Computational Auditory Scene Analysis: Principles, Practice and Applications

Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

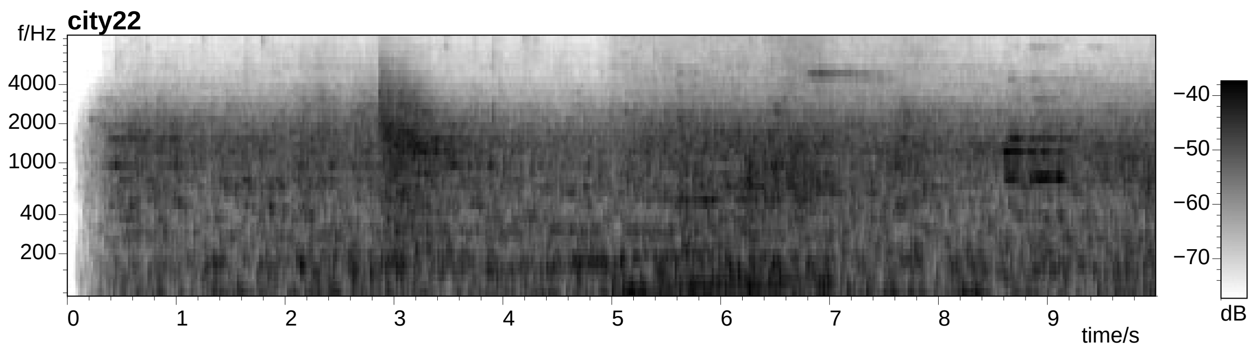
Outline

1. **Auditory Scene Analysis (ASA)**
2. **Computational ASA (CASA)**
3. **Context, expectation & predictions**
4. **Applications: speech recognition, indexing**
5. **Conclusions and open issues**

1 Auditory Scene Analysis

“The organization of sound scenes according to their inferred sources”

- **Sounds rarely occur in isolation**
  - need to ‘separate’ for useful information
- **Human audition is very effective**
  - it’s a shock that modeling it is so hard



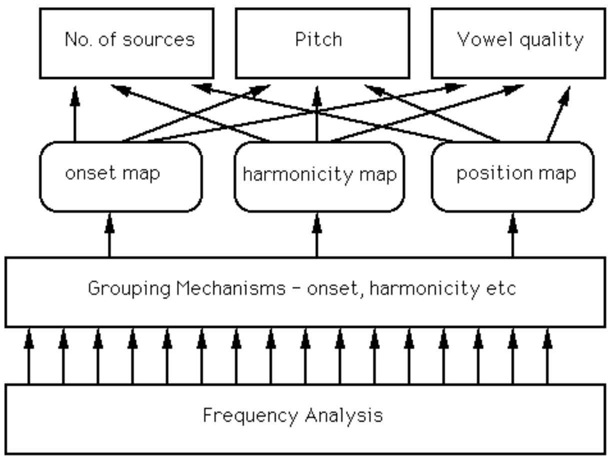
How can we separate sources?

- **In general, we can't**
  - mathematically, too few degrees of freedom
  - sensor noise limits separation
- **'Correct' analysis is subjectively defined**
  - sources exhibit independence, continuity, ...

  $\rightarrow$ *ecological* constraints enable organization
- **Goal not complete separation (reconstruction) but organization of available information into useful structures**
  - telling us useful things about the outside world
- **'The auditory system as a separation machine' (Alain de Cheveigné)**

Psychology of ASA

- **Extensive experimental research**
  - perception of simplified stimuli (sinusoids, noise)
- **“Auditory Scene Analysis” [Bregman 1990]**
  - first: break mixture into small *elements*
  - elements are *grouped* in to sources using *cues*
  - sources have aggregate attributes
- **Grouping ‘rules’ (Darwin, Carlyon, ...):**
  - common onset/offset/modulation,harmonicity, spatial location, ...



(from
Darwin 1996)

Thinking about information processing: Marr's levels-of-explanation

- **Three distinct aspects to info. processing**

Computational Theory	'what' and 'why'; the overall goal	Sound source organization
Algorithm	'how'; an approach to meeting the goal	Auditory grouping
Implementation	practical realization of the process.	Feature calculation & binding

Why bother?

- helps organize interpretation
- it's OK to consider levels separately, one at a time

Cues to grouping

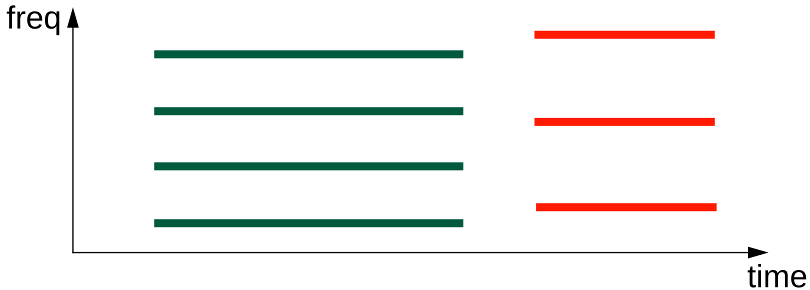
- **Common onset/offset/modulation (“fate”)**
- **Common periodicity (“pitch”)**

	Common onset	**Periodicity**
Computational theory	Acoustic consequences tend to be synchronized	(Nonlinear) cyclic processes are common
Algorithm	Group elements that start in a time range	? Place patterns ? Autocorrelation
Implementation	Onset detector cells Synchronized osc’s?	? Delay-and-mult ? Modulation spect

- **Spatial location (ITD, ILD, spectral cues)**
- **Sequential cues...**
- **Source-specific cues...**

Simple grouping

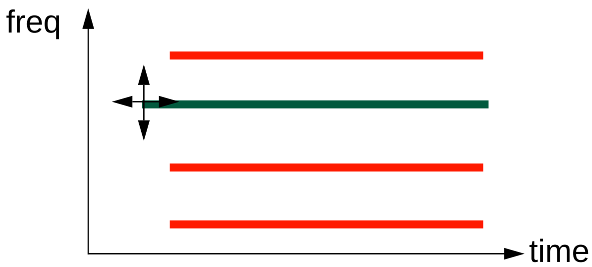
- **E.g. isolated tones**



Computational theory	• common onset • common period (harmonicity)
Algorithm	• locate elements (tracks) • group by shared features
Implementation	? exhaustive search • evolution in time

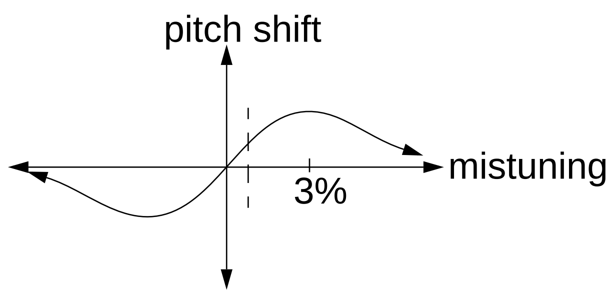
Complications for grouping:
1: Cues in conflict

- **Mistuned harmonic (Moore, Darwin..):**



  - harmonic usually groups by onset & periodicity
  - can alter frequency and/or onset time
  - 'degree of grouping' from overall pitch match

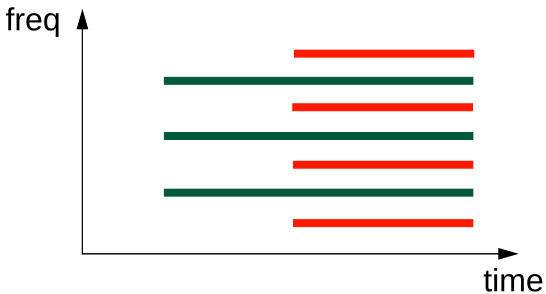
- **Gradual, various results:**



  - heard as separate tone, still affects pitch

Complications for grouping: 2: The effect of time

- **Added harmonics:**



  - onset cue initially segregates; periodicity eventually fuses

- **The effect of time**
  - some cues take time to become apparent
  - onset cue becomes increasingly distant...

