united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

SMR: 1343/10

EU ADVANCED COURSE IN COMPUTATIONAL NEUROSCIENCE

An IBRO Neuroscience School

(30 July - 24 August 2001)

"Computational Models of Auditory Organization"

presented by:

Dan ELLIS
Columbia University
Department of Electrical Engineering
500 West 120th Street, Room 1312
New York, NY 10027
U.S.A.

Computational Models of Auditory Organization

Dan Ellis
Electrical Engineering, Columbia University
http://www.ee.columbia.edu/~dpwe/

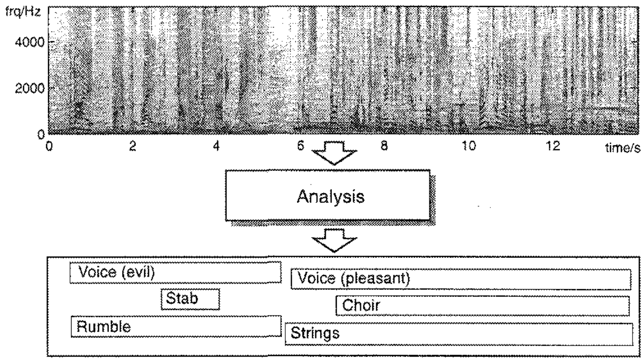
Outline

1. Sound organization
2. Human Auditory Scene Analysis (ASA)
3. Computational ASA (CASA)
4. CASA issues & applications
5. Summary

Outline

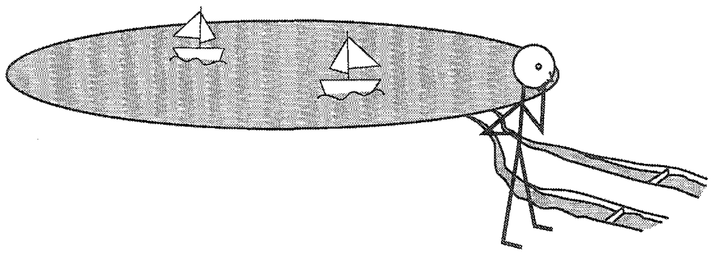
1. Sound organization
   - the information in sound
   - Marr's levels of explanation
2. Human Auditory Scene Analysis (ASA)
3. Computational ASA (CASA)
4. CASA issues & applications
5. Summary

❶ Sound organization



- **Central operation:**
  - continuous sound mixture
    → distinct objects & events
- **Perceptual impression is very strong**
  - but hard to 'see' in signal

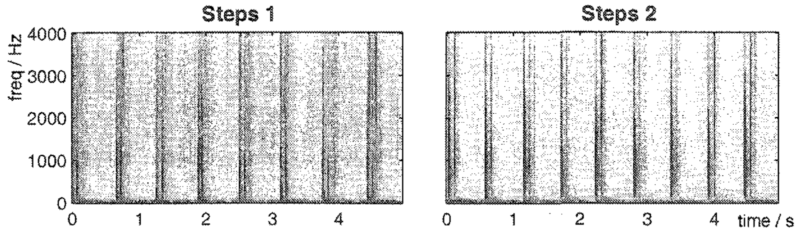
Bregman's lake



"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**
  - two sensors, N signals ...
- **Disentangling mixtures as primary goal**
  - perfect solution is not possible
  - need knowledge-based *constraints*

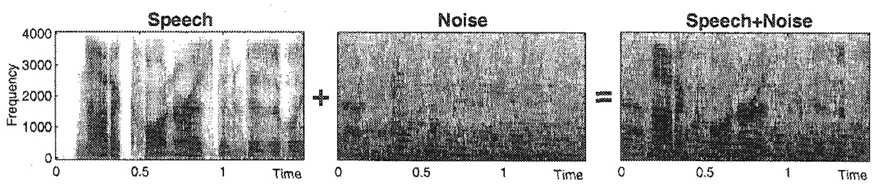
The information in sound



- **A sense of hearing is evolutionarily useful**
  - gives organisms 'relevant' information
- **Auditory perception is *ecologically* grounded**
  - scene analysis is preconscious ($\rightarrow$ illusions)
  - special-purpose processing reflects 'natural scene' properties
  - subjective *not* canonical (ambiguity)

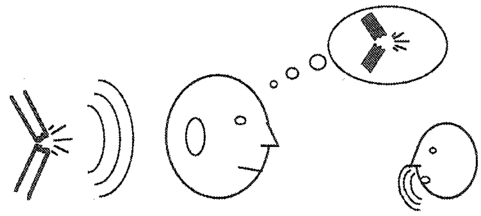
Sound mixtures

- **Sound 'scene' is almost always a mixture**
  - always stuff going on
  - sound is 'transparent'



