Auditory Scene Analysis: phenomena, theories and computational models

July 1998

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

Outline

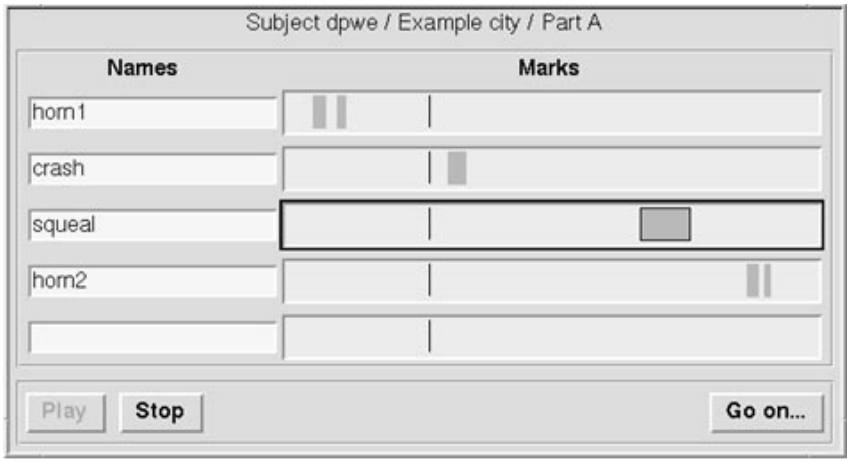
1. **The computational theory of ASA**
2. **Cues & grouping**
3. **Expectations & inference**
4. **Big issues**

Auditory Scene Analysis

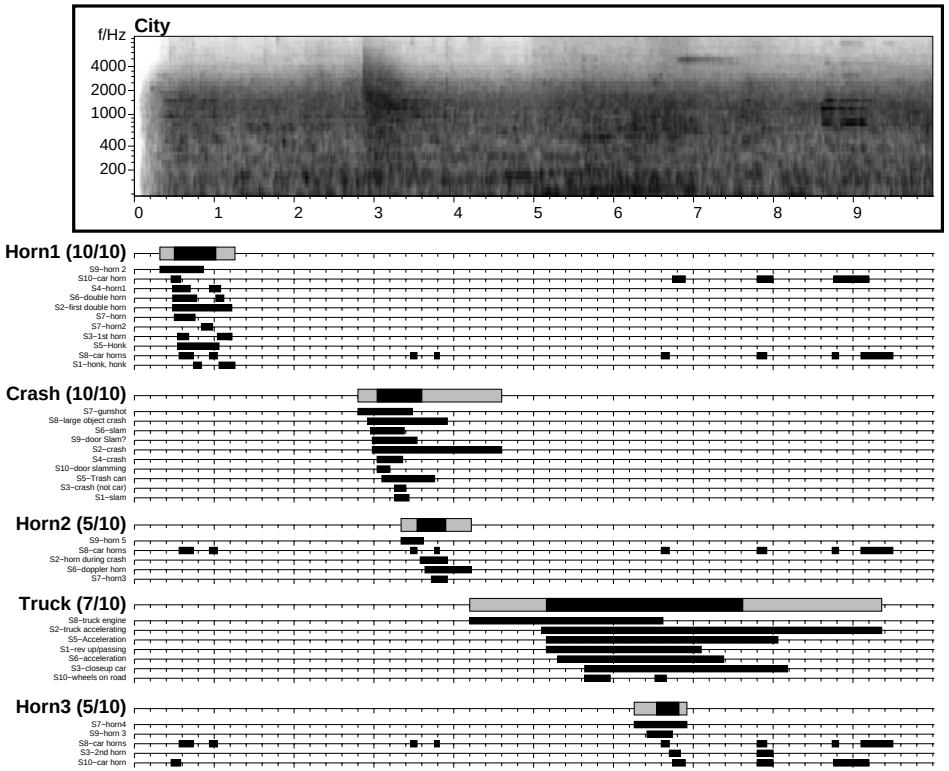
What does our sense of hearing do?