- **What is the impetus for fission?**
  - e.g. double vowels
  - depends on what you expect .. ?

Outline

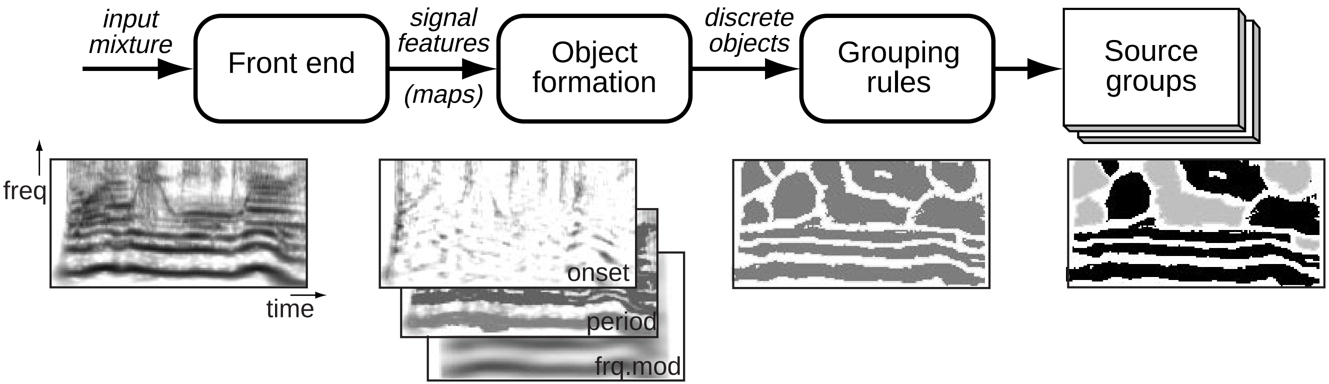
1. **Auditory Scene Analysis (ASA)**
2. **Computational ASA (CASA)**
   - A simple model of grouping
   - Other systems
3. **Context, expectation & predictions**
4. **Applications: speech recognition, indexing**
5. **Conclusions and open issues**

2 Computational Auditory Scene Analysis (CASA)

- **Automatic sound organization?**
  - convert an undifferentiated signal into a description in terms of different sources
- **Translate psych. rules into programs?**
  - representations to reveal common onset, harmonicity ...
- **Motivations & Applications**
  - it's a puzzle: new processing principles?
  - real-world interactive systems (speech, robots)
  - hearing prostheses (enhancement, description)
  - advanced processing (remixing)
  - multimedia indexing

A simple model of grouping

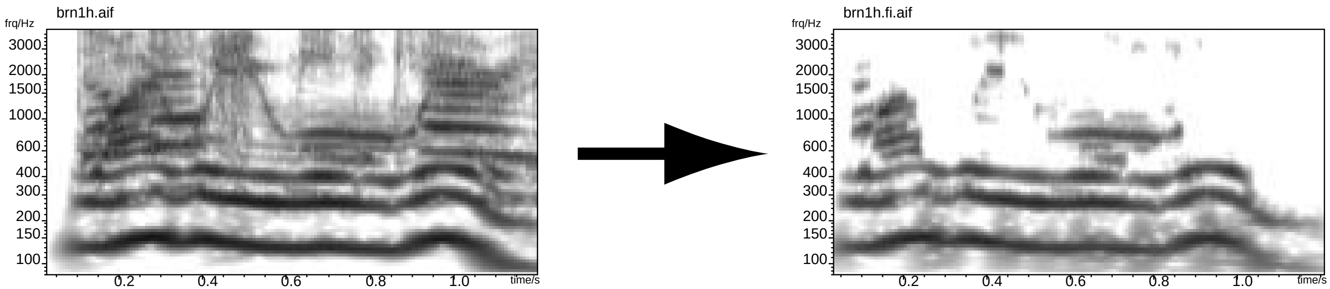
- **"Bregman at face value" (e.g. Brown 1992):**



  - feature maps
  - periodicity cue
  - common-onset boost
  - resynthesis

Grouping model results

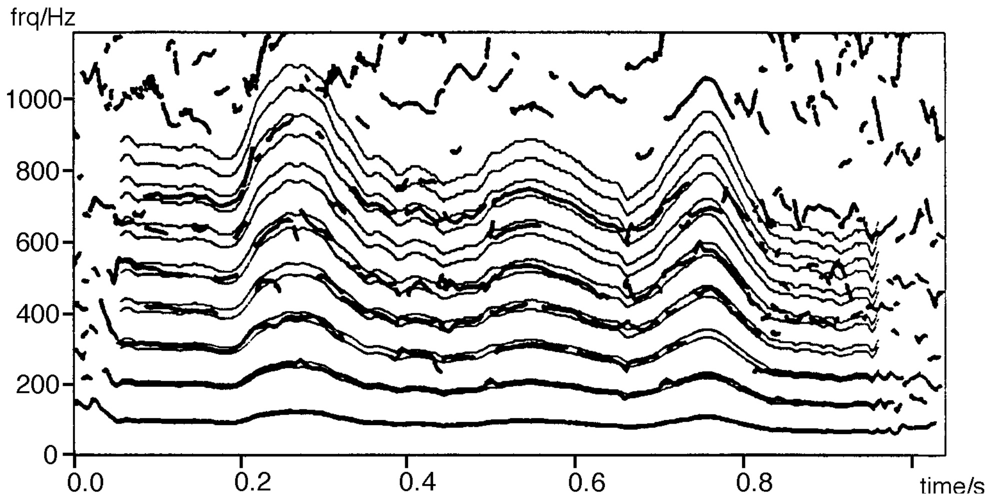
- **Able to extract voiced speech:**



- **Periodicity is the primary cue**
  - how to handle aperiodic energy?
- **Limitations**
  - resynthesis via filter-mask
  - *only* periodic targets
  - robustness of discrete objects

