Sound content analysis for indexing and understanding

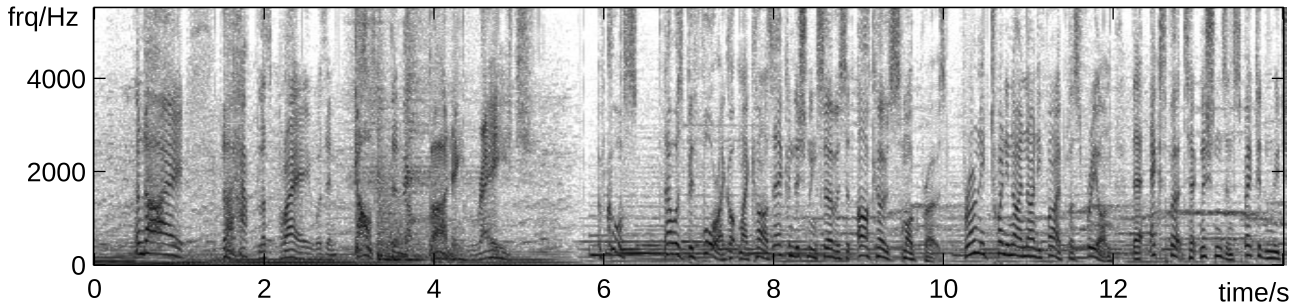
Dan Ellis
International Computer Science Institute, Berkeley CA
<dpwe@icsi.berkeley.edu>

Outline

1. **Sound content analysis**
2. **Speech recognition**
3. **Auditory scene analysis**
4. **Audio content indexing**
5. **Conclusions**

1 Sound content analysis

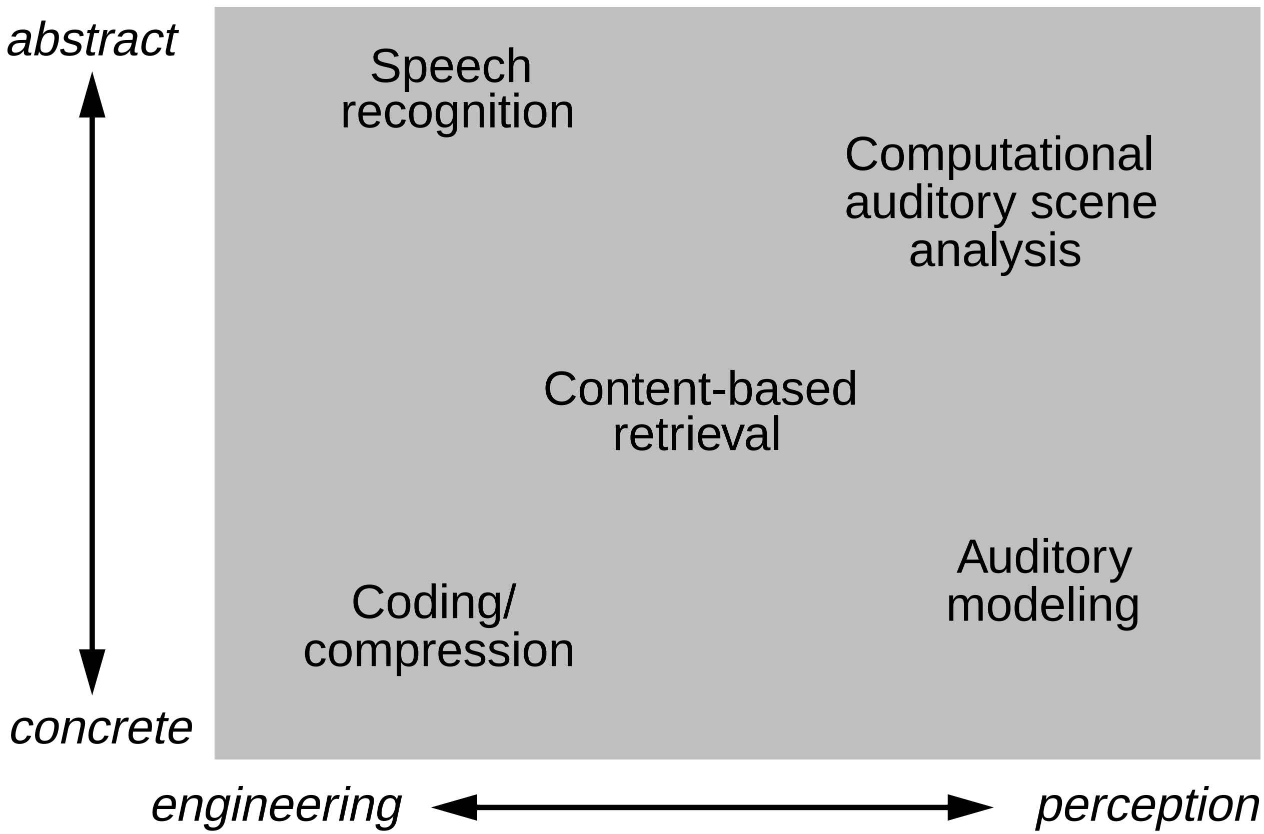
- **Overall goal: ‘Useful’ data from sound**