- **Need information related to our 'world model'**
  - i.e. separate objects
  - a wolf howling in a blizzard is the same as a wolf howling in a rainstorm
  - whole-signal statistics won't do this
- **'Separateness' is similar to independence**
  - objects/sounds that change in isolation
  - but: depends on the situation e.g. passing car vs. mechanic's diagnosis

Vision and hearing



- **Hearing and seeing are complementary**
  - hearing is omindirectional
  - hearing works in the dark
- **Reveal different things about the world**
  - vision is good for examining static situations
  - physical motion almost always makes sound

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

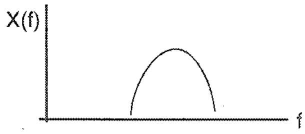
- **Three distinct aspects to info. processing**

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

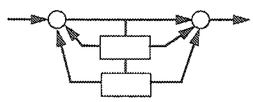
Why bother?
- helps organize interpretation
- it's OK to consider levels separately, one at a time

An example: Neural inhibition *

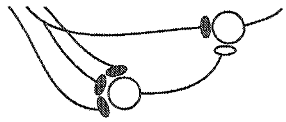
Computational theory — Frequency-domain processing



Algorithm — Discrete-time filtering (subtraction)



Implementation — Neurons with GABAergic inhibitions

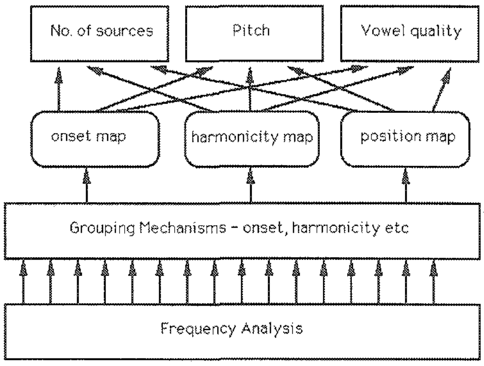


Outline

1. Sound organization
2. **Human Auditory Scene Analysis (ASA)**
   - experimental psychoacoustics
   - grouping and cues
   - perception of complex scenes
3. Computational ASA (CASA)
4. CASA issues & applications
5. Summary

② Human Sound Organization

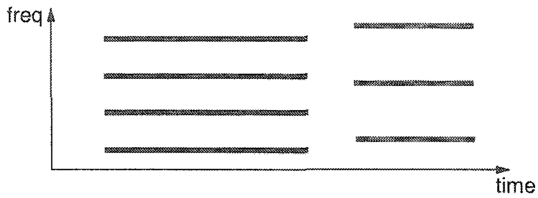
- **"Auditory Scene Analysis" [Bregman 1990]**
  - break mixture into small *elements* (in time-freq)
  - elements are *grouped* in to sources using *cues*
  - sources have aggregate *attributes*
- **Grouping 'rules' (Darwin, Carlyon, ...):**
  - cues: common onset/offset/modulation, harmonicity, spatial location, ...



(from Darwin 1996)

Cues to simultaneous grouping

- **Common attributes and 'fate'**

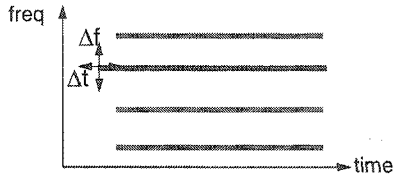


	Common onset	Periodicity
Computational theory	Acoustic consequences tend to be synchronized	(Nonlinear) cyclic processes are common
Algorithm	Group elements that start in a time range	Place patterns? Autocorrelation?
Implementation	Onset detector cells Synchronized osc's?	Delay-and-mult? Modulation spect?

- **+ Spatial location (ITD, ILD, spectral)...**

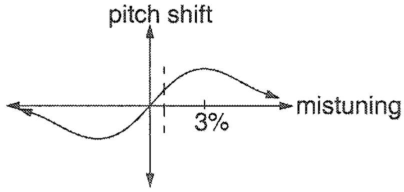
Complications for grouping: 1: Cues in conflict

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



  - determine how Δt and Δf affect
    - segregation of harmonic
    - pitch of complex

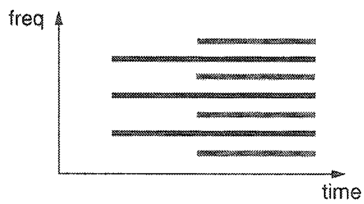
- **Gradual, various results:**



http://www.dcs.shef.ac.uk/~martin/MAD/docs/mad.htm

Complications for grouping: 2: The effect of time

- **Added harmonics:**



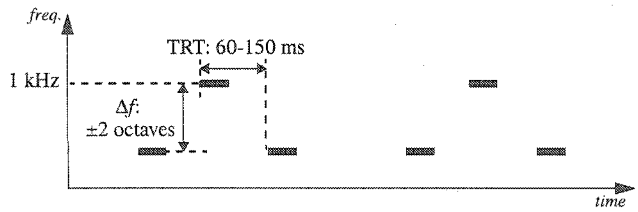
  - onset cue initially segregates; periodicity eventually fuses

- **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 .. ?