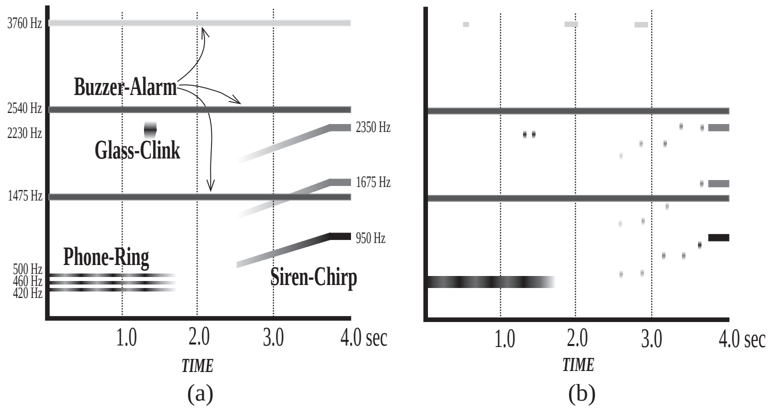
Other CASA systems (1)

- **Weintraub 1985**
  - separate male & female voices
  - find periodicities in each frequency channel by auto-coincidence
  - number of voices is ‘hidden state’
- **Cooke 1991**
  - ‘Synchrony strands’ auditory model
  - Fusing resolved harmonics and AM formants
  - led to [Brown 1992]



Other CASA systems (2)

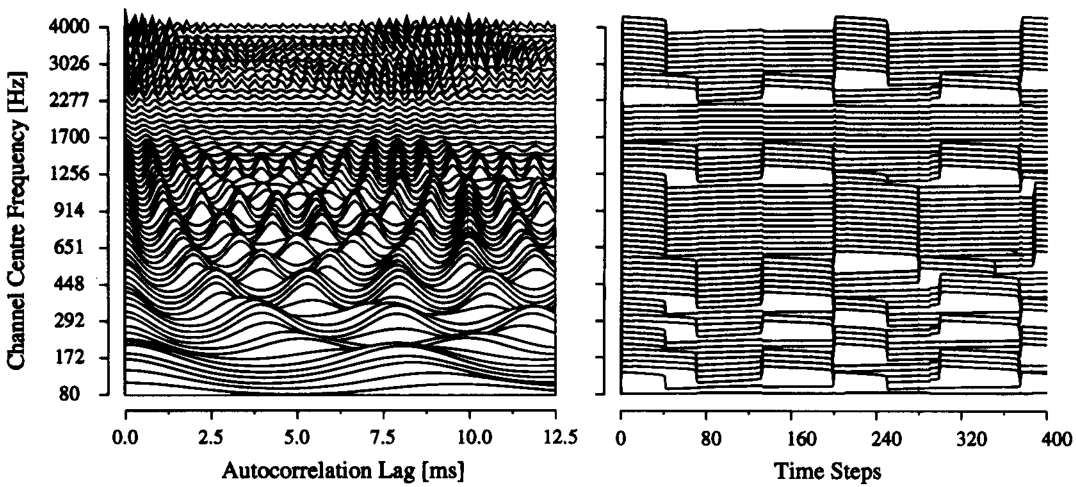
- **Okuno, Nakatani &c (1994-)**
  - ‘tracers’ follow each harmonic + noise ‘agent’
  - residue-driven: account for whole signal
- **Klassner 1996**
  - search for a combination of templates
  - high-level hypotheses permit front-end tuning



- **Ellis 1996**
  - model for events perceived in dense scenes
  - prediction-driven: observations - hypotheses

Other signal-separation approaches

- **HMM decomposition (RK Moore '86)**
  - recover combined source states directly
- **Blind source separation (Bell & Sejnowski '94)**
  - find exact separation parameters by maximizing statistic e.g. signal independence
- **Neural models (Malsburg, Wang & Brown)**
  - avoid implausible AI methods (search, lists)
  - oscillators substitute for iteration?



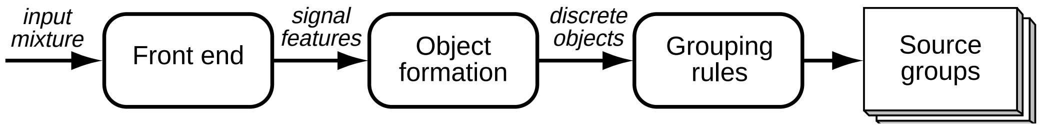
Outline

1. **Auditory Scene Analysis (ASA)**
2. **Computational ASA (CASA)**
3. **Context, expectation & predictions**
   - the effect of context
   - streaming, illusions and restoration
   - prediction-driven (PD) CASA
4. **Applications: speech recognition, indexing**
5. **Conclusions and open issues**

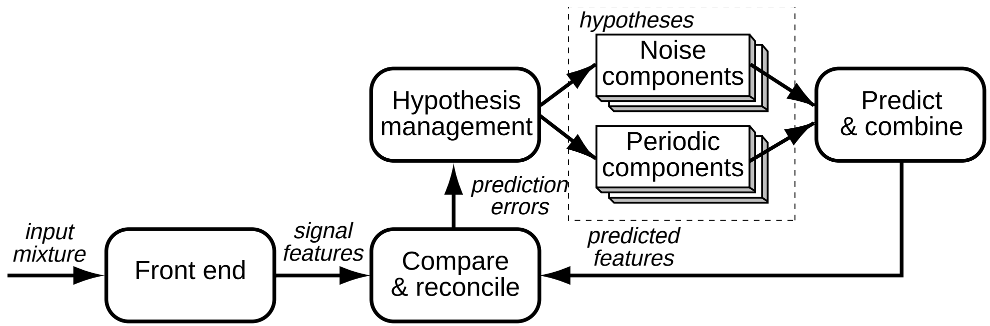
3 Context, expectations & predictions

Perception is not *direct*
but a *search* for *plausible hypotheses*

- **Data-driven...**



vs. Prediction-driven

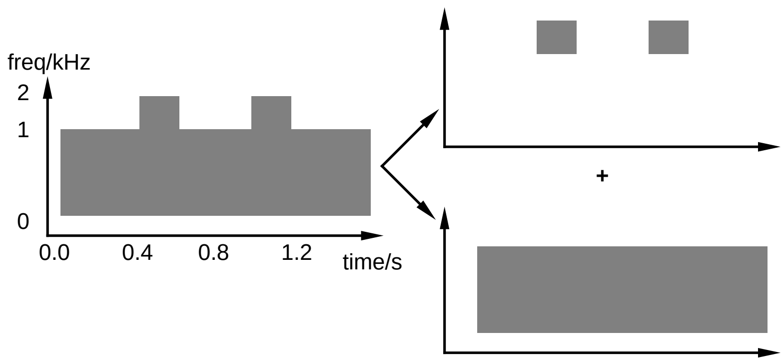


- **Motivations**
  - detect non-tonal events (noise & clicks)
  - support 'restoration illusions'...
    $\rightarrow$ hooks for high-level knowledge
  + 'complete explanation', multiple hypotheses, resynthesis

The effect of context

- **Context can create an 'expectation': i.e. a bias towards a particular interpretation**
- **e.g. Bregman's "old-plus-new" principle:**

