Automatic audio analysis for content description & indexing

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

Outline

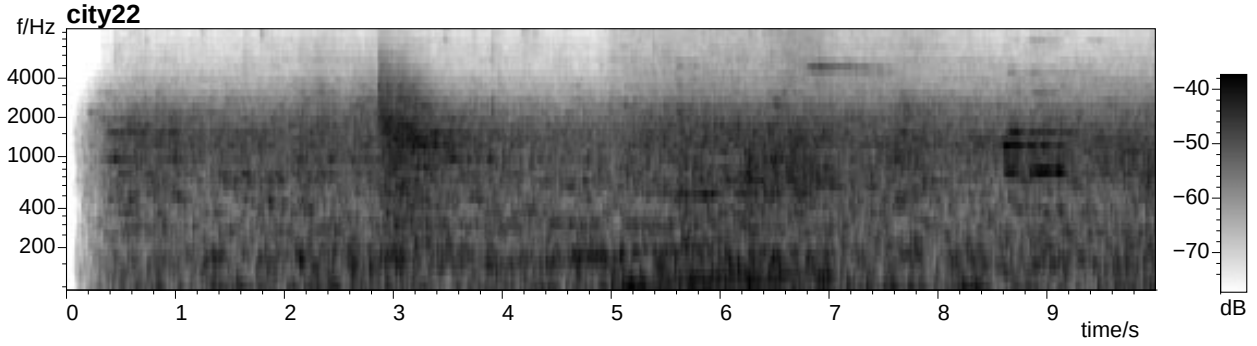
1. **Auditory Scene Analysis (ASA)**
2. **Computational ASA (CASA)**
3. **Prediction-driven CASA**
4. **Speech recognition & sound mixtures**
5. **Implications for content analysis**

1 Auditory Scene Analysis

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

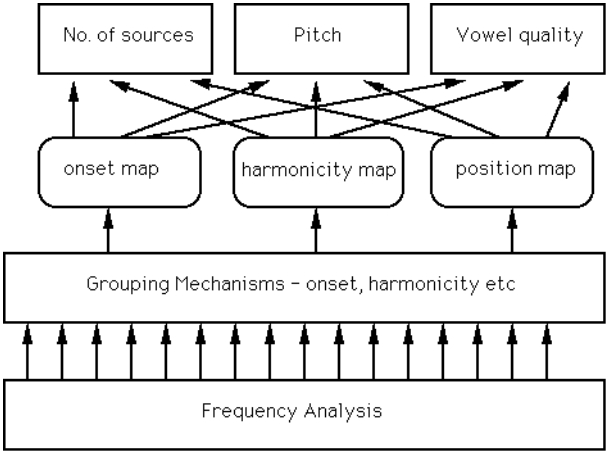
- **Sounds rarely occur in isolation**
  - organization required for useful information
- **Human audition is very effective**
  - unexpectedly difficult to model
- **‘Correct’ analysis defined by goal**
  - source shows independence, continuity

→ecological constraints enable organization



Psychology of ASA

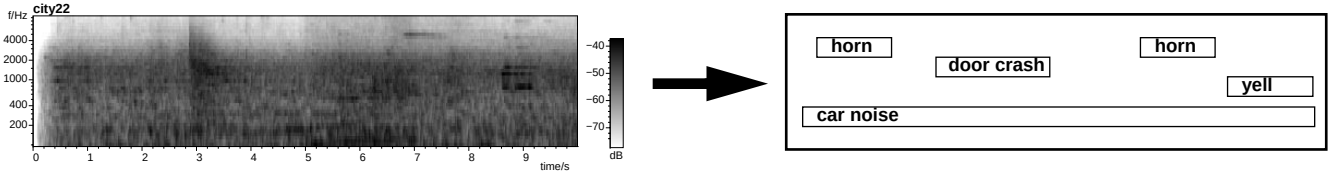
- **Extensive experimental research**
  - organization of 'simple pieces' (sinusoids & white noise)
  - streaming, pitch perception, 'double vowels'
- **"Auditory Scene Analysis" [Bregman 1990]**
  → grouping 'rules'
  - common onset/offset/modulation, harmonicity, spatial location
- **Debated... (Darwin, Carlyon, Moore, Remez)**



(from Darwin 1996)

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 ...