- recover useful information
- ... about objects of interest
- ... in a wide range of circumstances

Measuring objects in an auditory scene:



Subjective analysis of auditory scenes



- **Subjects identify structures in dense scenes with high agreement**

Outline

1. **The computational theory of ASA**
   - ASA and CASA
   - The grouping paradigm
   - Marr's three levels of explanation
2. **Cues & grouping**
3. **Expectations & inference**
4. **Big issues**

Auditory Scene Analysis (ASA)

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

- **Real-world sounds rarely occur in isolation**

  →a useful sense of hearing must be able to segregate mixtures

  - people (and ...) do this very well; unexpectedly difficult to model
  - depends on:

    subjective definition of relevant sources

    regularity/constraints of real-world sounds

- **Studied via experimental psychology**
  - characterize ‘rules’ for organizing simple pieces (tones, noise bursts, clicks) i.e. ‘reductive’ approach

Computational Auditory Scene Analysis (CASA)

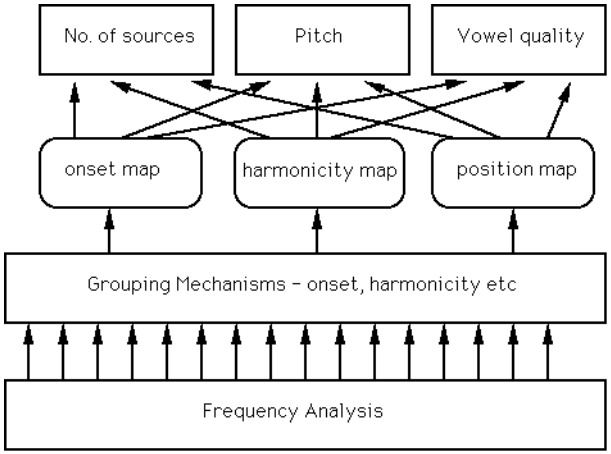
- **Psychological 'rules' suggest computer implementation**
  - .. but many practical problems arise!
- **Motivations:**
  Practical applications
  - real-world interactive systems
  - indexing of media databases
  - hearing prostheses

  Crossover opportunities
  - unknown signal/information processing principles?

  Benefits for theory
  - implementations are very revealing

The grouping paradigm

- **Standard theory of ASA (Bregman, Darwin &c):**
  - sound mixture is broken up into small elements e.g. time-frequency 'cells'
  - each element has a number of feature dimensions (amplitude, ITD, period)
  - elements are grouped together according to their features to form larger structures
  - resulting groups have overall attributes (pitch, location)



(from Darwin 1996)

Marr's levels-of-explanation of information processing

- **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? - to help organize understanding
- avoid confusion/wasted effort

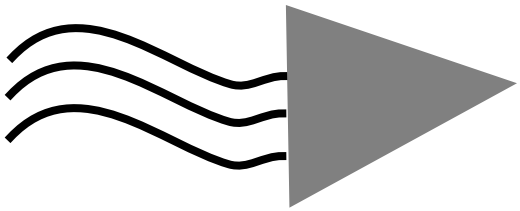
→use as an analysis tool...

Level 1: Computational theory

- **The underlying regularities that make the problem possible**
  - i.e. the 'ecological' facts
- **Implicit definition of "what is a source?":**

  Independence of attributes between sources

  Continuity of attributes for each source



+ other source-specific constraints

Level 2: Algorithm

- **A particular approach to exploiting the constraints of the computational theory**
  - both process & representation
- **Audition: the "elements-then-grouping" approach**
  - could have been otherwise e.g. templates
- **Often the focus of analysis**
  - but: debate is muddled without a clear computational theory

Level 3: Implementation

- **A specific realization of the algorithm**
  - computer programs
  - neurons
  - ...
- **Can be analyzed separately?**
  - provided epiphenomena are correctly assigned
- **Needs context of algorithm, computational theory**

  "You cannot understand stereopsis simply by thinking about neurons"

The advantage of the appropriate level

- **Computational theory**
  - determines the purpose of the process; provides focus necessary for analysis

  e.g. biosonar: benefit of hyperresolution
- **Algorithm**
  - abstraction that is still specific, transferable