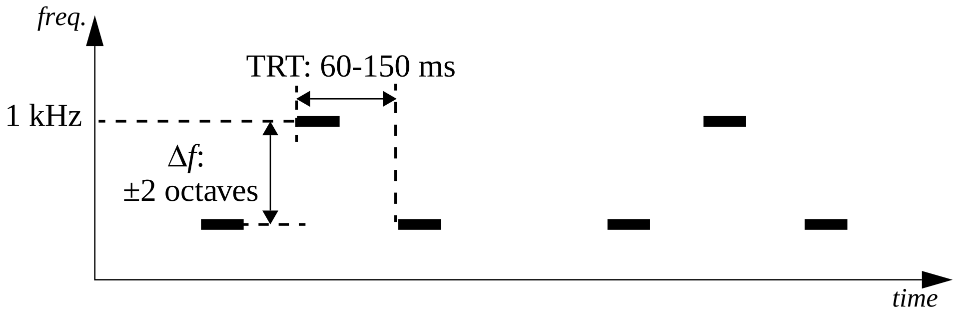
  A change in a signal will be interpreted as an *added* source whenever possible



- a different division of the same energy depending on what preceded it

Streaming

- **Successive tone events form separate streams**



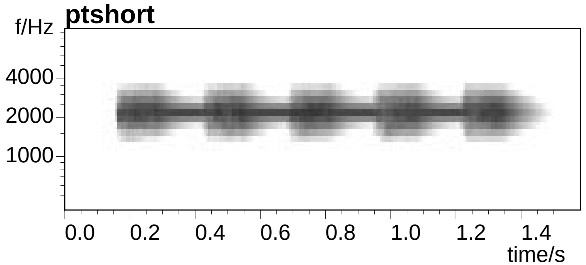
- **Order, rhythm &c *within*, not *between*, streams**

Computational theory	Consistency of properties for successive source events
Algorithm	• ‘expectation window’ for known streams (widens with time)
Implementation	• competing time-frequency affinity weights...

Restoration & illusions

- **Direct evidence may be masked or distorted**
  $\rightarrow$make best guess using available information

- **E.g. the 'continuity illusion':**



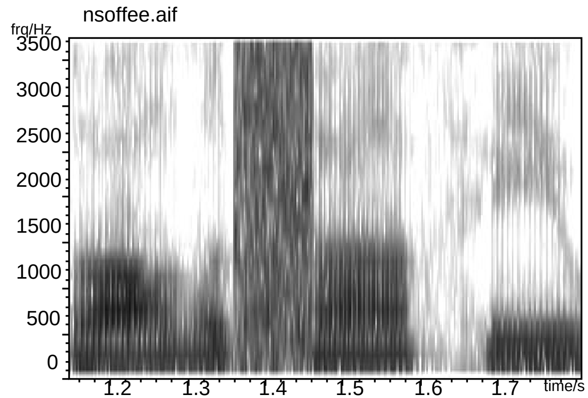
  - tones alternates with noise bursts
  - noise is strong enough to mask tone
    ... so listener discriminate presence
  - continuous tone distinctly perceived
    for gaps ~100s of ms

$\rightarrow$ **Inference acts at low, preconscious level**

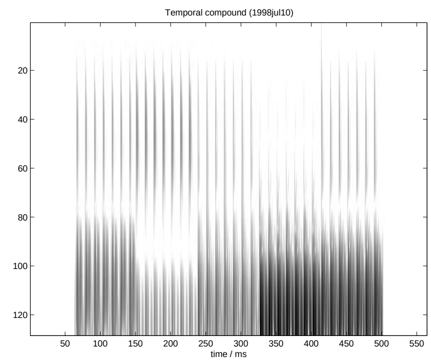
Speech restoration

- **Speech provides very strong bases for inference (coarticulation, grammar, semantics):**