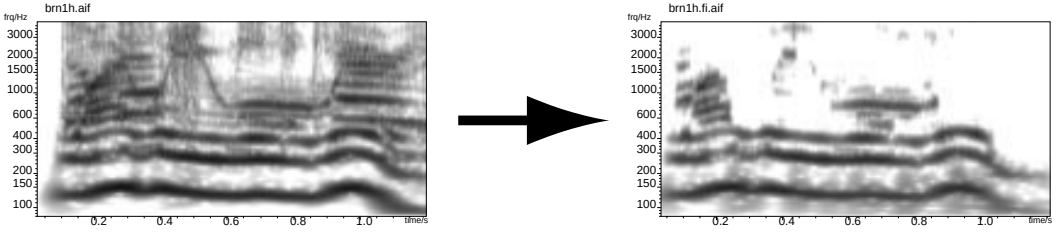
- **Motivations & Applications**
  - it's a puzzle: new processing principles?
  - real-world interactive systems (speech, robots)
  - hearing prostheses (enhancement, description)
  - advanced processing (remixing)
  - multimedia indexing...

CASA survey

- **Early work on co-channel speech**
  - listeners benefit from pitch difference
  - algorithms for separating periodicities
- **Utterance-sized signals need more**
  - cannot predict number of signals (0, 1, 2 ...)
  - birth/death processes
- **Ultimately, more constraints needed**
  - nonperiodic signals
  - masked cues
  - ambiguous signals

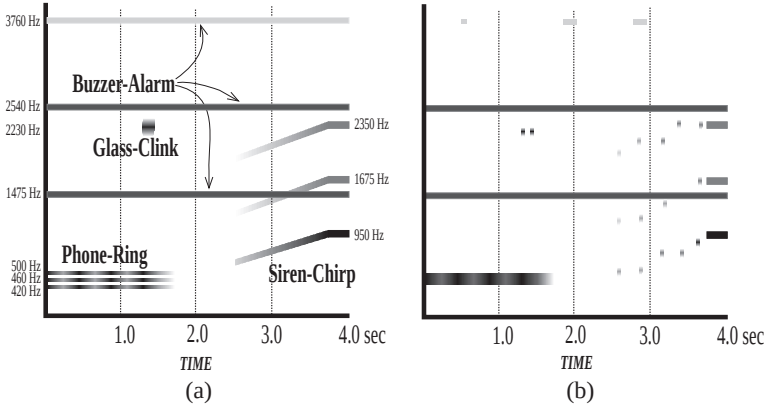
CASA1: Periodic pieces

- **Weintraub 1985**
  - separate male & female voices
  - find periodicities in each frequency channel by auto-coincidence
  - number of voices is ‘hidden state’
- **Cooke & Brown (1991-3)**
  - divide time-frequency plane into elements
  - apply grouping rules to form sources
  - pull single periodic target out of noise



CASA2: Hypothesis systems

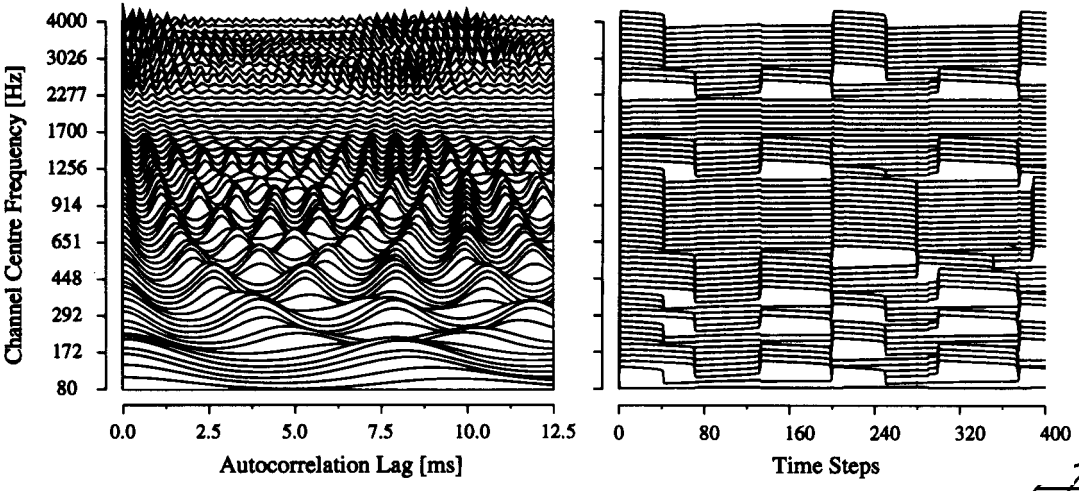
- **Okuno et al. (1994-)**
  - 'tracers' follow each harmonic + noise 'agent'
  - residue-driven: account for whole signal
- **Klassner 1996**
  - search for a combination of templates
  - high-level hypotheses permit front-end tuning