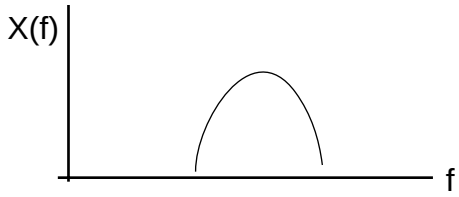
  e.g. autocorrelation for pitch
- **Implementation**
  - explain ‘epiphenomena’

  e.g. ‘subjective octave’ from refractory period

An example: Neural inhibition

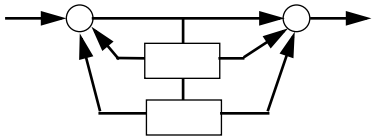
Computational theory

Frequency-domain processing



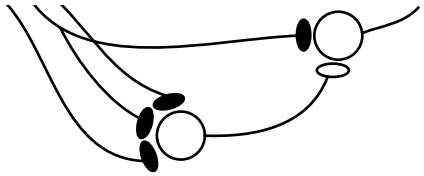
Algorithm

Discrete-time filtering (subtraction)



Implementation

Neurons with GABAergic inhibitions



Summary 1

- Acoustic scenes are very complex
- .. but the auditory system extracts useful information
- Grouping is the main focus of Auditory Scene Analysis
- .. but it fits into a larger Marrian framework

Outline

1. The computational theory of ASA
2. **Cues & grouping**
   - Cue analysis
   - Simple scenes
   - Models
   - Complications: interaction, ambiguity, time
3. Expectations & inference
4. Big issues

Cues to grouping

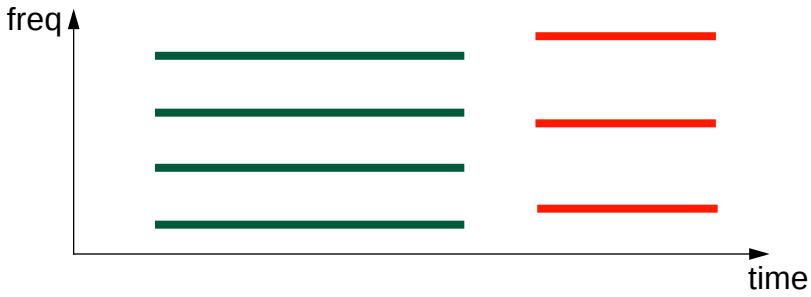
- **Common onset/offset/modulation (“fate”)**
- **Common periodicity (“pitch”)**

	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**



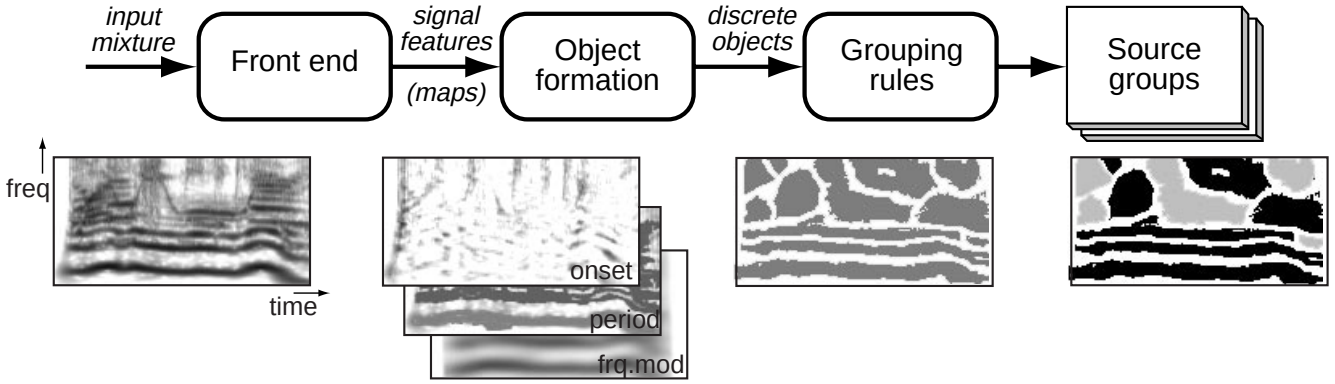
Computational theory
- common onset
- common period (harmonicity)

