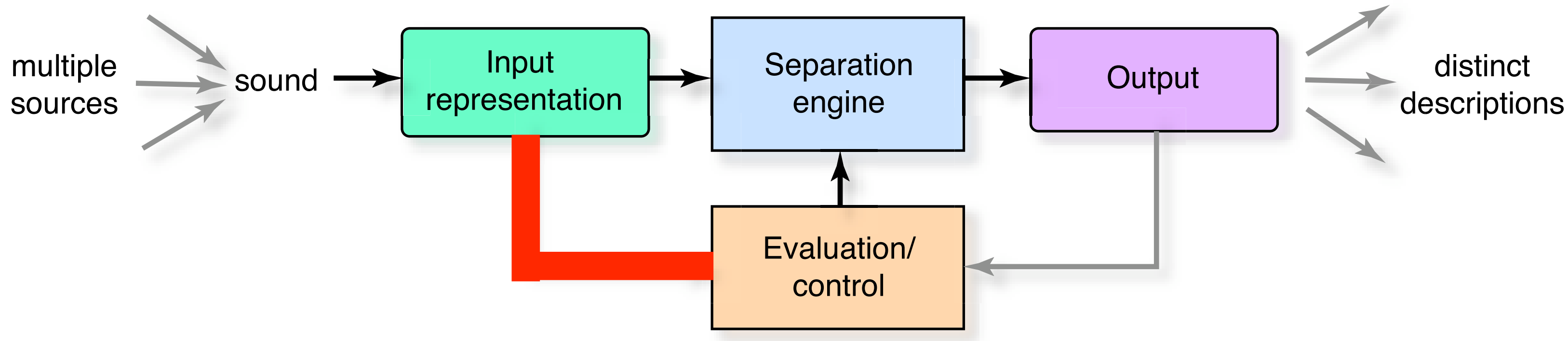
Scene Analysis Systems

- "Scene Analysis"
  - not necessarily separation, recognition, ...
  - scene = overlapping objects, ambiguity
- General Framework:



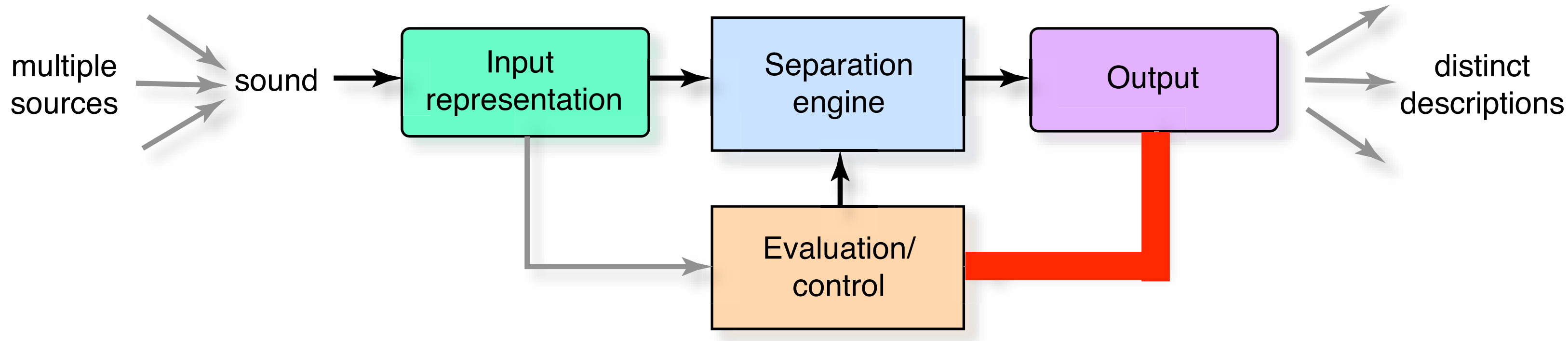
  - distinguish input and output representations
  - distinguish engine (algorithm) and control (constraints, "computational model")