Sequential grouping: Streaming

- **Successive tone events form separate streams**

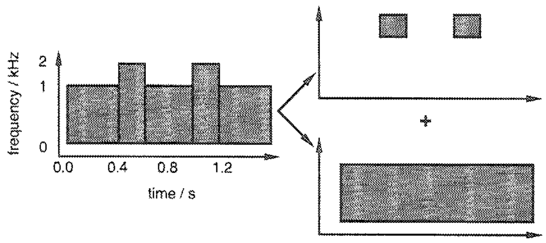


- **Order, rhythm &c *within*, not *between*, streams**

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

The effect of context

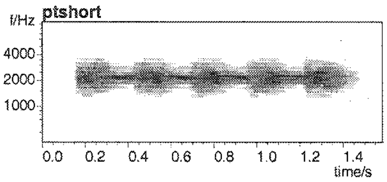
- **Context can create an 'expectation': i.e. a bias towards a particular interpretation**
- **e.g. Bregman's "old-plus-new" principle:**
  A change in a signal will be interpreted as an *added* source whenever possible



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

Restoration & illusions

- **'Illusions' illuminate algorithm**
  - what model would 'misbehave' this way?
- **E.g. the 'continuity illusion':**
  - making 'best guess' for masked information



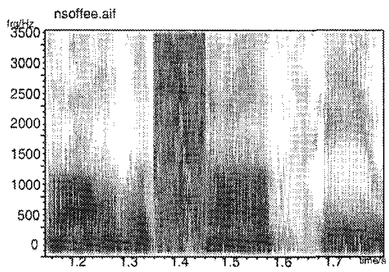
  - tones alternates with noise bursts
  - noise is strong enough to mask tone
  - continuous tone distinctly perceived for gaps ~100s of ms

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

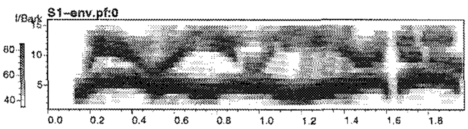
Speech restoration

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

- **Phonemic restoration**



- **Sinewave speech (duplex?)**

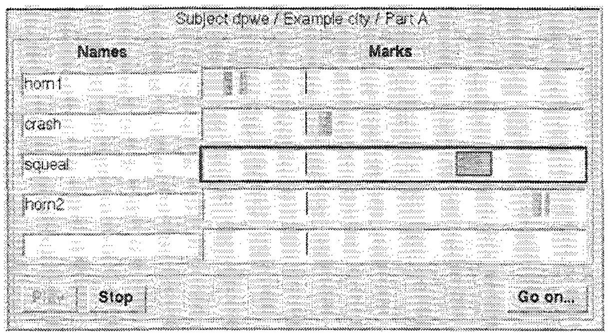


Ground truth in complex scenes

*

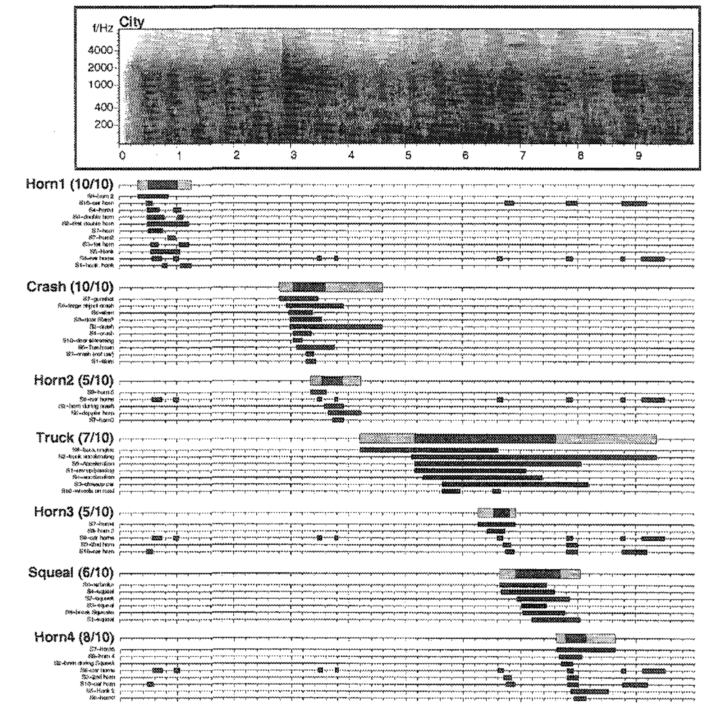
- **What do people hear in sound mixtures?**
  - do interpretations match?