Algorithm
- locate elements (tracks)
- group by shared features

Implementation
- ? exhaustive search
- evolution in time

Computer models of grouping

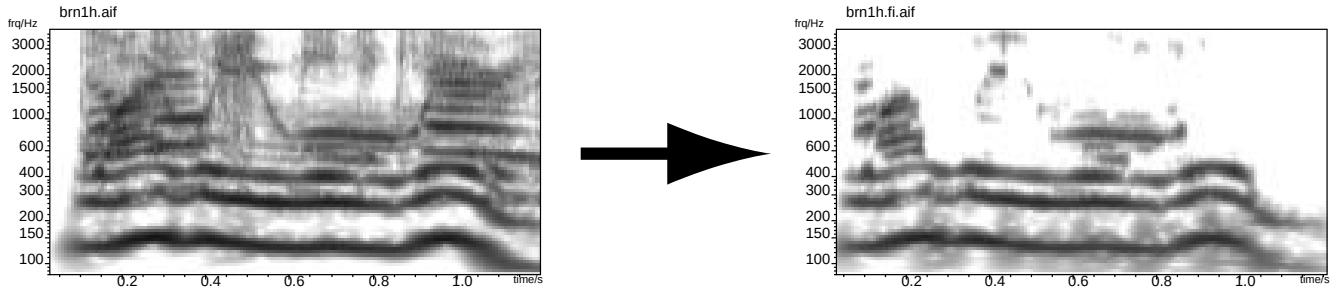
- **“Bregman at face value” (e.g. Brown 1992):**



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

Grouping model results

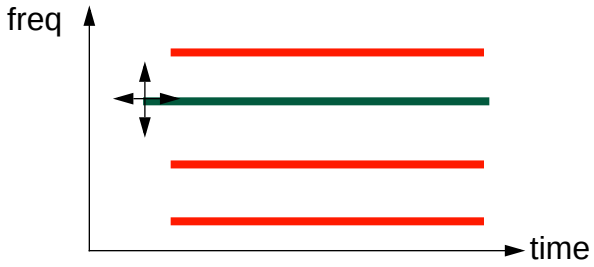
- **Able to extract voiced speech:**



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

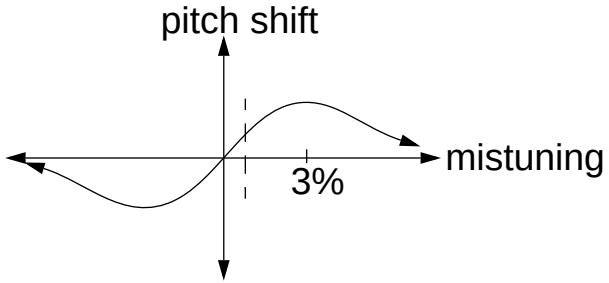
Complications for grouping: 1: Cues in conflict

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



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

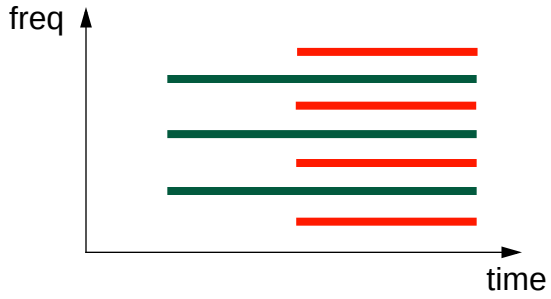
- **Gradual, various results:**



  - 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 .. ?