  - which depends on the goal

- **Involving:**
  - continuous $\rightarrow$ discrete
  - source separation
  - extract ‘semantic’ content < words / actions/events

The space of sound analysis research



Outline

1. **Sound content analysis**
2. **Speech recognition**
   - Classic speech recognition
   - The connectionist-HMM hybrid
   - Strength through combinations
3. **Auditory scene analysis**
4. **Audio content indexing**
5. **Conclusions**

Speech recognition: Dictation

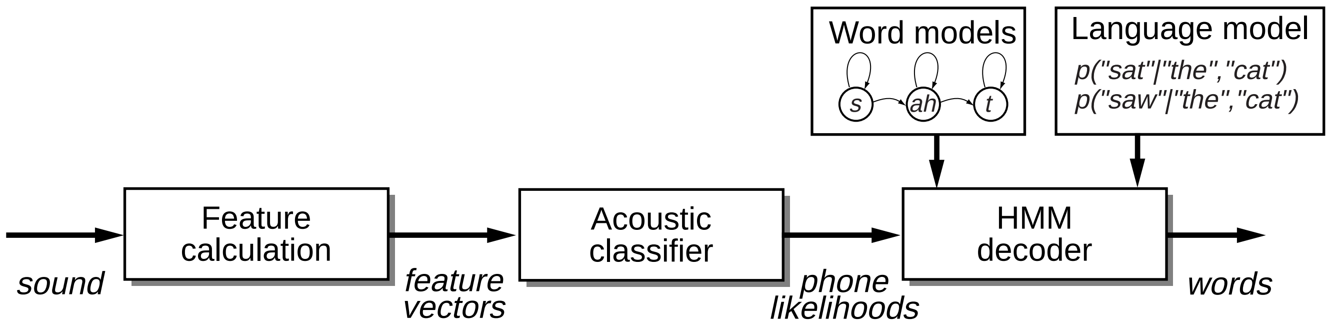
- **Observations $X = \{X_1..X_N\} \rightarrow$ States $S = \{S_1..S_N\}$**

$$
\begin{aligned}
S^* &= \underset{S}{\operatorname{argmax}}\ P(S|X) \\
&= \underset{S}{\operatorname{argmax}}\ \frac{P(S,X)}{P(X)} \\
&= \underset{S}{\operatorname{argmax}}\ \prod_i P(X_i|S_i) \cdot P(S_i|S_{i-1})
\end{aligned}
$$