Human and Machine Scene Analysis



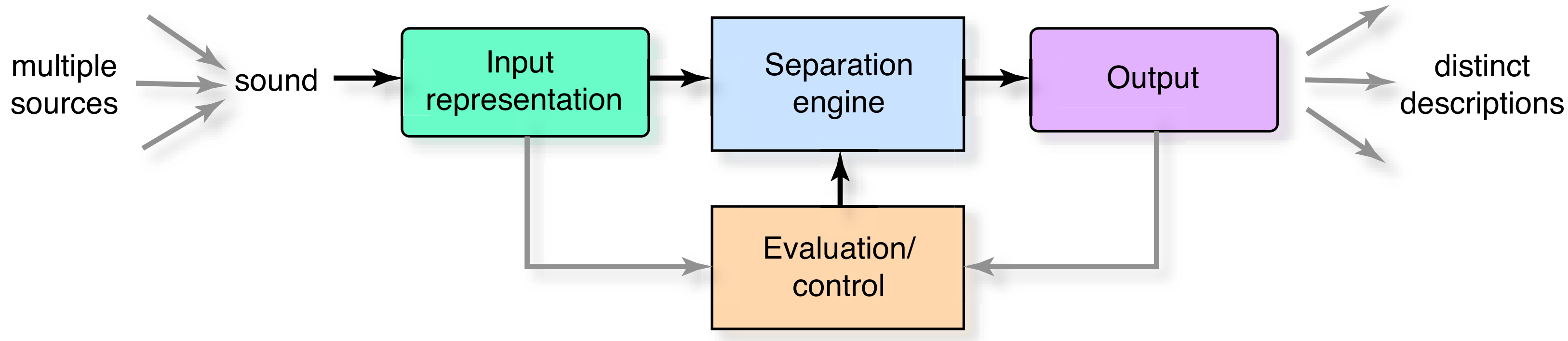
- CASA (Brown'92 et seq.):
    - Input: Periodicity, continuity, onset "maps"
    - Output: Waveform (or mask)
    - Engine: Time-frequency masking
    - Control: "Grouping cues" from input
        - or: spatial features (Roman, ...)

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

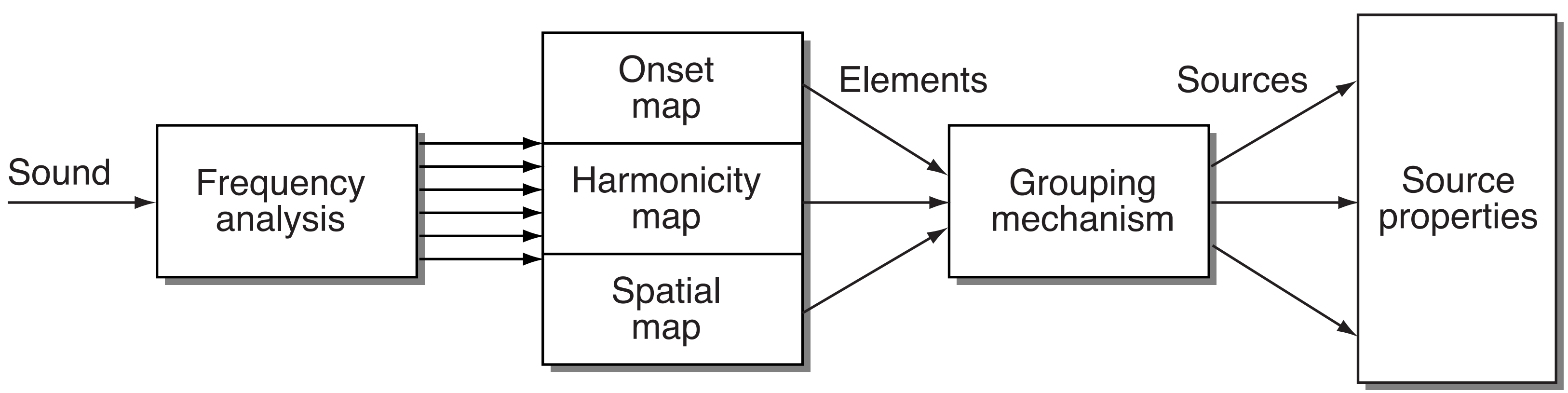


- CASA (e.g. Brown'92):
- ICA (Bell & Sejnowski et seq.):
- **Human Listeners:**
  - Input: excitation patterns ...
  - Output: percepts ...
  - Engine: ?
  - Control: find a plausible explanation