Summary 2

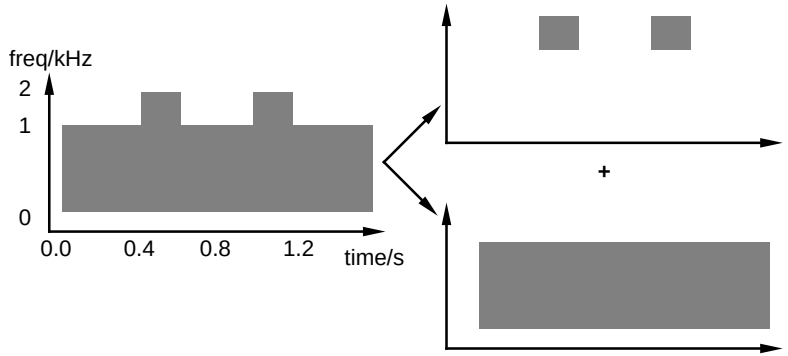
- Known grouping cues make sense
- Simple examples are straightforward
- Models can be implemented directly
- .. but problematic situations abound

Outline

1. The computational theory of ASA
2. Cues & grouping
3. **Expectations & inference**
   - “Old-plus-new”
   - Streaming
   - Restoration & illusions
   - Top-down models
4. Big issues

The effect of context

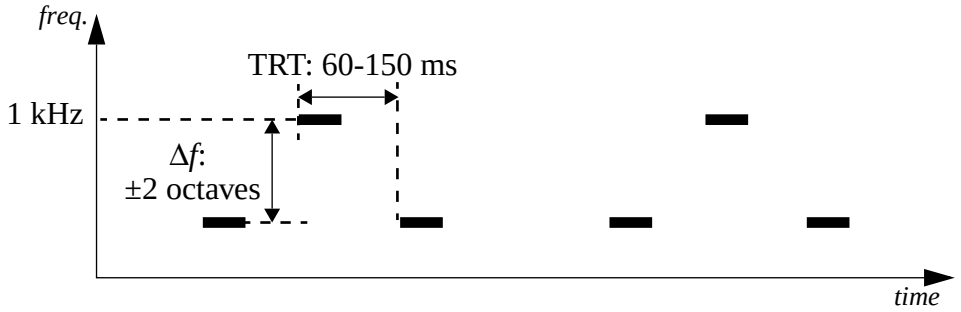
- **Context can create an 'expectation':**
  i.e. a bias towards a particular interpretation
- **e.g. Bregman's "old-plus-new" principle:**
  A change in a signal will be interpreted as an *added* source whenever possible



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

Streaming

- **Successive tone events form separate streams**

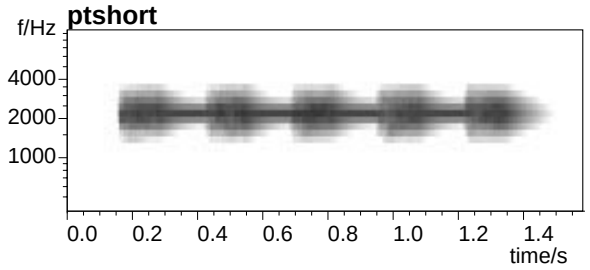


- **Order, rhythm &c *within*, not *between*, streams**

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**
  →make best guess using available information
- **E.g. the 'continuity illusion':**



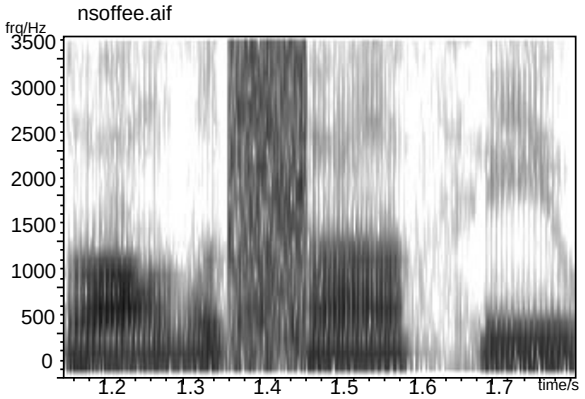
  - tones alternates with noise bursts
  - noise is strong enough to mask tone
    ... so listener discriminate presence
  - continuous tone distinctly perceived
    for gaps ~100s of ms