Markov assumption

acoustic prob. **transition prob.**

- **State sequence $\{S_i\}$ (e.g. phones) define words**

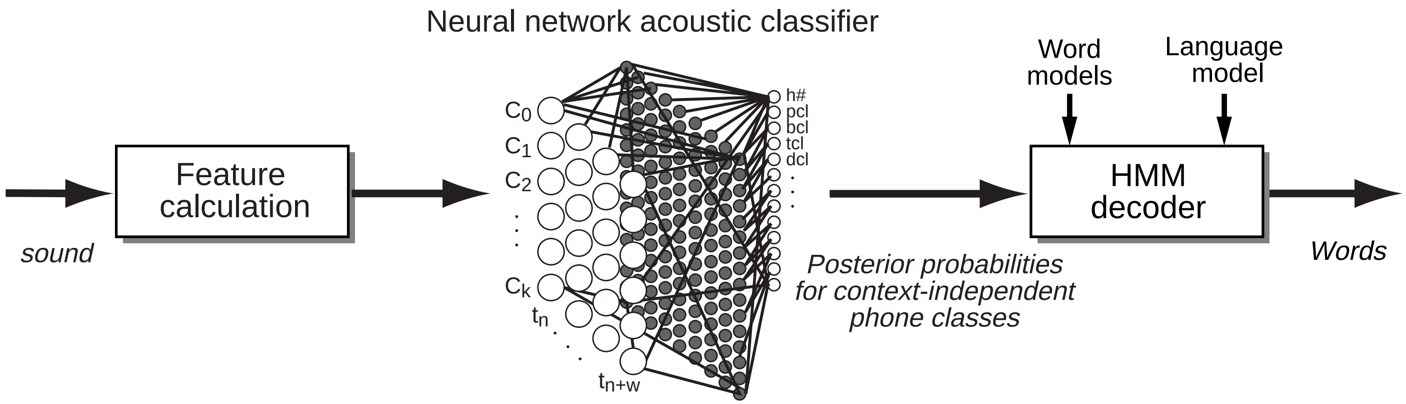


- **Training (on large datasets) is the key**
  - EM iteration for acoustic & transition probs.

The connectionist-HMM hybrid

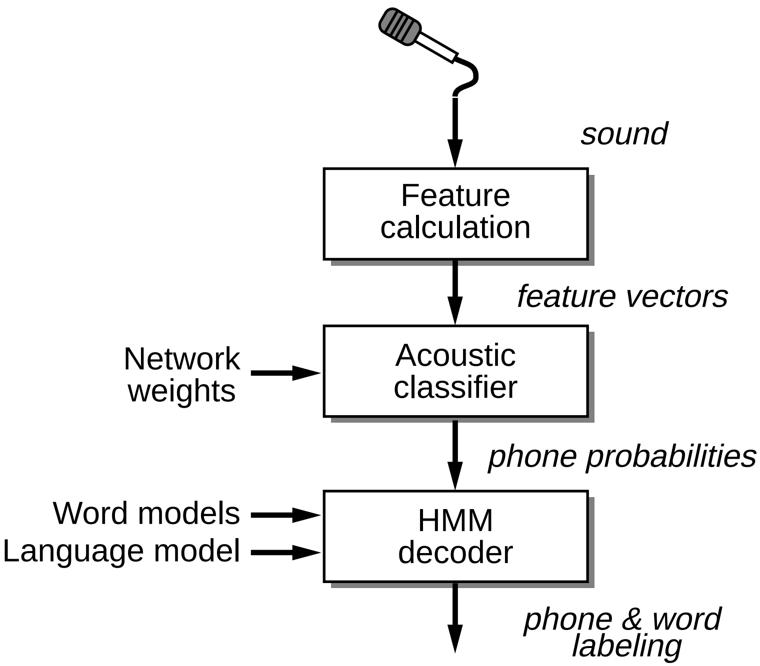
(Morgan & Bourlard, 1995)