→ **Listening tests to collect 'perceived events':**



Ground-truth results

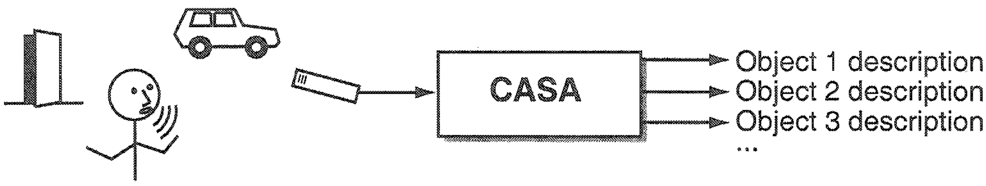
*



Outline

1. Sound organization
2. Human Auditory Scene Analysis (ASA)
3. **Computational ASA (CASA)**
   - bottom-up models
   - top-down predictions
   - other approaches
4. CASA issues & applications
5. Summary

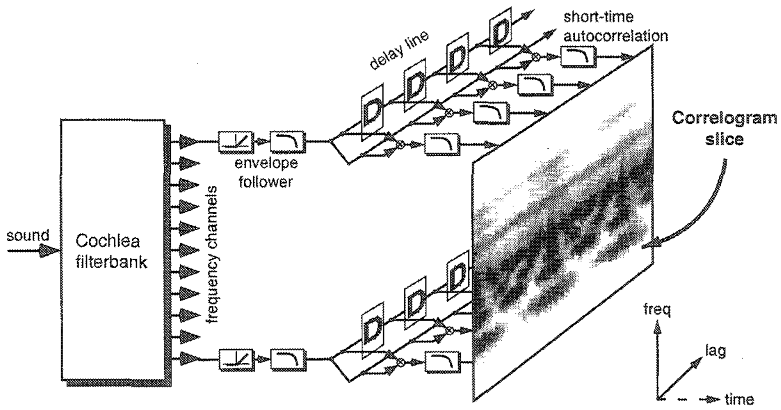
3 Computational ASA



- **Goal: Automatic sound organization ; Systems to 'pick out' sounds in a mixture**
  - ... like people do
- **E.g. voice against a noisy background**
  - to improve speech recognition
- **Approach:**
  - psychoacoustics describes grouping 'rules'
  - ... just implement them?

CASA front-end processing

- **Correlogram: Loosely based on known/possible physiology**

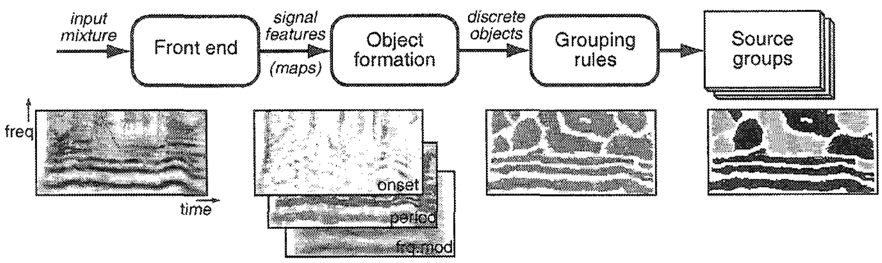


  - linear filterbank cochlear approximation
  - static nonlinearity
  - zero-delay slice is like spectrogram
  - periodicity from delay-and-multiply detectors

The Representational Approach

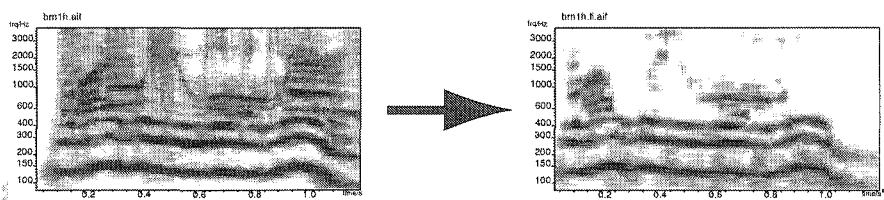
(Brown & Cooke 1993)