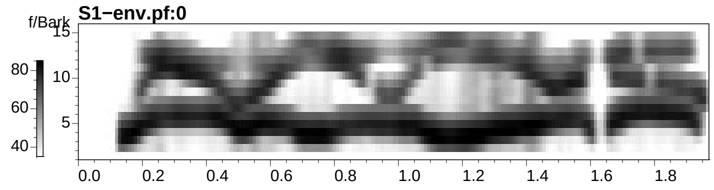
- **Phonemic restoration**



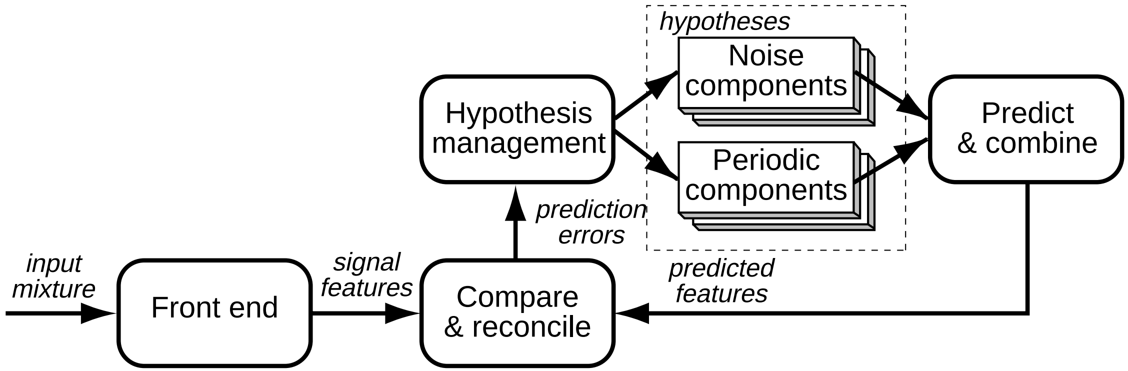
- **Temporal compounds**



- **Sinewave speech (duplex?)**



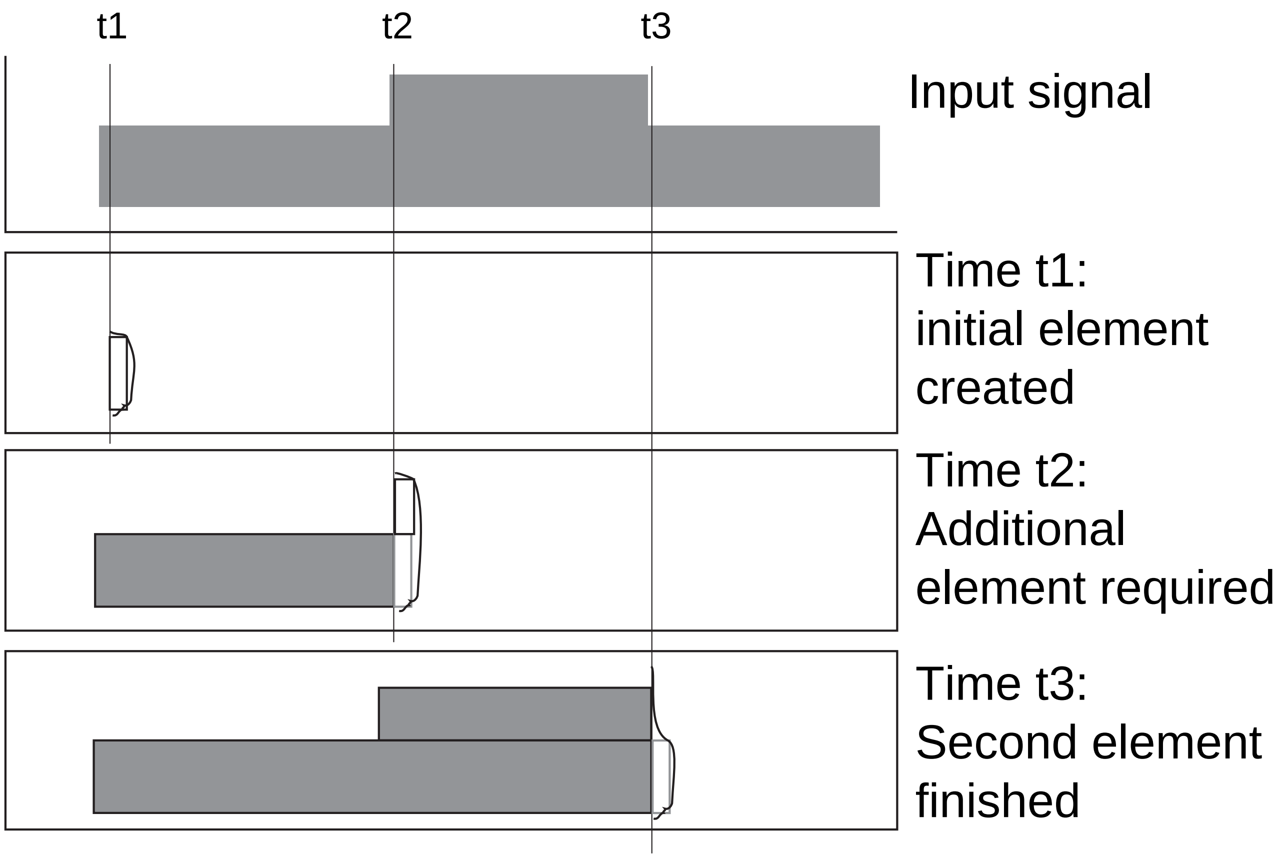
Modeling top-down processing: ‘Prediction-driven’ CASA (PDCASA):



- **An approach as well as an implementation...**
- **Key features:**
  - ‘complete explanation’ of all scene energy
  - vocabulary of periodic/noise/transient elements
  - multiple hypotheses
  - explanation hierarchy

PDCASA for old-plus-new

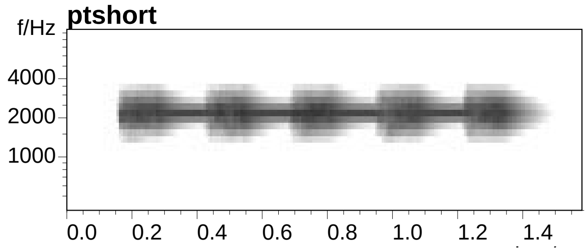
- **Incremental analysis**



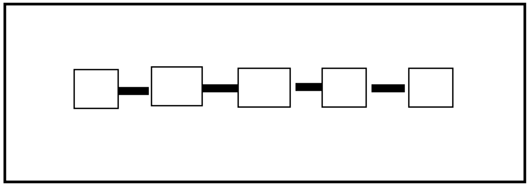
PDCASA for the continuity illusion

- **Subjects hear the tone as continuous**

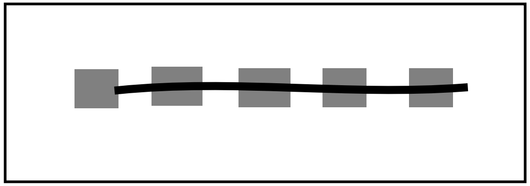
  ... if the noise is a plausible masker



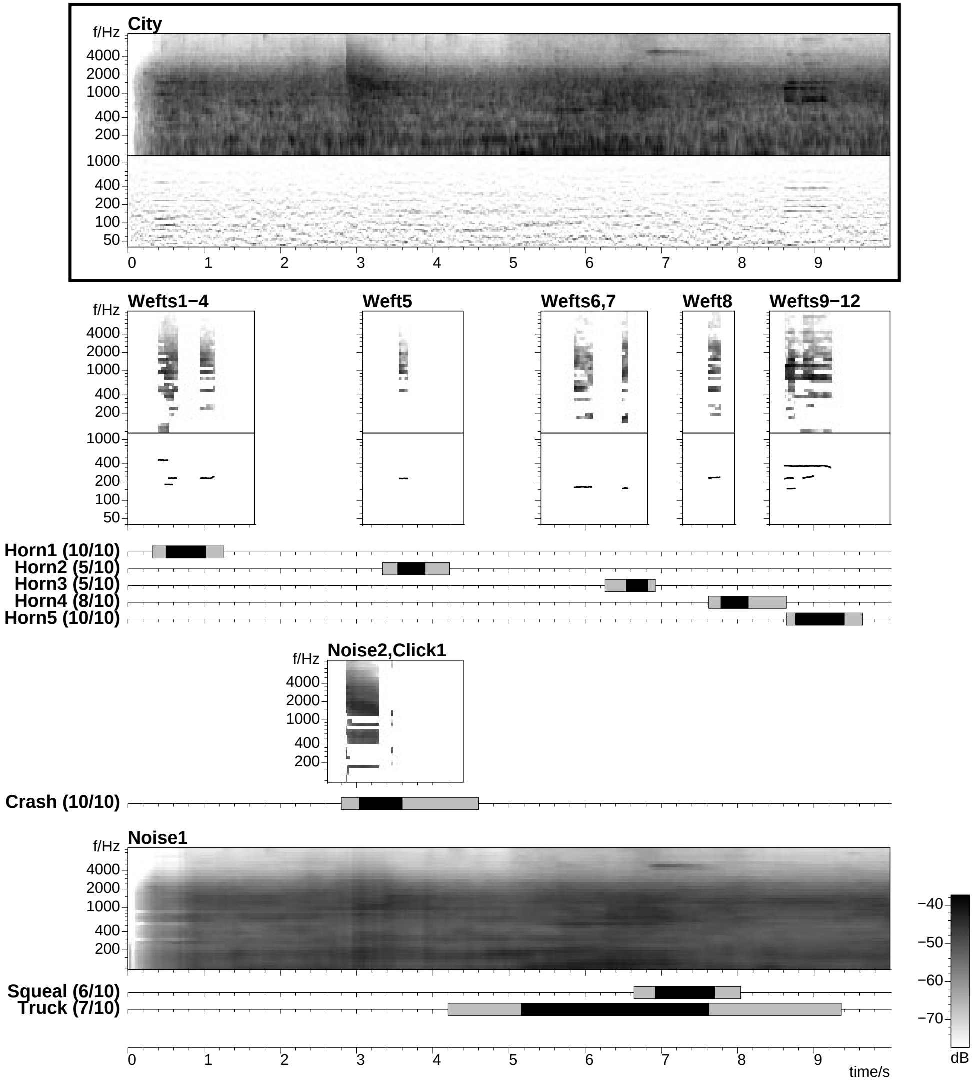
- **Data-driven analysis gives just visible portions:**