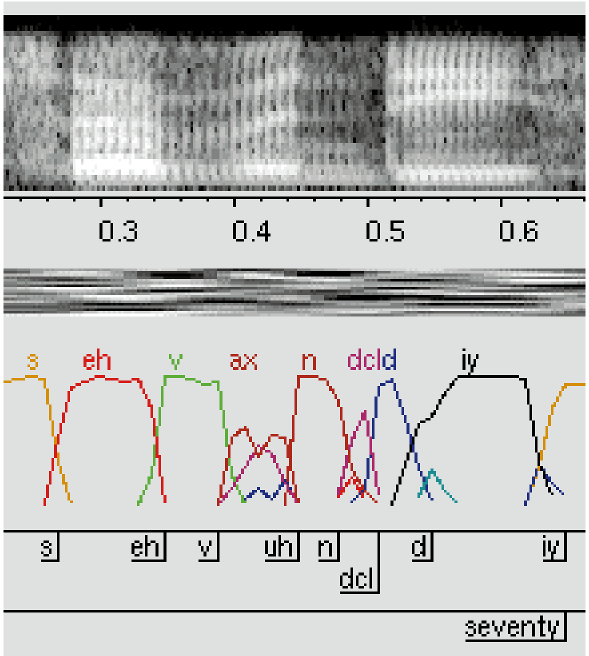
- ***P*(*X*$_i$|*S*$_i$) is acoustic *likelihood* model**
  - model distribution with, e.g., Gaussian mixtures
- **Replace with *posterior*, *P*(*S*$_i$|*X*$_i$):**



  - neural network estimates phone given acoustics
  - discriminative
- **Simpler structure for research**