- **Implement psychoacoustic theory**

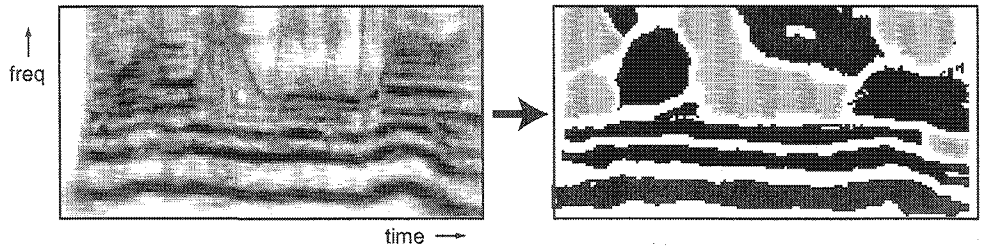


  - 'bottom-up' processing
  - uses common onset & periodicity cues

- **Able to extract voiced speech:**



Problems with 'bottom-up' CASA

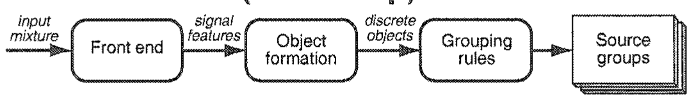


- **Circumscribing time-frequency elements**
  - need to have 'regions', but hard to find
- **Periodicity is the primary cue**
  - how to handle aperiodic energy?
- **Resynthesis via masked filtering**
  - cannot separate within a single t-f element
- **Bottom-up leaves no ambiguity or context**
  - how to model illusions?

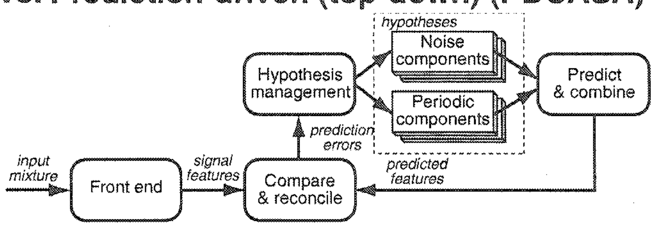
Adding top-down cues

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

- **Data-driven (bottom-up)...**



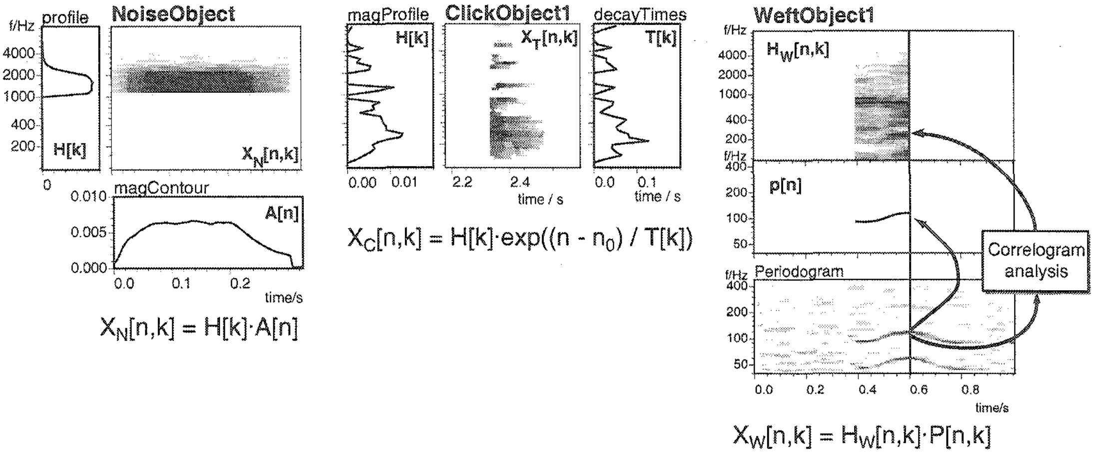
vs. Prediction-driven (top-down) (PDCASA)



- **Motivations**
  - detect non-tonal events (noise & click elements)
  - support 'restoration illusions'...
    → hooks for high-level knowledge
  + 'complete explanation', multiple hypotheses, ...

Generic sound elements for PDCASA

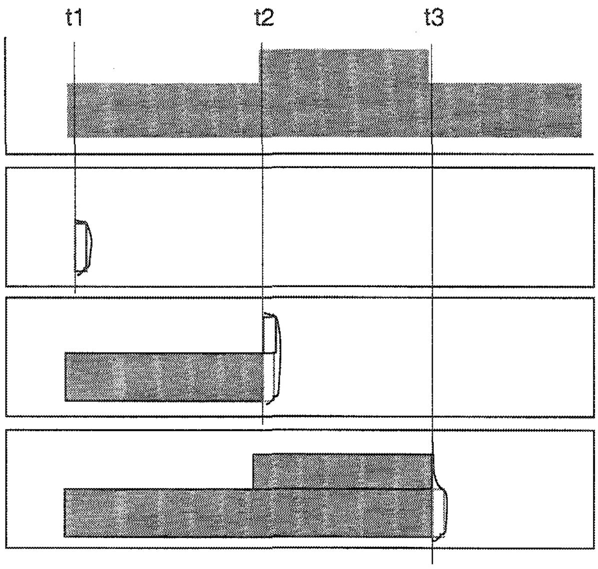
- **Goal is a representational space that**
  - covers real-world perceptual sounds
  - minimal parameterization (sparseness)
  - separate attributes in separate parameters