2. Auditory Scene Analysis

(Bregman 1990)

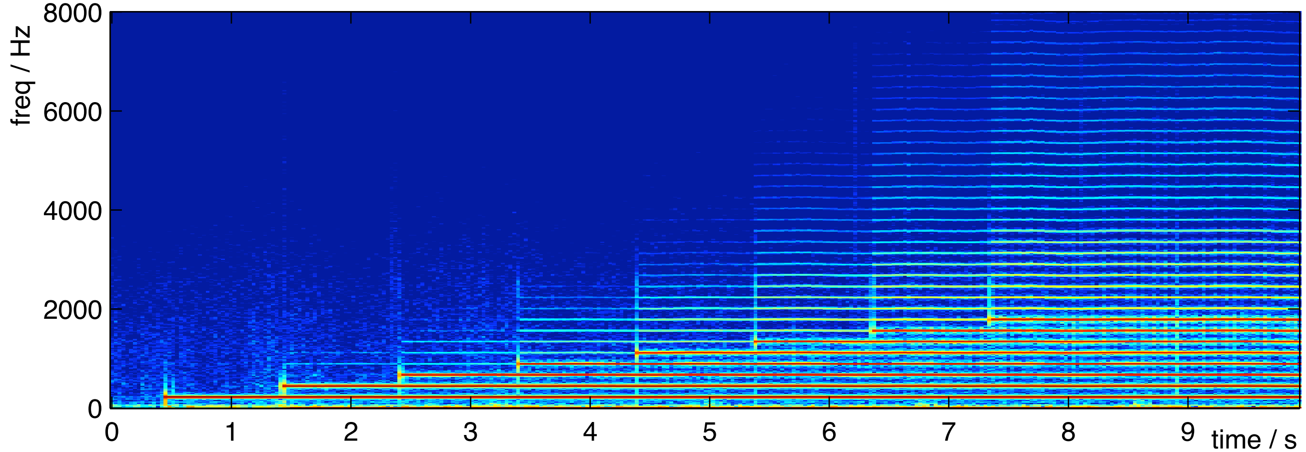
- How do people analyze sound mixtures?
  - break mixture into small elements (in time-freq)
  - elements are grouped in to sources using cues
  - sources have aggregate attributes
- Grouping rules (Darwin, Carlyon, ...):
  - cues: common onset/offset/modulation, harmonicity, spatial location, ...



(after Darwin 1996)