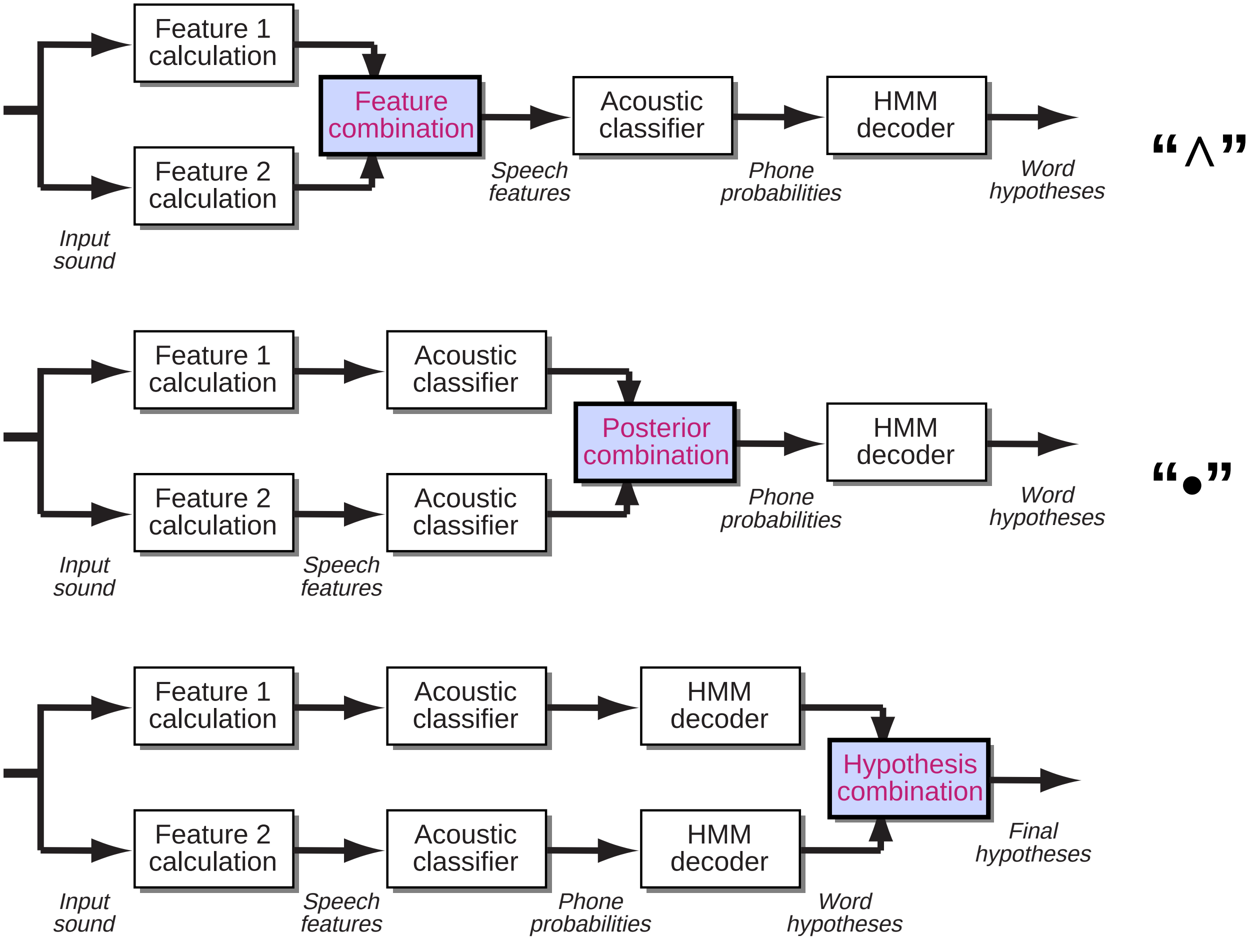
Visualizing speech recognition





Combination schemes

- **How to use complementary features?**



Combining feature streams

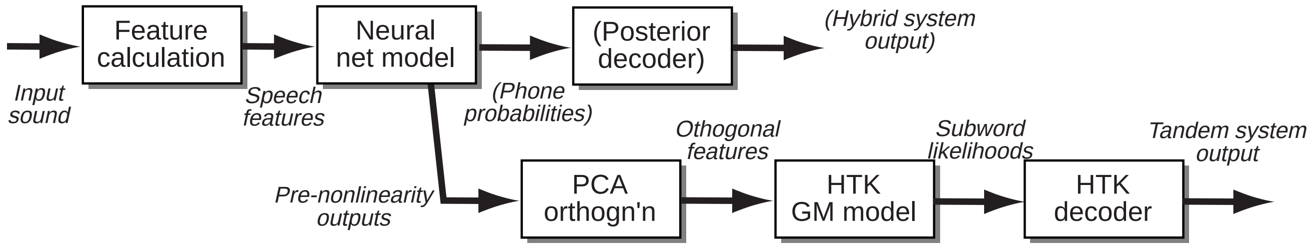
- **How to allocate feature dimensions to models?**
  - lower-dimension models train more quickly
  - higher-dimension models find more interactions
- **PLP & MSG for Aurora (digits in noise):**
  - PLP are ‘conventional’ features
  - MSG developed as robust alternative
  - Evaluate by word-error rate (WER) compared to default baseline

Features	*Parameters*	*baseline WER ratio*
plp12•dplp12	136k	97.6%
plp12^dplp12	124k	89.6%
msg3a•msg3b	145k	101.1%
msg3a^msg3b	133k	85.8%
plp12•dplp12•msg3a•msg3b	281k	76.5%
plp12^dplp12^msg3a^msg3b	245k	74.1%
plp12^dplp12•msg3a^msg3b	257k	63.0%

Tandem connectionist models

(with Hermansky et al., OGI)

- **How can we combine neural net & GM models?**



  - (GMM system does not know they are phones)