- **Prediction-driven can infer masking:**

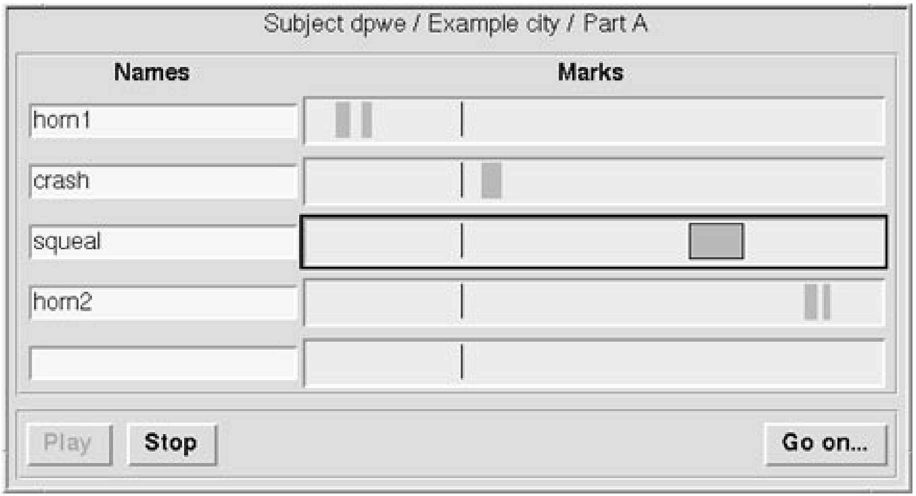


PDCASA analysis of a complex scene



Problems in PDCASA

- **Subjective ground-truth in mixtures?**
  - listening tests collect ‘perceived events’:



- **Other problems**
  - error allocation
  - source hierarchy
  - rating hypotheses
  - resynthesis

Marrian analysis of PDCASA

- **Marr invoked to separate high-level function from low-level details**

Computational theory	• Objects persist predictably • Observations interact irreversibly
Algorithm	• Build hypotheses from generic elements • Update by prediction-reconciliation
Implementation	???

"It is not enough to be able to describe the response of single cells, nor predict the results of psychophysical experiments.

Nor is it enough even to write computer programs that perform approximately in the desired way:

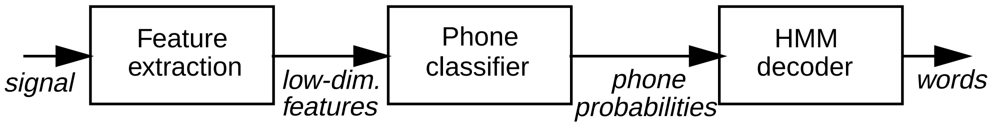
One has to do all these things at once, and also be very aware of the computational theory..."

Outline

1. Auditory Scene Analysis (ASA)
2. Computational ASA (CASA)
3. Context, expectation & predictions
4. **Applications**
   - CASA for speech recognition
   - CASA for audio indexing
5. Conclusions and open issues

4.1 Applications: Speech recognition

- **Conventional speech recognition:**



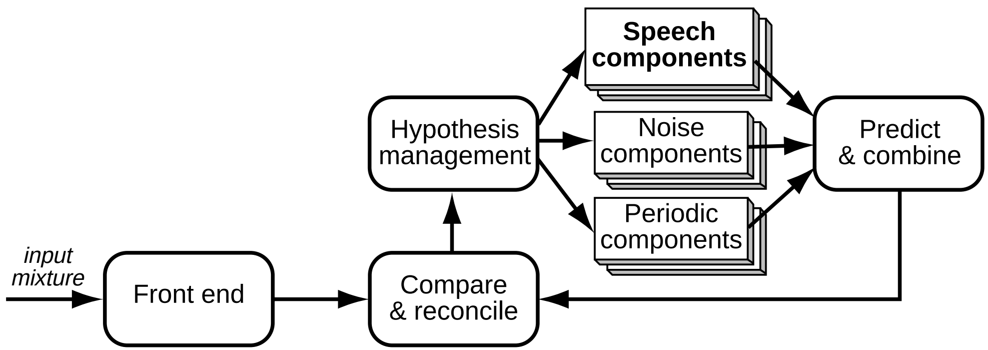
  - signal assumed entirely speech
  - find valid segmentation using discrete labels
  - class models from training data

- **Some problems:**
  - need to ignore lexically-irrelevant variation (microphone, voice pitch etc.)
  - compact feature space $\rightarrow$ everything speech-like

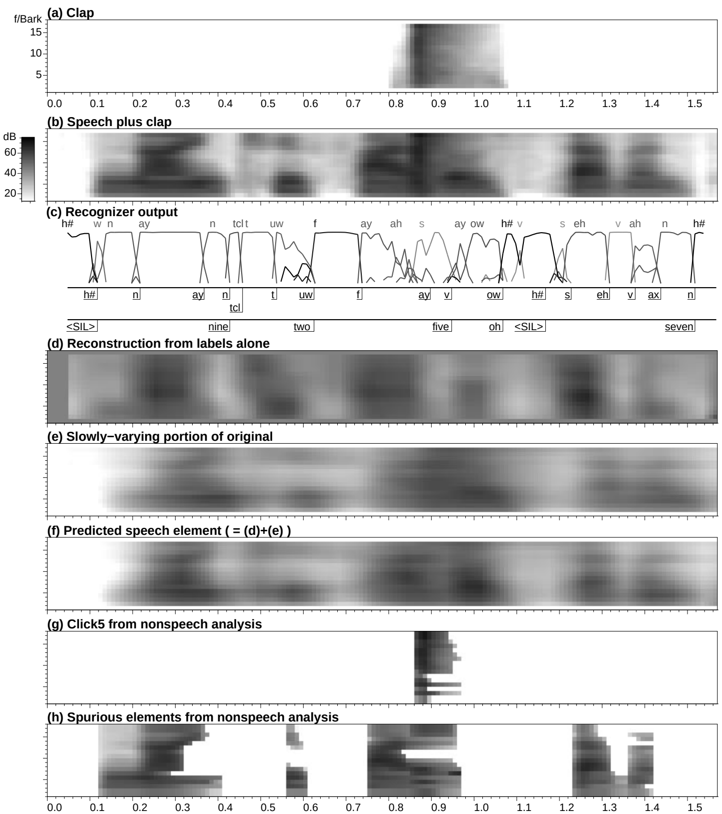
- **Very fragile to nonspeech, background**
  - scene-analysis methods very attractive...