- **Object hierarchies built on top...**

PDCASA for old-plus-new

- **Incremental analysis**



Input signal

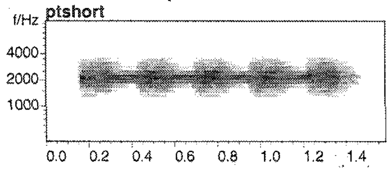
Time t1:
initial element
created

Time t2:
Additional
element required

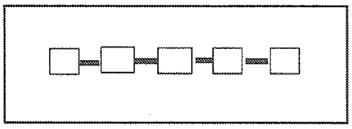
Time t3:
Second element
finished

PDCASA for the continuity illusion

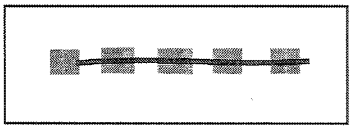
- **Subjects hear the tone as continuous**
  ... if the noise is a plausible masker



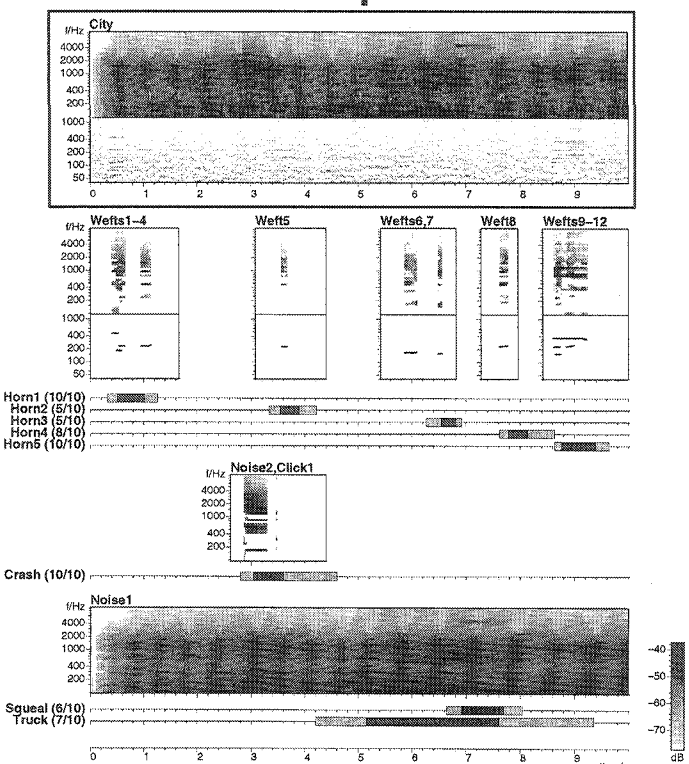
- **Data-driven analysis gives just visible portions:**



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



PDCASA and complex scenes



Marrian analysis of PDCASA *

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

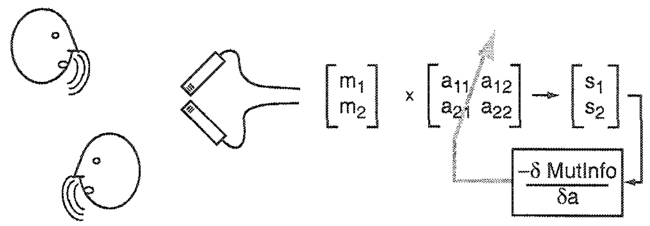
Computational theory	• Objects persist predictably • Observations interact irreversibly
Algorithm	• Build hypotheses from generic elements • Update by prediction-reconciliation
Implementation	???

"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..."

Other approaches: ICA *

(Bell & Sejnowski etc.)

- **General idea:**
  Drive a parameterized separation algorithm to maximize indepdendence of outputs

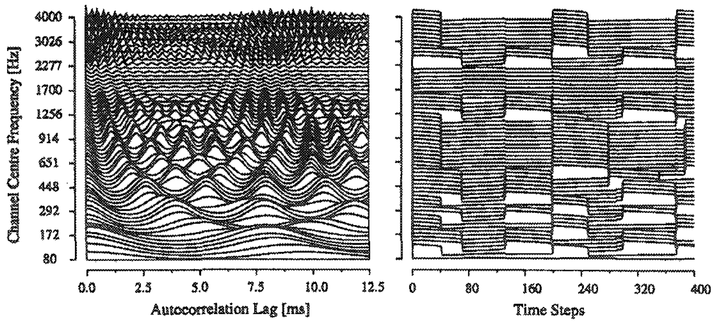


- **Attractions:**
  - mathematically rigorous, minimal assumptions
- **Problems:**
  - limitations of separation algorithm (N x N)
  - essentially bottom-up

Other approaches: Neural Oscillators *

(Malsburg, Wang & Brown)

- **Locally-excited, globally-inhibited networks form separate phases of synchrony**