→ **Inference acts at low, preconscious level**

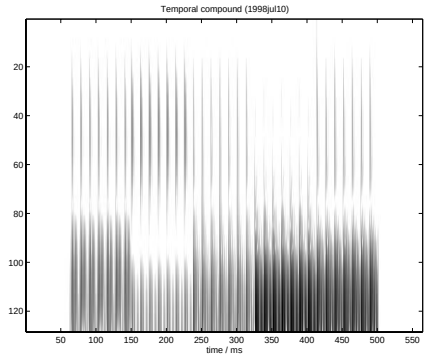
Speech restoration

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

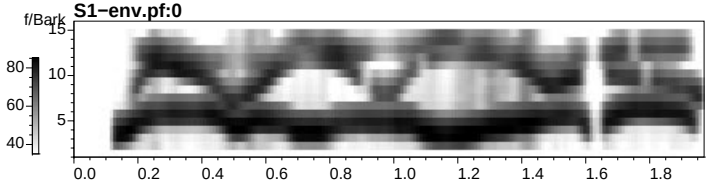
- **Phonemic restoration**



- **Temporal compounds**



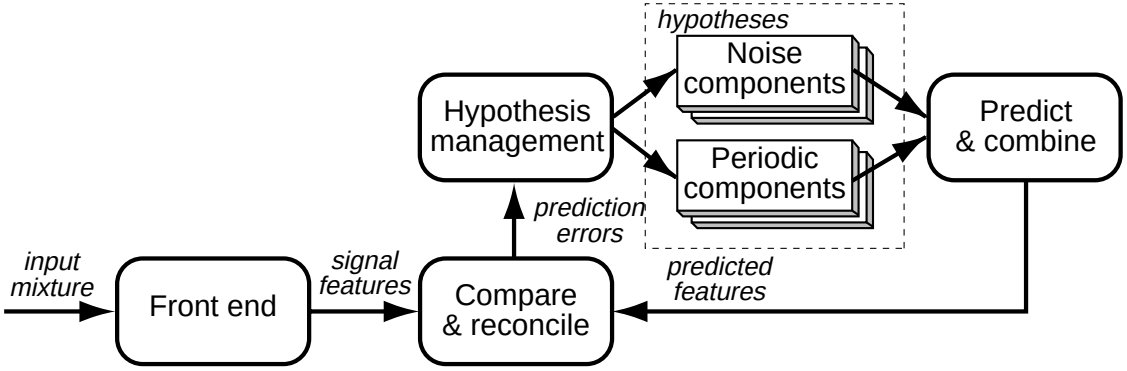
- **Sinewave speech (duplex?)**



Models of top-down processing

Perception as a *search* for *plausible explanations*

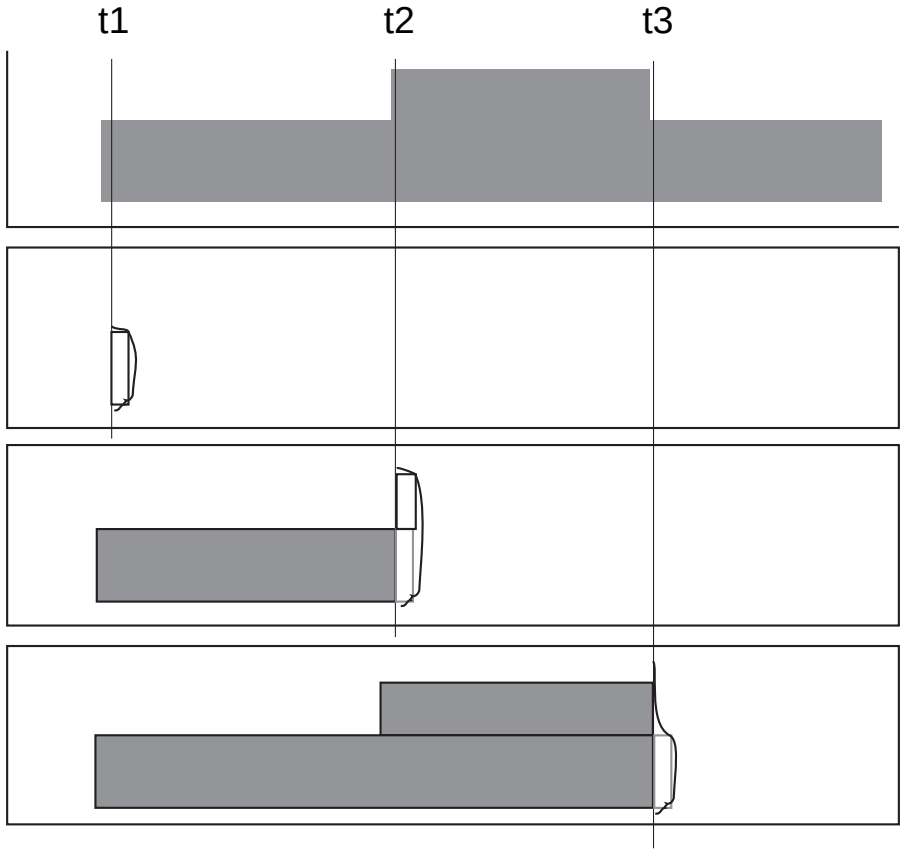
- **'Prediction-driven' CASA (PDCASA):**