CASA for speech recognition

- **Data-driven: CASA as preprocessor**
  - problems with ‘holes’ (but: Okuno)
  - doesn’t exploit knowledge of speech structure
- **Missing data (Cooke &c, de Cheveigné)**
  - CASA cues distinguish present/absent
  - use ‘aware’ classifier
- **Prediction-driven: speech as component**
  - same ‘reconciliation’ of speech hypotheses
  - need to express ‘predictions’ in signal domain



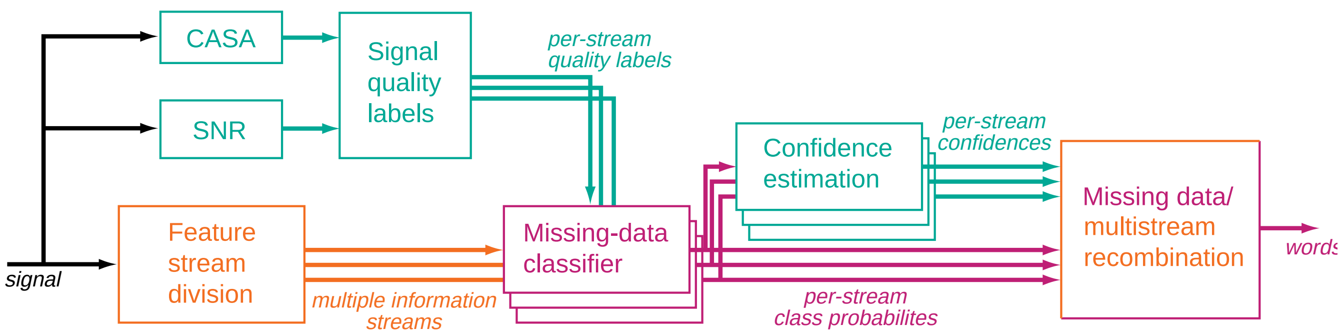
Example of speech & nonspeech



- **Problems:**
  - undoing classification & normalization
  - finding a starting hypothesis
  - granularity of integration

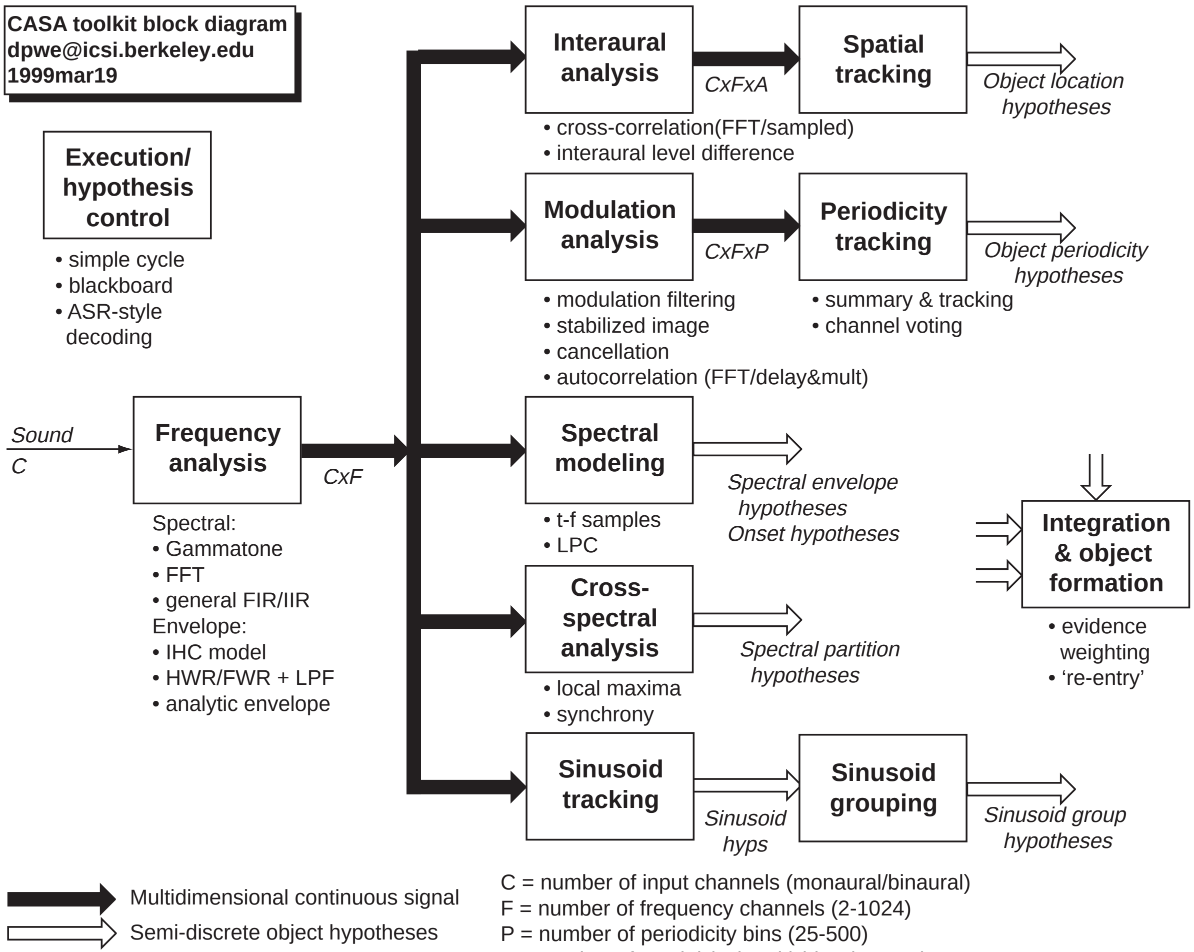
CASA & Missing Data

- **CASA indicates source energy regions**
  - e.g. the time-frequency mask of [Brown 1992]
- **'Missing data' theory permits inference:**
  - skip dimensions of an uncorrelated Gaussian
  - perform full integral over unknown range
  - 'data imputation' e.g. for deltas, cepstra
- **... or just weighting of information streams**
  - 4 band recognizer [Berthommier et al. 1998]
- **RESPITE project**



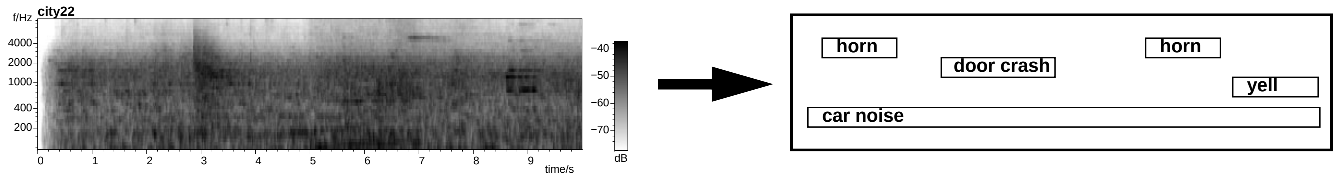
The RESPITE CASA Toolkit

(Barker, Ellis & Cooke 1999)



C = number of input channels (monaural/binaural)
F = number of frequency channels (2-1024)
P = number of periodicity bins (25-500)
A = number of spatial (azimuth) bins (16-256)