Grouping cues

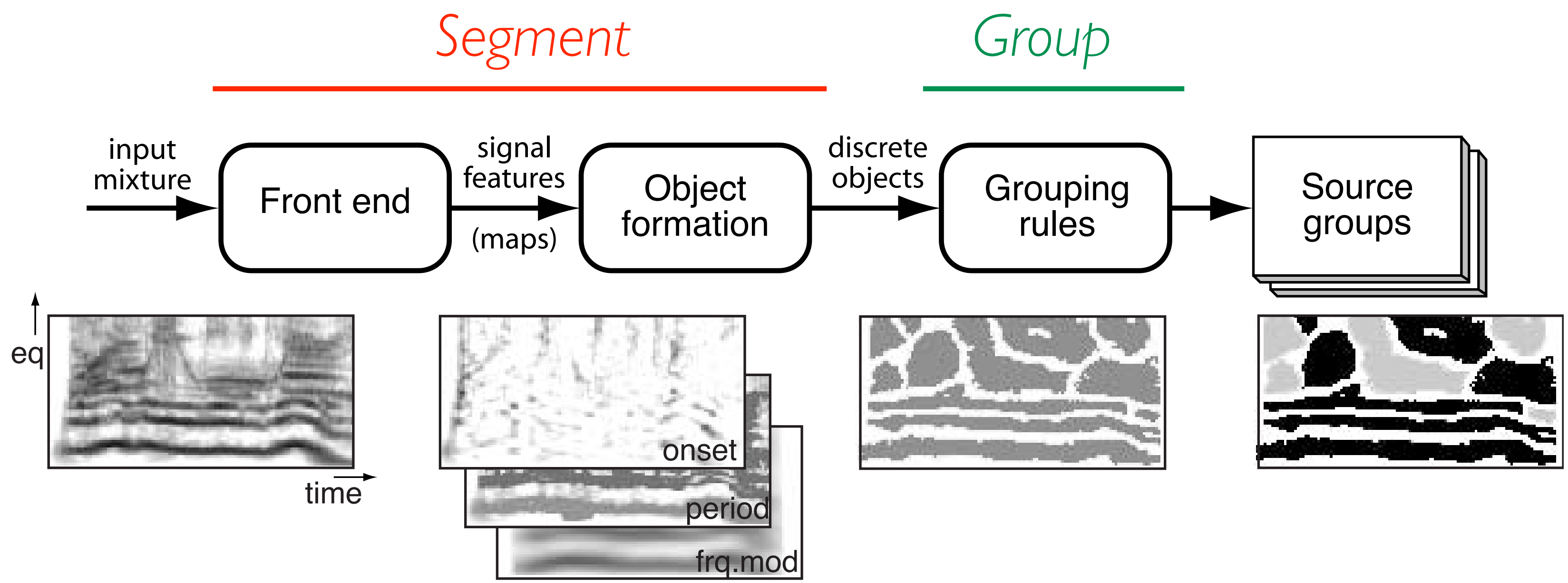
- Main cues: Harmonicity + Onset



(from Pierce 1980)

  - not necessarily consistent!
- Other cues:
  - spatial information
  - 'schema' – learned patterns
- Cues ≈ constraints

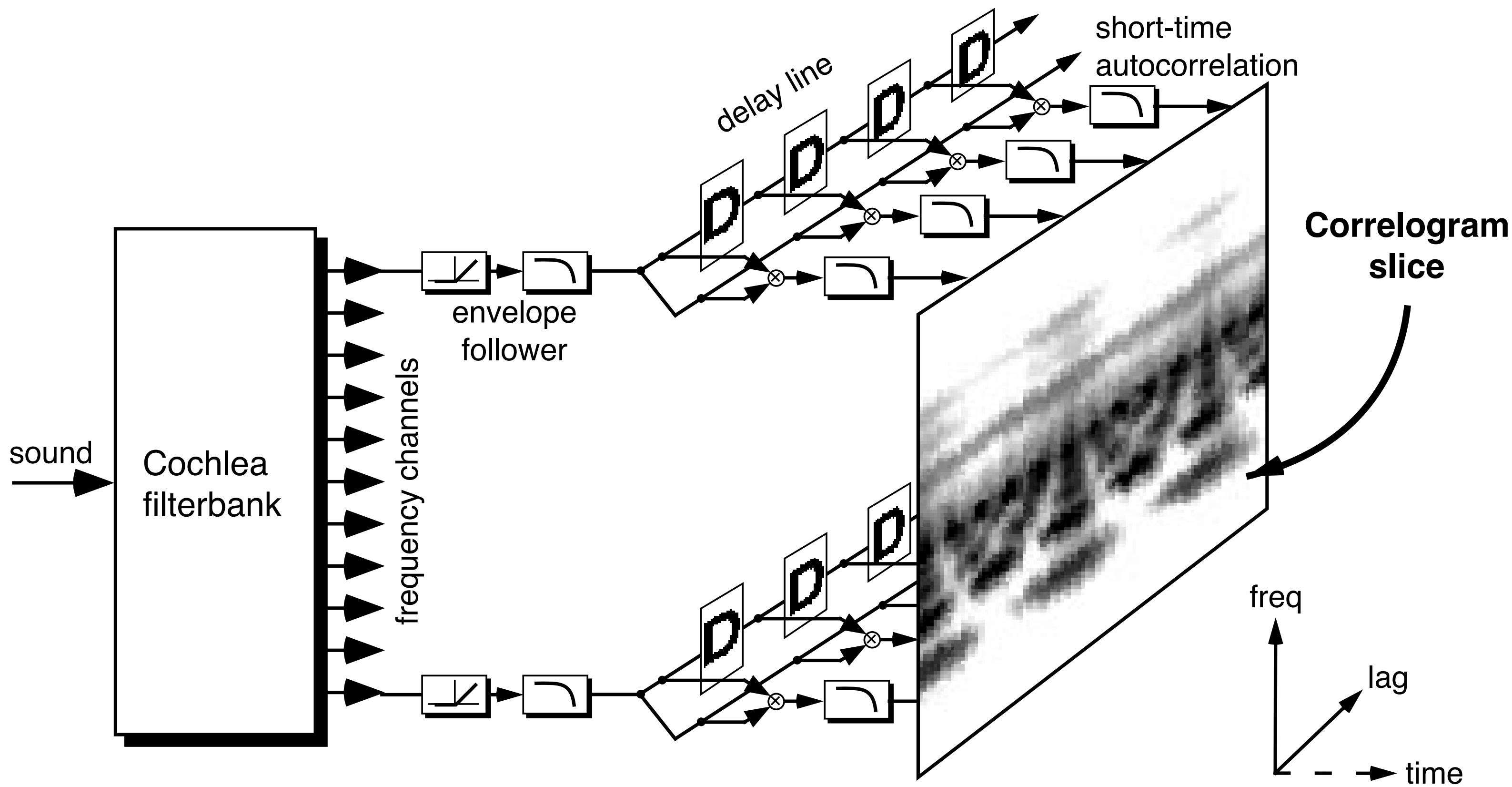
Bottom-up CASA
(Brown'92, Hu & Wang'02)