- **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

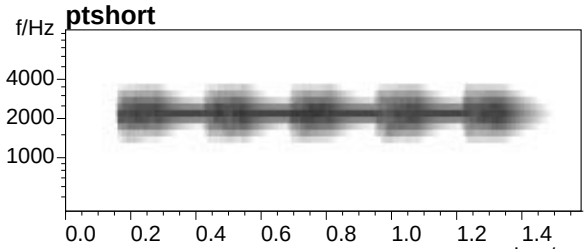
- **Incremental analysis**



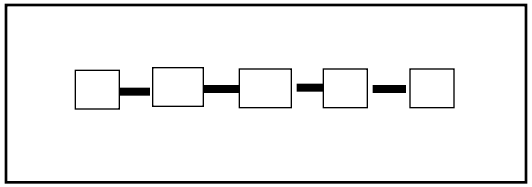
PDCASA for the continuity illusion

- **Subjects hear the tone as continuous**

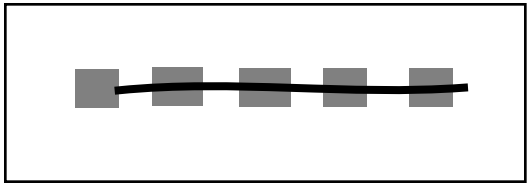
  ... if the noise is a plausible masker