- **Advantages:**
  - avoid implausible AI methods (search, lists)
  - oscillators substitute for iteration
- **Only concerns the implementation level?**

Outline

1. Sound organization
2. Human Auditory Scene Analysis (ASA)
3. Computational ASA (CASA)
4. **CASA issues & applications**
   - learning
   - missing data
   - audio information retrieval
   - the machine listener
5. Summary

Learning & acquisition

*

- **The speech recognition lesson: How to exploit large databases?**
- **'Maximum likelihood' sound organization (e.g. Roweis)**
  - learn model→sound distributions $P(X|M)$ by analyzing isolated sound databases
  - combine models with physics: $P(X|\{M_i\})$
  - learn patterns of model combinations $P(\{M_i\})$
  - search for most likely combinations of models to explain observed sound mixtures
    $\max P(\{M_i\}|X) = P(X|\{M_i\}) \cdot P(\{M_i\})$
- **Short-term learning**
  - hearing a particular source can alter short-term interpretations of mixtures

Missing data recognition

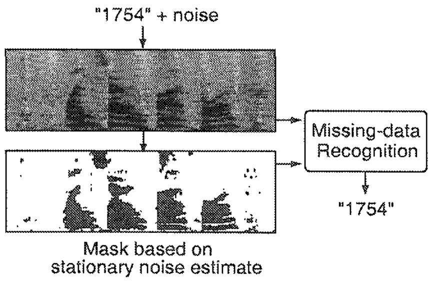
*

(Cooke, Green, Barker... @ Sheffield)

- **Energy overlaps in time-freq. hide features**
  - some observations are effectively missing
- **Use missing feature theory...**
  - integrate over missing data dimensions x_m

$$p(x|q) = \int p(x_g|x_m, q)p(x_m|q)dx_m$$

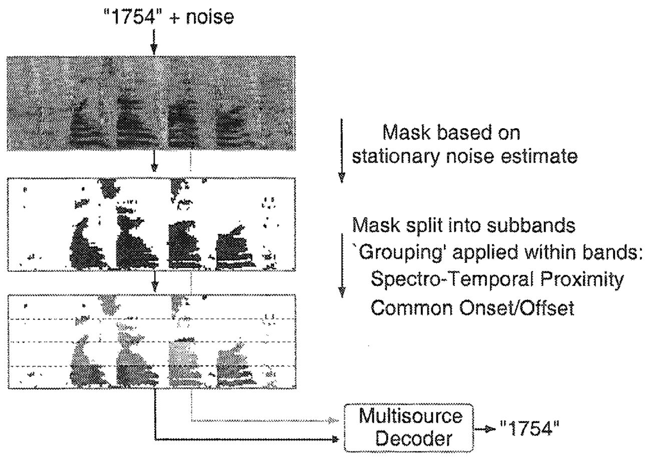
- **Effective in speech recognition**
  - trick is finding good/bad data mask



Multi-source decoding

(Jon Barker @ Sheffield)

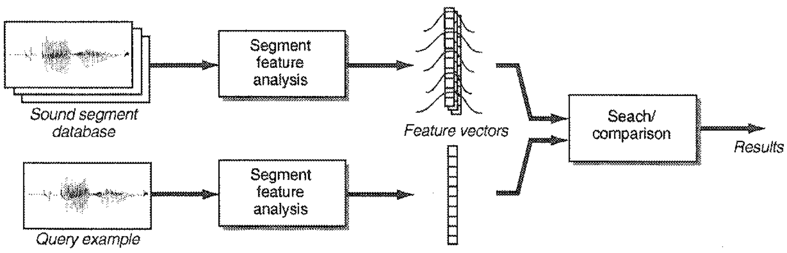
- **Search of sound-fragment interpretations**



- **CASA for masks/fragments**
  - larger fragments → quicker search
- **Use with nonspeech models?**

Audio Information Retrieval

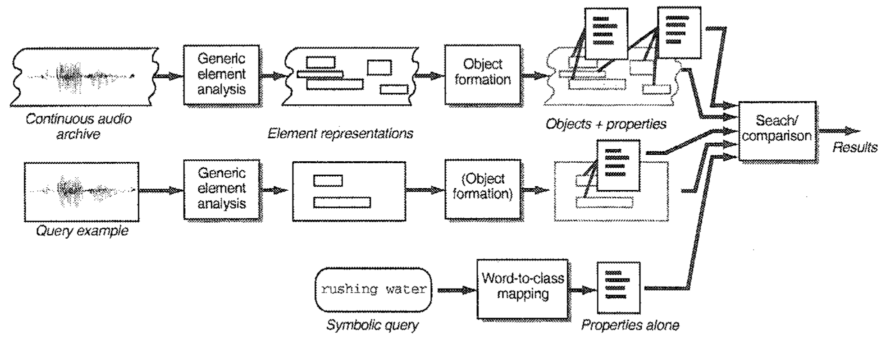
- **Searching in a database of audio**
  - speech .. use ASR
  - text annotations .. search them
  - sound effects library?
- **e.g. Muscle Fish "SoundFisher" browser**
  - define multiple 'perceptual' feature dimensions
  - search by proximity in weighted feature space