- **Literal implementation of psychoacoustics**
  - segment time-frequency into elements
  - group into sources
- **Output via time-frequency masking**
  - i.e. time-varying filter

Correlogram front-end
(Slaney'90 et seq.)

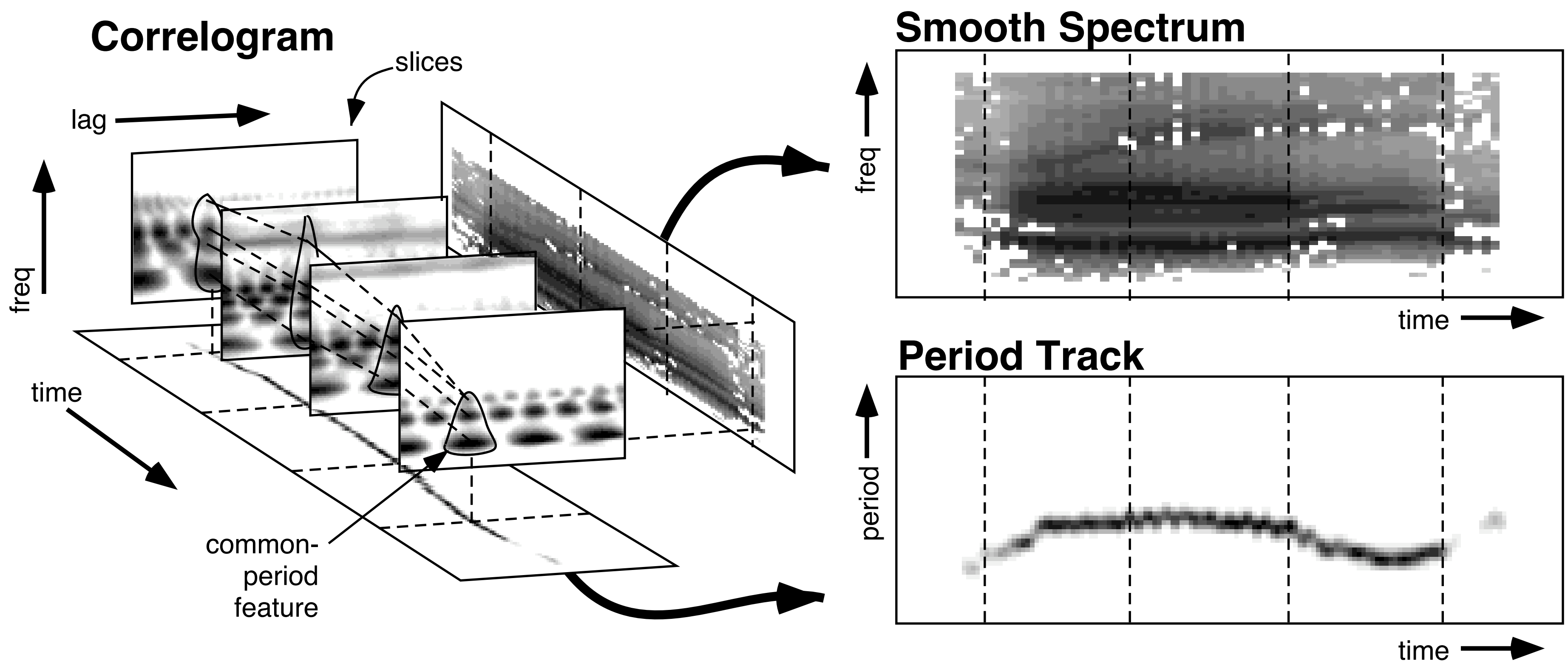
- Periodic modulation as 3rd separating axis
  - envelope to handle unresolved harmonics