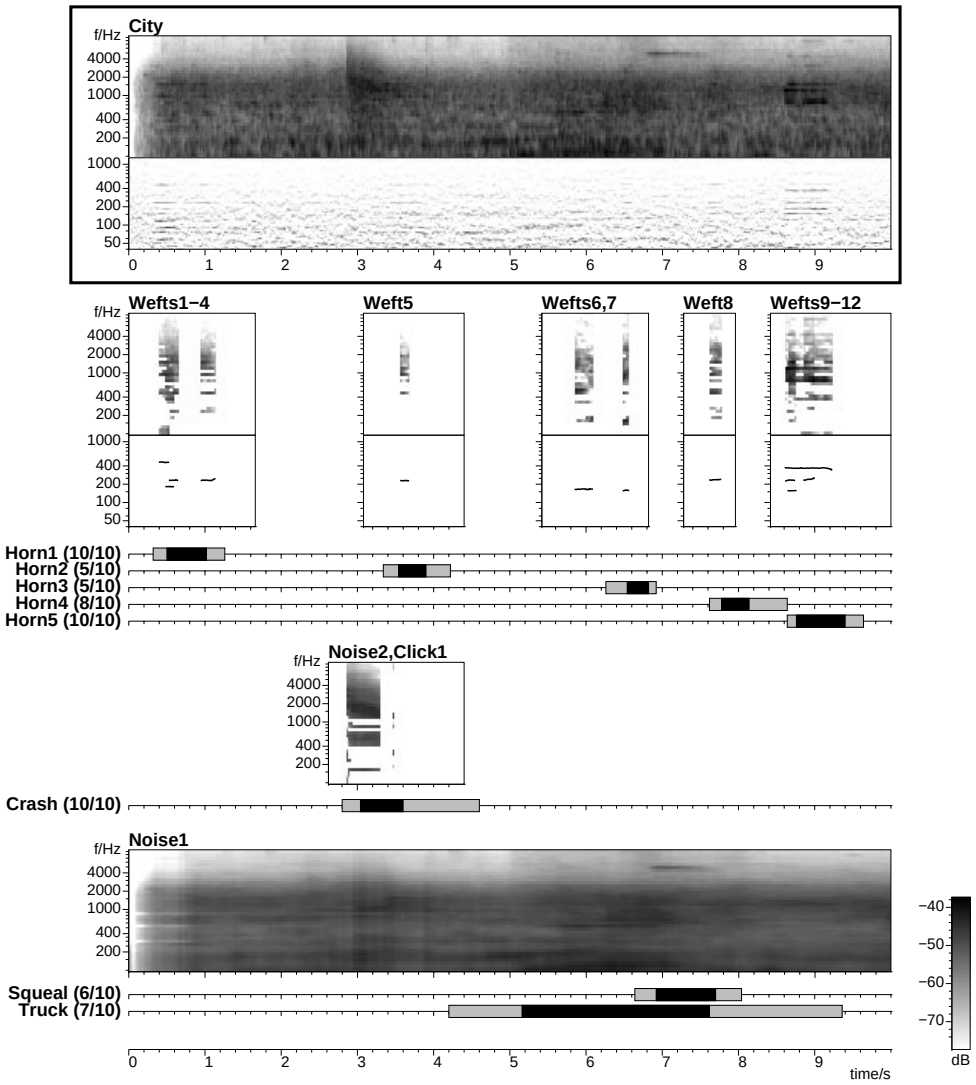
- **Data-driven analysis gives just visible portions:**



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



PDCASA analysis of a complex scene



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..."

Summary 3

- **Perceptual processing is highly context-dependent**
- **Auditory system will use prior knowledge to fill-in gaps (subconsciously)**
- **Prediction-reconciliation models can encompass this behavior**

Outline

1. The computational theory of ASA
2. Cues & grouping
3. Expectations & inference
4. **Big issues**
   - the state of ASA and CASA
   - outstanding issues
   - discussion points

The current state of ASA and CASA

- **ASA**
  - detailed descriptions of “in vitro” tests
  - some quite subtle effects explained (DV beats)

  but: how to extend to complex scenarios?
- **CASA**
  - numerous models, some convergence (mainly periodicity-based)
  - best results sound impressive (least plausible systems!)
  - applications in speech recognition?

  but: domains limited, poor robustness

Big issues in CASA:

- **Plausibility**
  - correct level for human correspondence?
  - which phenomena are important to match?
  - how to implement symbolic-style processing?
- **Top-down vs. bottom-up**
  - different approaches to ambiguity, latency
  - how far down for top-down?
  - how far ‘up’ for high level?
  - choice between extraction & inference?
- **Integrating multiple cues (e.g. binaural)**
- **Other debates:**
  - what is the real goal?
  - resynthesis
  - evaluation

Big issues in ASA & CASA:

- **Knowledge:**
  how to acquire, represent & store ...
  - short-term: context
  - long-term: memories
  - abstract: classes, generalities
- **Attention:**
  - what does it mean in these models?
  - limitation or important principle?

Conclusions

- **Real-world sounds are complex; scene-analysis is required**
- **We know certain cues & some rules, but real situations raise contradictions**
- **Current models handle ‘obvious’ cases; robustness & generality are hard**
- **Many issues remain**

Discussion points

- **Are Marr's levels important? Useful? Can you study levels in isolation?**
- **What do restoration phenomena imply about internal representations?**
- **Do we have an adequate account of an ASA algorithm? e.g. where do hypotheses come from?**
- **How important/challenging are phenomena like duplex perception, sinewave speech etc.?**