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

CASA3: Other approaches

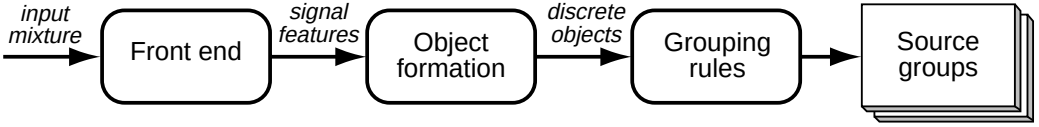
- **Blind source separation (Bell & Sejnowski)**
  - find exact separation parameters by maximizing statistic e.g. signal independence
- **HMM decomposition (RK Moore)**
  - recover combined source states directly
- **Neural models (Malsburg, Wang & Brown)**
  - avoid implausible AI methods (search,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



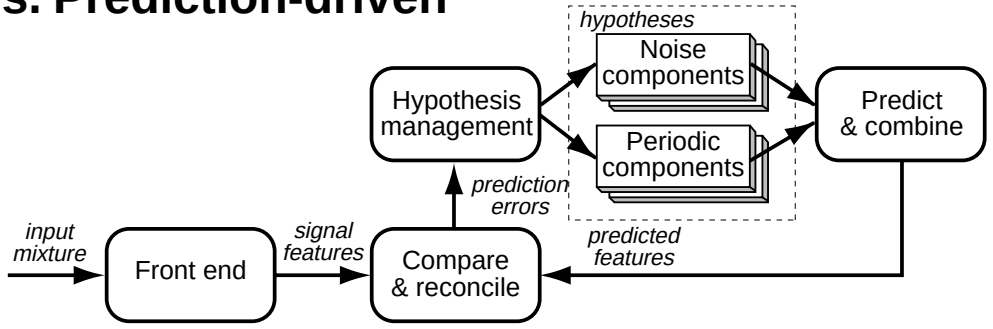
3 Prediction-driven CASA

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

- **Data-driven...**



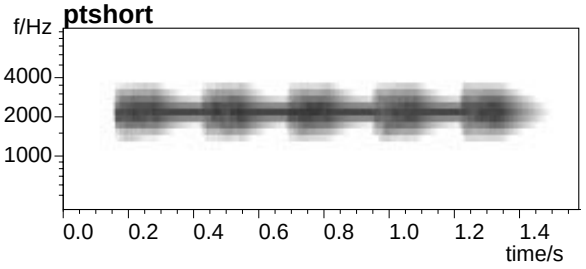
vs. Prediction-driven



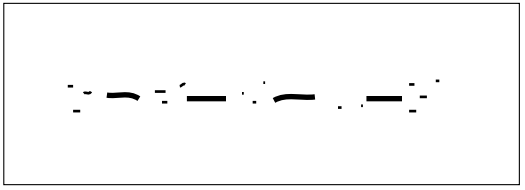
- **Motivations**
  - detect non-tonal events (noise & clicks)
  - support ‘restoration illusions’...
    $\rightarrow$ hooks for high-level knowledge
  + ‘complete explanation’, multiple hypotheses, resynthesis