"Weft" Periodic Elements

(Ellis'96)

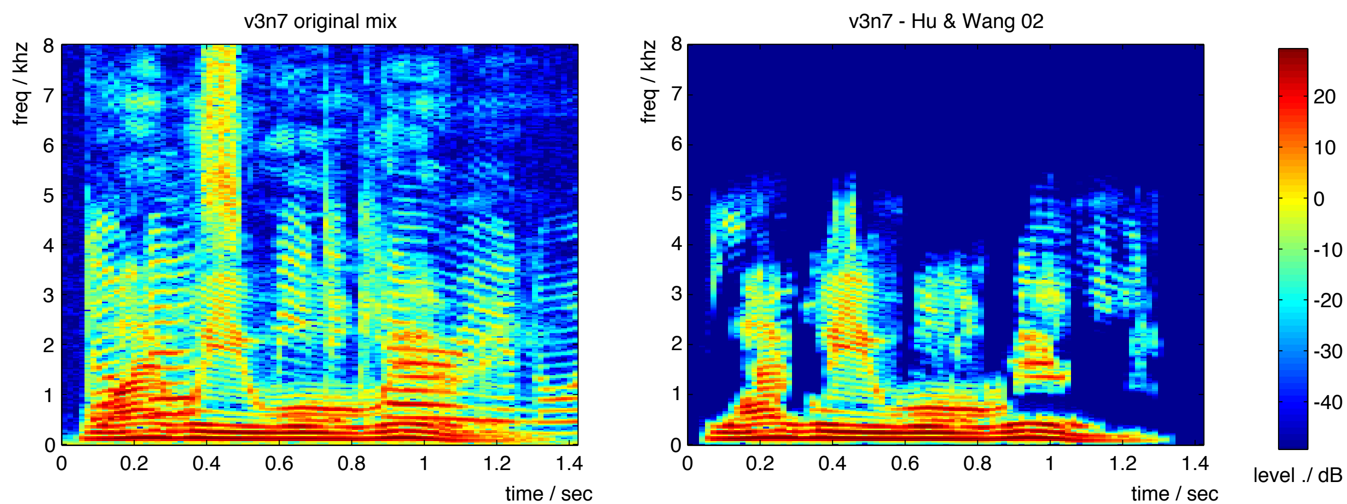
- Represent harmonics without grouping?