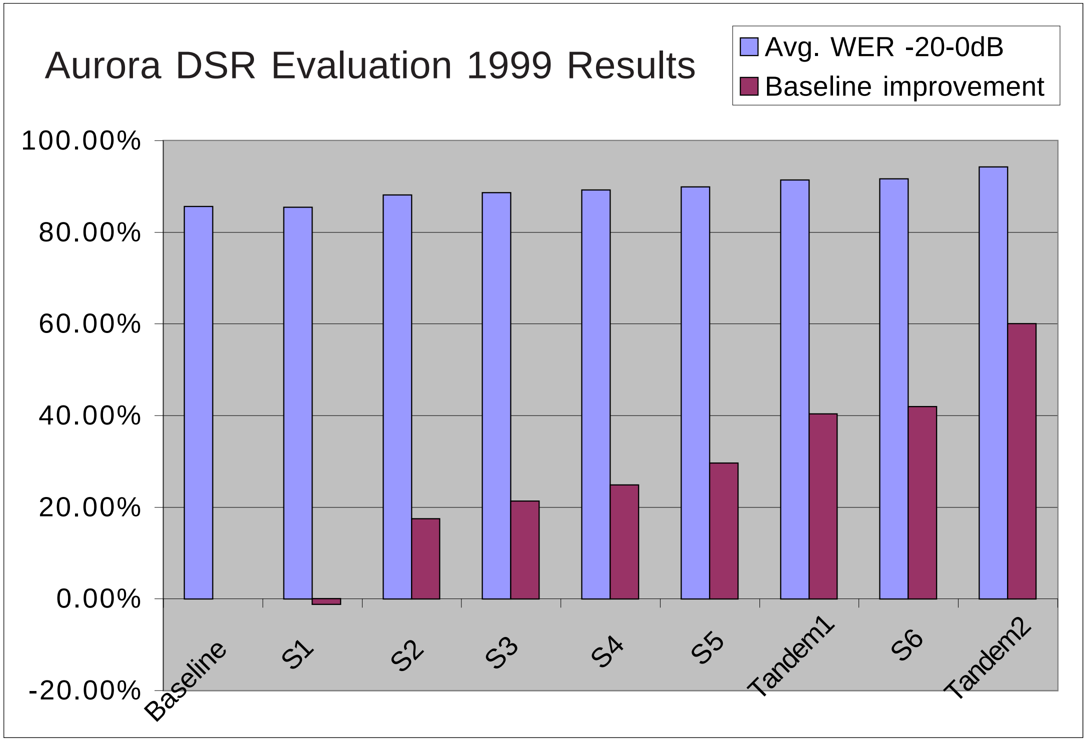
- **Result: better performance than either alone!**
  - neural net has trained discriminatively
  - GMM HMMs learn context-dependent structure

  →extract complementary info from training data

System-features	*baseline WER ratio*
HTK-mfcc	100.0%
Hybrid-mfcc	84.6%
Tandem-mfcc	64.5%
Tandem-plp+msg	47.2%

Aurora "Distributed SR" evaluation

- **7 telecoms company submissions:**



  - Tandem systems from OGI-ICSI-Qualcomm

Outstanding issues in speech recognition

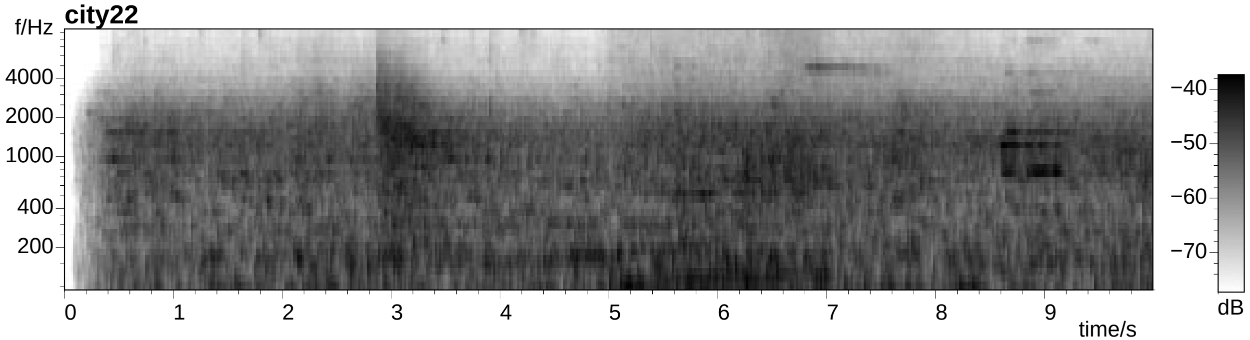
- **Are we on the right path?**
  - useful dictation products exist
  - evaluation results improve every year
  - .. but appear to be asymptoting
- **Is dictation enough?**
  - a useful focus initially
  - .. but not speech *understanding*
  - .. and has skewed research
- **What should be our research priorities?**
  - straight ASR research is hard to fund
  - need to look at harder domains
  - need to connect it to applications