Analyzing the continuity illusion

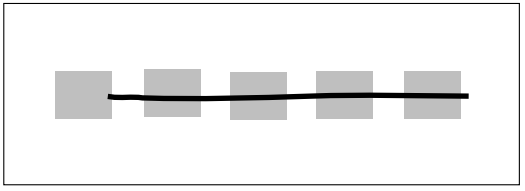
- **Interrupted tone heard as continuous**
  - .. if the interruption could be a masker



- **Data-driven just sees gaps**



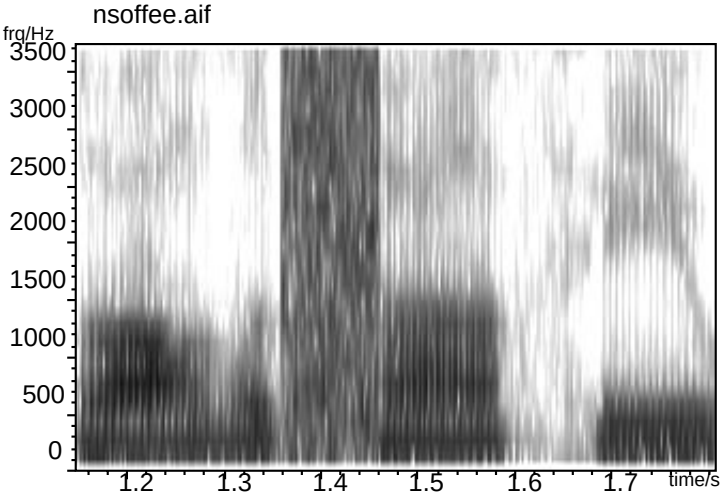
- **Prediction-driven can accommodate**



  - special case or general principle?

Phonemic Restoration (Warren 1970)

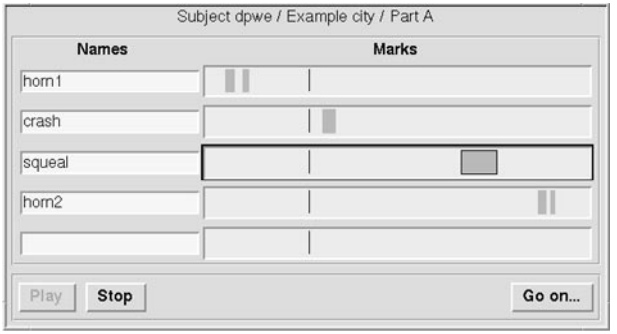
- **Another ‘illusion’ instance**
- **Inference relies on high-level semantics**



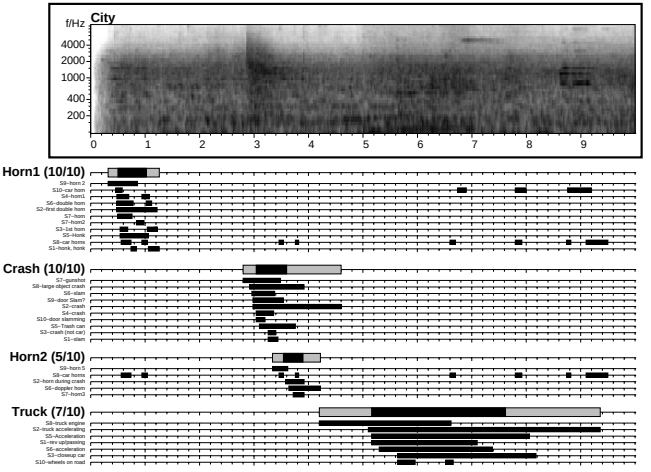
- **Incorporating knowledge into models?**

Subjective ground-truth in mixtures?