  - hard to separate multiple pitch tracks

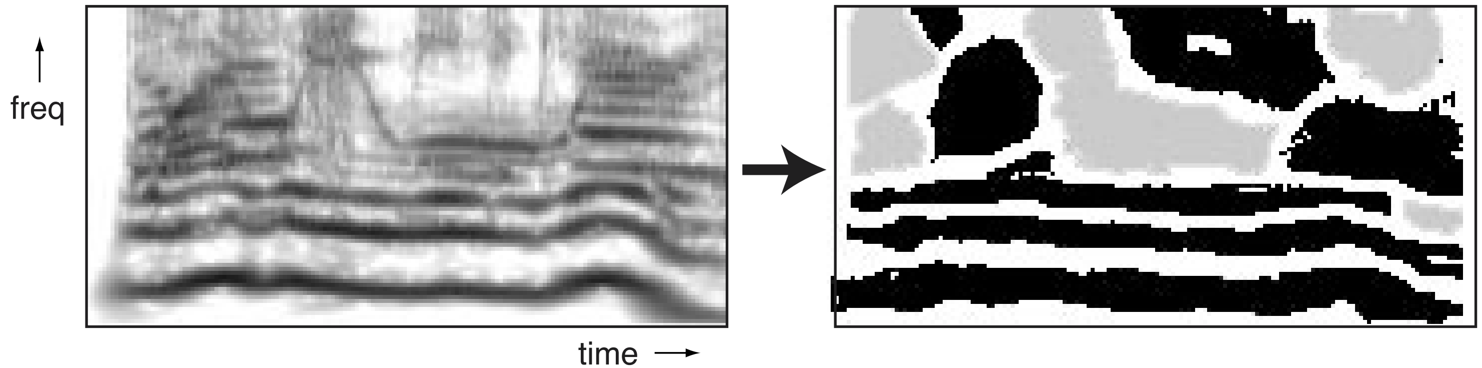
CASA Output

- Time-Frequency masked reconstruction



  - works surprisingly well (for speech?)
  - cannot undo overlapping energy (< 20%?)
  - applicable to reverberation also?
- Or: parametric **resynthesis**
  - e.g. 'wefts', speech synthesizer

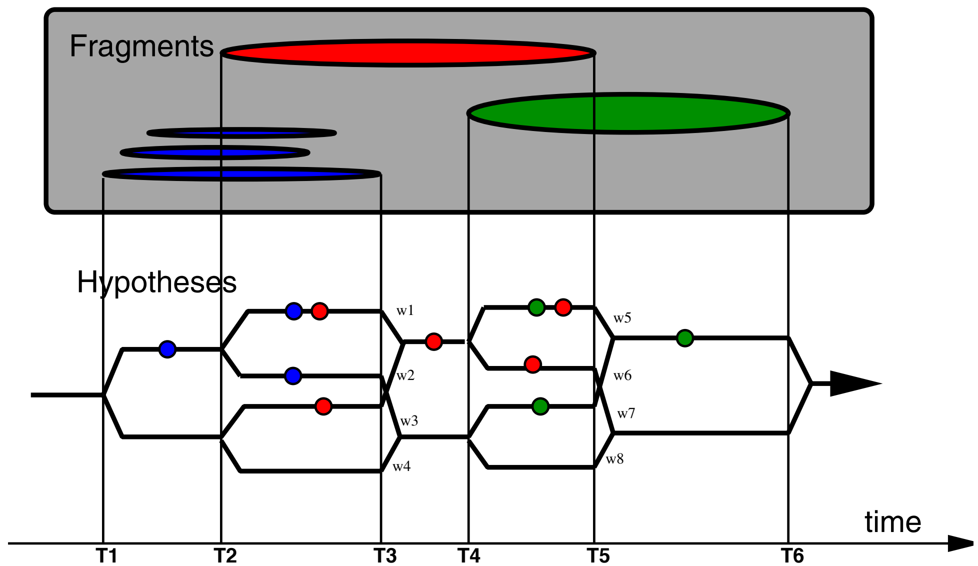
Challenges for CASA



- **Circumscribing time-frequency elements**
  - need to have 'regions', but hard to find
- **Periodicity is the primary cue**
  - how to handle aperiodic energy?
- **Bottom-up leaves no ambiguity or context**
  - how to model illusions/interpolations?
- **Need to group over longer timespans**
  - local properties not enough

Model-based integration

- How to represent high-level constraints?
  How to integrate disparate fragments?
- "Speech fragment decoder" (Barker et al. '05)