  - features are 'global' for each soundfile, no attempt to separate mixtures

CASA for audio retrieval

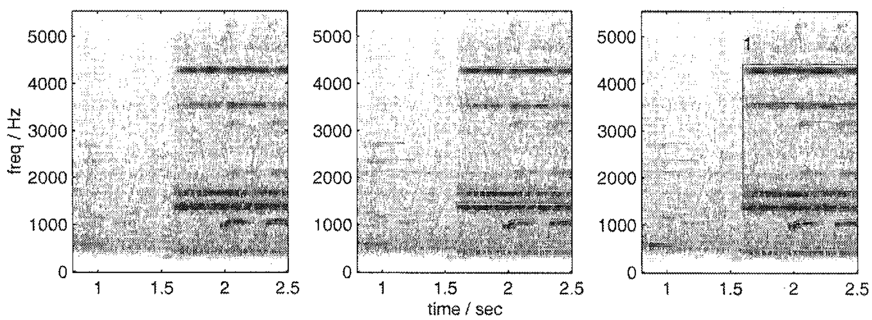
- **When audio material contains mixtures, global features are insufficient**
- **Retrieval based on element/object analysis:**



  - features are calculated over grouped subsets

Alarm sound detection

- **Alarm sounds have particular structure**
  - people 'know them when they hear them'
- **Isolate alarms in sound mixtures**



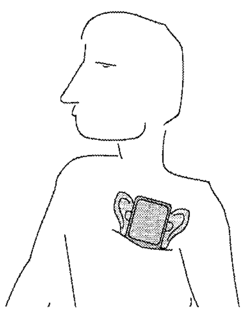
  - representation of energy in time-frequency
  - formation of atomic elements
  - grouping by common properties (onset &c.)
  - classify by attributes...
- **Key: recognize *despite* background**

Future prosthetic listening devices

- **CASA to replace lost hearing ability**
  - sound mixtures are difficult for hearing impaired
- **Signal enhancement**
  - resynthesize a single source without background
  - (need very good resynthesis)
- **Signal understanding**
  - monitor for particular sounds (doorbell, knocks) & translate into alternative mode (vibro alarm)
  - real-time textual descriptions i.e. "automatic subtitles for real life"

The 'Machine listener'

- **Goal: An auditory system for machines**
  - use same environmental information as people
- **Aspects:**
  - recognize spoken commands (but not others)
  - track 'acoustic channel' quality (for responses)
  - categorize environment (conversation, crowd...)
- **Scenarios**



  - personal listener → summary of your day
  - autonomous robots: need awareness

Outline

1. Information from sound
2. Human Auditory Scene Analysis (ASA)
3. Computational ASA (CASA)
4. CASA issues & applications
5. **Summary**

5 Summary

- **Sound contains lots of information**
  ... but it's not easy to extract
- **We know a little about how humans hear**
  ... at least for simplified sounds
- **We have some ways to copy it**
  ... which we hope to improve
- **CASA would have many useful applications**
  ... machines to listen and remember for us

Further reading

[BarkCE00] J. Barker, M.P. Cooke & D. Ellis (2000). "Decoding speech in the presence of other sound sources," *Proc. ICSLP-2000*, Beijing. ftp://ftp.icsi.berkeley.edu/pub/speech/papers/icslp00-msd.pdf

[Breg90] A.S. Bregman (1990). *Auditory Scene Analysis: the perceptual organization of sound*, MIT Press.

[BrowC94] G.J. Brown & M.P. Cooke (1994). "Computational auditory scene analysis," *Computer Speech and Language* 8, 297-336.

[Chev00] A. de Cheveigné (2000). "The Auditory System as a Separation Machine," Proc. Intl. Symposium on Hearing. http://www.ircam.fr/pcm/cheveign/sh/ps/ATReats98.pdf

[CookE01] M. Cooke, D. Ellis (2001). "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication* (accepted for publication). http://www.ee.columbia.edu/~dpwe/pubs/tcfkas.pdf

[DarC95] C.J. Darwin, R.P. Carlyon (1995). "Auditory Grouping," in *The Handbook of Perception and Cognition*, Vol 6, Hearing (ed: B.C.J. Moore), Academic Press, 387-424.

[Ellis99] D.P.W. Ellis (1999). "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures," *Speech Communications* 27. http://www.icsi.berkeley.edu/~dpwe/research/spcomcasa98/spcomcasa98.pdf