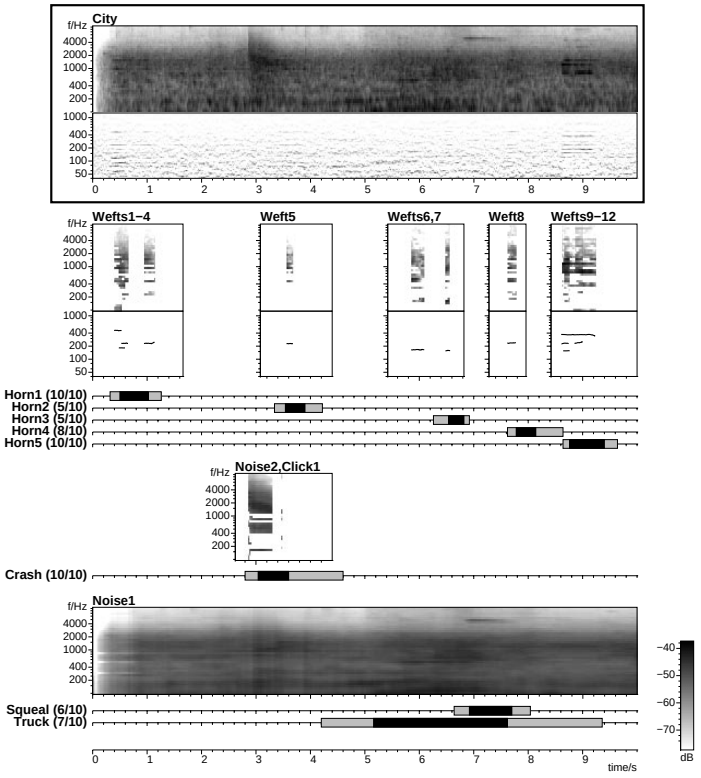
- **Listening tests collect ‘perceived events’:**



- **Consistent answers:**



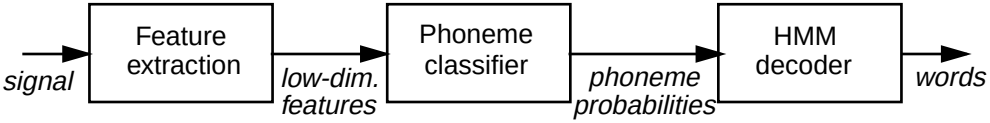
PDCASA example: City-street ambience



- **Problems**
  - error allocation
  - source hierarchy
  - rating hypotheses
  - resynthesis

4 Speech recognition & sound mixtures

- **Conventional speech recognition:**



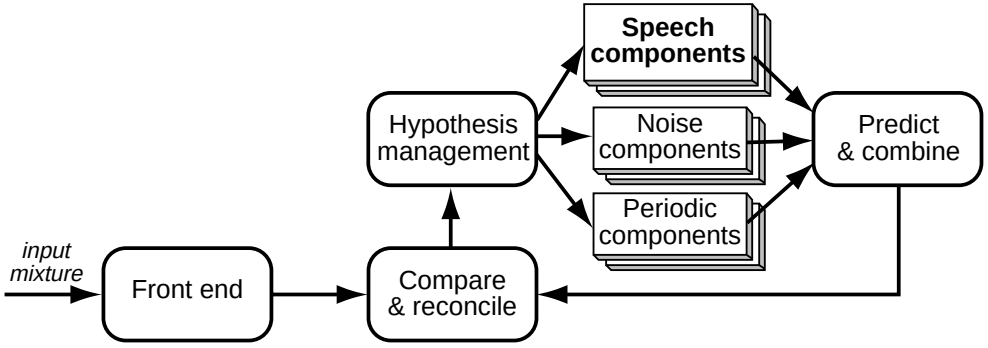
  - signal assumed entirely speech
  - find valid labelling by discrete labels
  - class models from training data