  - model of source (e.g. speech recognition HMM) to say which parts go together

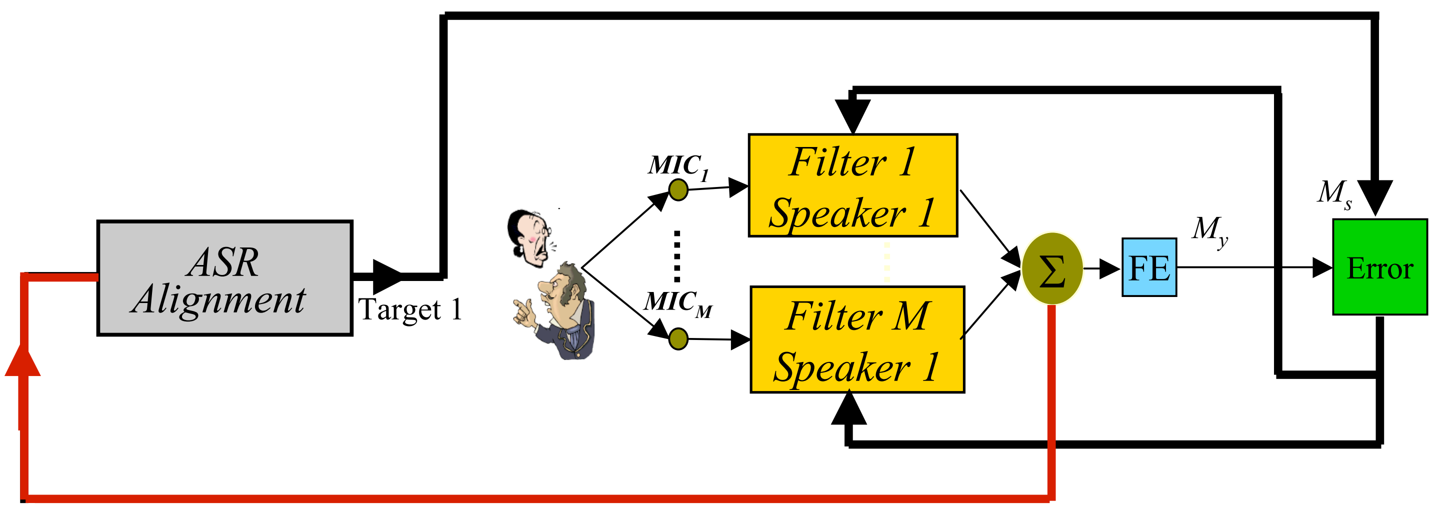
Disambiguation

- Scene ⇒ multiple possible explanations
  Analysis ⇒ choose most reasonable one
- Most reasonable means...
  - consistent with grouping cues (CASA)
  - independent sources (ICA)
  - consistent with experience ... (human)
  - $\max P(\{S_i\} | X) \propto P(X | \{S_i\})\, P(\{S_i\})$

    combination physics source models
- i.e. some kind of constraints to disambiguate
  - Learning as the source of this disambiguation knowledge

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:**



from Manuel Reyes's WASPAA 2003 presentation

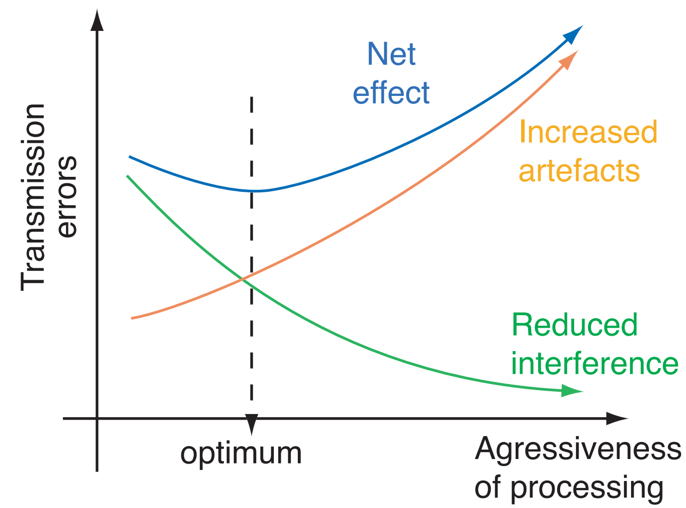
  - speech recognizer's best-match provides optimization target

Obliteration and Outputs

- **Perfect separation is rarely possible**
  - e.g. cancellation 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
    - synthesize with "minimally informative noise"

Practical CASA?

- **When will CASA be useful?**
  - no agreed way to measure progress!
  - intelligibility is a novel idea
- **Obstacles:**
  - graceful degradation
    - effect of distortions
  - unpitched sounds
  - computation
    - look-ahead
  - integration with multichannel techniques
    - sequential or all-at-once?



Current CASA work

- Handling unvoiced events (OSU)
- Partial recognition & grouping (Sheffield)
- Model-based separation (Columbia)
- Spatial cues?
- Dereverberation?

Conclusions

- Framework for scene analysis
  - Input, Output, Engine, Control
- Auditory Scene Analysis
  - in humans and machines
- Scene analysis as Disambiguation
  - finding the additional constraints
- Big problems still to overcome