Outline

1. Sound content analysis
2. Speech recognition
3. **Auditory scene analysis**
   - Psychological phenomena
   - Computational modeling
   - Prediction-driven analysis
   - Incorporating speech
4. Audio content indexing
5. Conclusions

3 Auditory Scene Analysis (ASA)

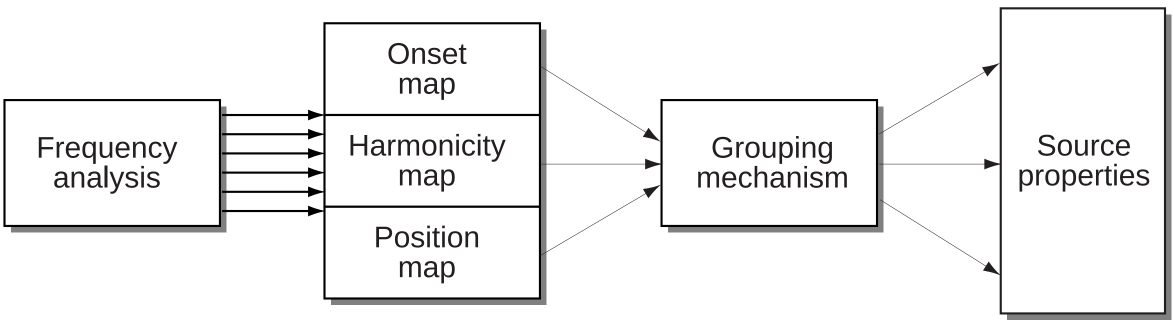
"The organization of sound scenes according to their inferred sources"



- **Sounds rarely occur in isolation**
  - need to 'separate' for useful information
- **Human audition is very effective**
  - computational models have a lot to learn

Psychology of ASA

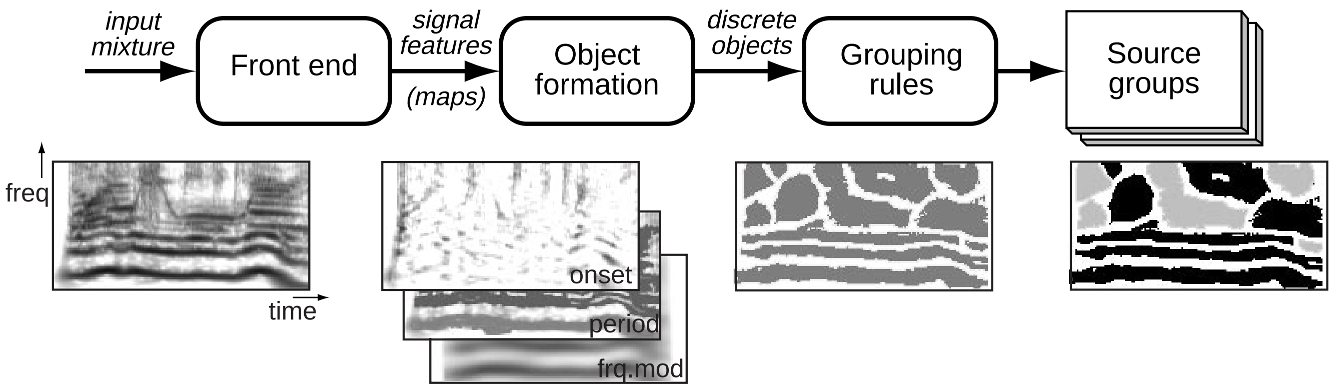
- **Extensive experimental research**
  - perception of simplified stimuli (sinusoids, noise)
- **“Auditory Scene Analysis” [Bregman 1990]**
  - first: break mixture into small *elements*
  - elements are *grouped* in to sources using *cues*
- **Grouping ‘rules’ (Darwin, Carlyon, ...):**
  - common onset/offset/modulation,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