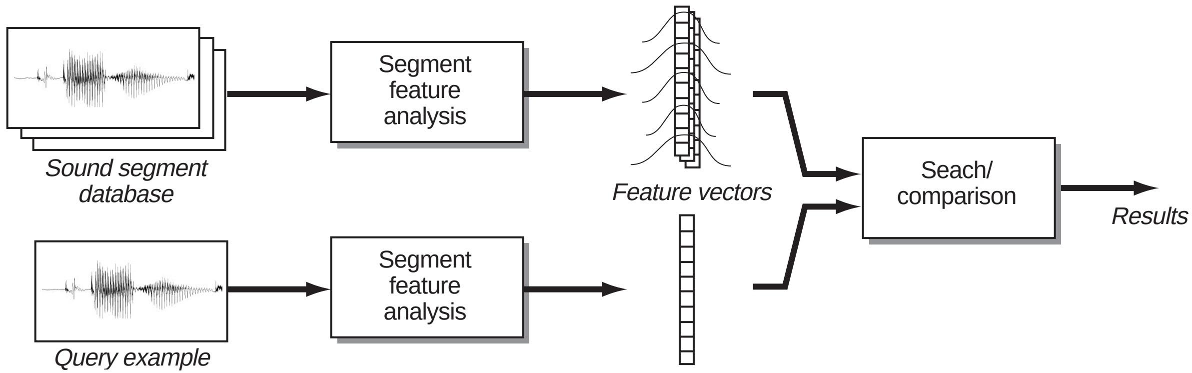
4.2 Applications: audio indexing



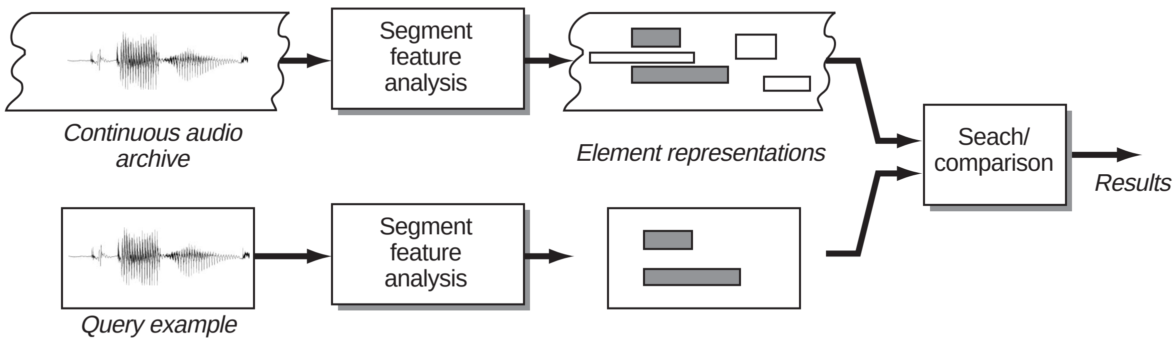
- **Current approaches**
  - speech recognition (Informedia etc.)
  - whole-sample statistics (Muscle Fish)
- **What are the ‘objects’ in a soundtrack?**
  - i.e. the analog of words in text IR
  - subjective definition → need auditory model
- **Problems**
  - parts vs. wholes
  - general vs. specific
  - how to be ‘data-driven’

Element-based audio indexing

- **Segment-level features**



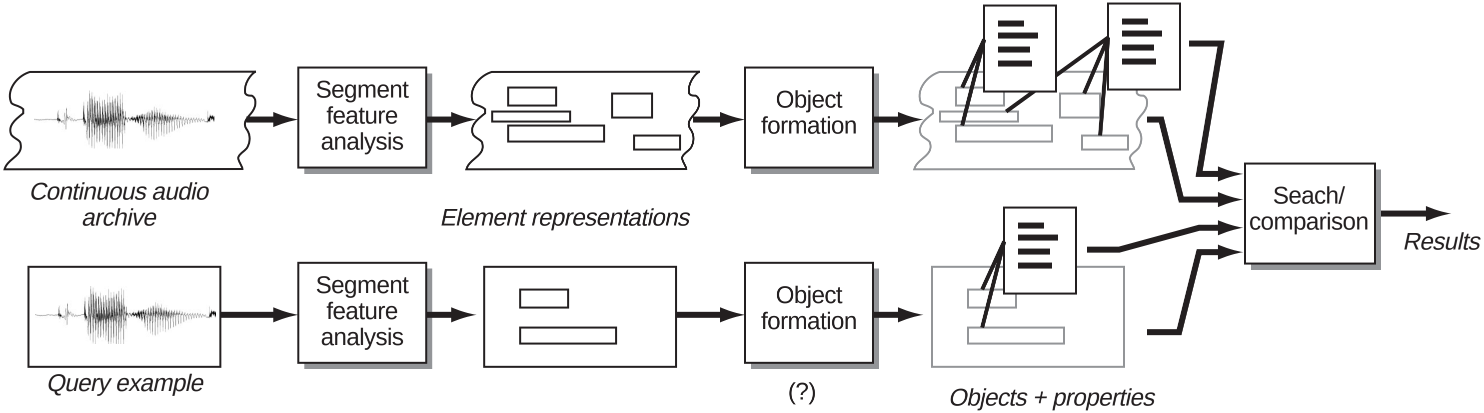
- **Using 'generic sound elements'**



  - search for subset (but: masked features?)
  - how to generalize?
  - how to use segment-style features?

Object-based audio indexing

- **Organizing elements into objects reveals higher-order properties**



- **How to form objects?**
  - heuristics (onset, harmonicity, continuity)
  - machine learning:
    associative recall, clustering, 'data mining'
- **Which higher-order properties?**
  - current wisdom (brightness, roughness...)
  - psychoacoustics
  - (semi) data-driven hierarchies

Open issues in automatic indexing

- **How to do CASA for element descriptions?**
  - PDCASA: ‘generic’ primitives
    + constraining hierarchy
  - (semi?) automatic learning of object structure
- **Classification**
  - connecting subjective & objective properties
    $\rightarrow$ finding subjective invariants, prominence
  - representation of sound-object ‘classes’
  - matching incompletely-described objects
- **Queries**
  - .. by example (which part?)
  - .. by symbolic descriptions of classes?
- **Related applications**
  - ‘structured audio encoder’
  - semantic hearing aid / robot listener

Outline

1. Auditory Scene Analysis (ASA)
2. Computational ASA (CASA)
3. Context, expectation & predictions
4. Applications: speech recognition, indexing
5. **Conclusions and open issues**

5 CASA: Open issues

- **We're still looking for the right perspective**
  - bottom up vs. top down
  - physiology, psychology, levels of description
- **What is the goal?**
  - simulating listeners on contrived tasks?
  - solving practical engineering problems?
  - laying the conceptual groundwork
- **How to evaluate CASA work?**
  - evaluation is critical for a healthy field
  - .. but people have to agree on a task
  - subjectively defined → listening tests
- **Looming on the horizon...**
  - learning
  - attention

Conclusions

- **Auditory organization is required in real environments**
- **We don’t know how listeners do it!**
  - plenty of modeling interest
- **Prediction-reconciliation can account for ‘illusions’**
  - use ‘knowledge’ when signal is inadequate
  - important in a wider range of circumstances?
- **Speech & speech recognizers**
  - urgent application for CASA
  - good source of signal knowledge?
- **Automatic indexing implies ‘synthetic listener’**
  - need to solve a lot of modeling issues
  - the next big thing?