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

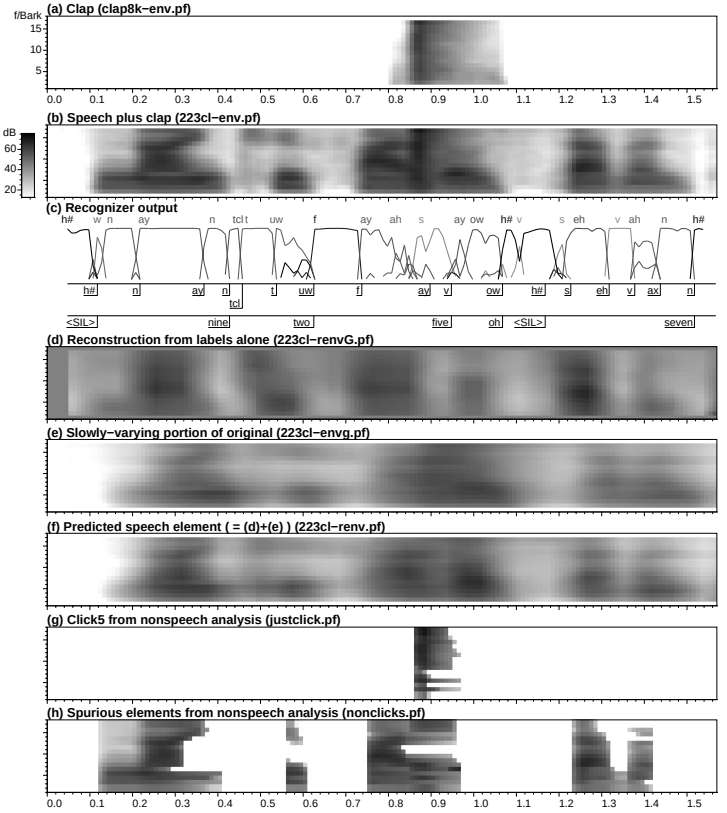
- **Very fragile to nonspeech, background**
  - scene-analysis methods very attractive...

CASA for speech recognition

- **Data-driven: CASA as preprocessor**
  - problems with ‘holes’ (but: Cooke, Okuno)
  - doesn’t exploit knowledge of speech structure
- **Prediction-driven: speech as component**
  - same ‘reconciliation’ of speech hypotheses
  - need to express ‘predictions’ in signal domain

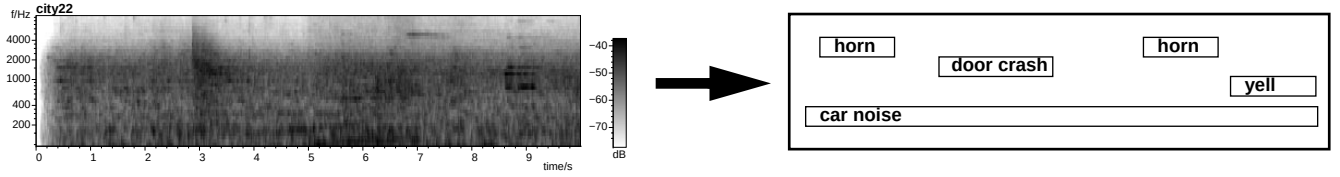


Example of speech & nonspeech



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

5 Implications for content analysis: Using CASA to index soundtracks



- **What are the 'objects' in a soundtrack?**
  - subjective definition → need auditory model
- **Segmentation vs. classification**
  - low-level cues → locate events
  - higher-level 'learned' knowledge to give semantic label (footstep, crash)

  ... AI complete?
- **But: hard to separate**
  - illusion phenomena suggest auditory *organization* depends on *interpretation*

Using speech recognition for indexing

- **Active research area:**
  Access to news broadcast databases
  - e.g. Informedia (CMU), ThisL (BBC+...)
  - use LV-CSR to transcribe,
    then text-retrieval to find
  - 30-40% word error rate, still works OK
- **Several systems at NIST TREC workshop**
- **Tricks to 'ignore' nonspeech/poor speech**

Open issues in automatic indexing

- **How to do ASA?**
- **Explanation/description hierarchy**
  - PDCASA: ‘generic’ primitives
    + constraining hierarchy
  - subjective & task-dependent
- **Classification**
  - connecting subjective & objective properties
    $\rightarrow$ finding subjective invariants, prominence
  - representation of sound-object ‘classes’
- **Resynthesis?**
  - a ‘good’ description should be adequate
  - provided in PDCASA, but low quality
  - requires good knowledge-based constraints

6 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 recognizers are a good source of knowledge**
- **Automatic indexing implies 'synthetic listener'**
  - need to solve a lot of modeling issues