(after Darwin, 1996)

Computational Auditory Scene Analysis (CASA)

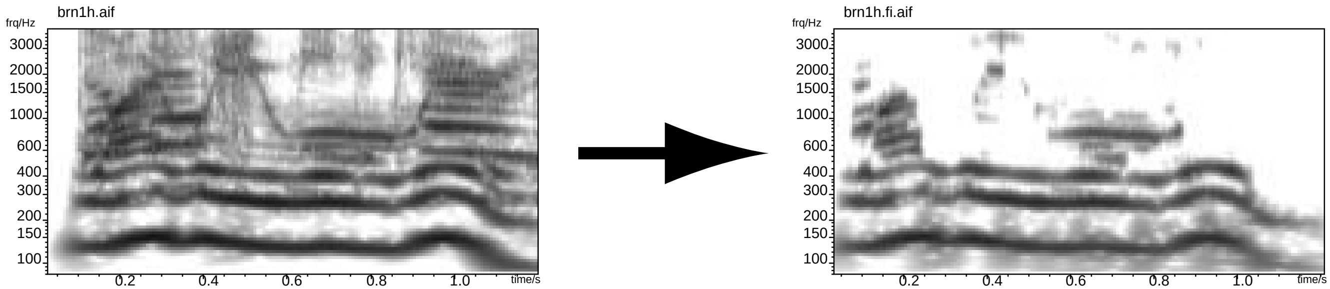
- **Literal model of Bregman... (e.g. Brown 1992):**



- **Goals**
  - identify and segregate different sources
  - resynthesize separate outputs!

Grouping model results

- **Able to extract voiced speech:**



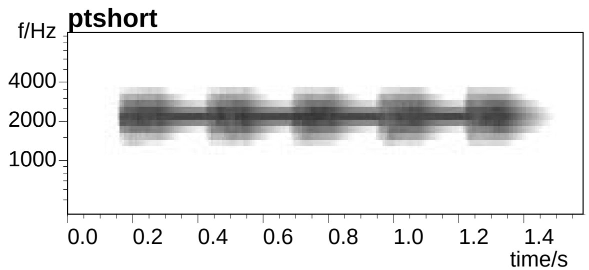
- **Limitations**
  - resynthesis via filter-mask
  - *only* periodic targets
  - robustness of discrete objects

Context, expectations & predictions

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

- **Bregman's "old-plus-new" principle:**
  A change in a signal will be interpreted as an *added* source whenever possible

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

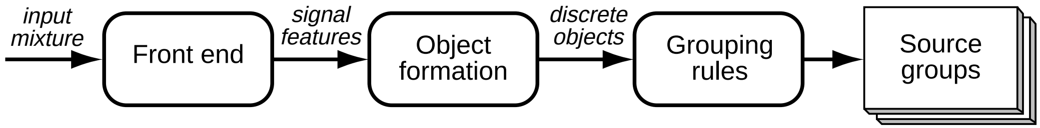


  - tones alternates with noise bursts
  - noise is strong enough to mask tone
    ... so listener discriminate presence
  - continuous tone perceived for gaps ~100s of ms

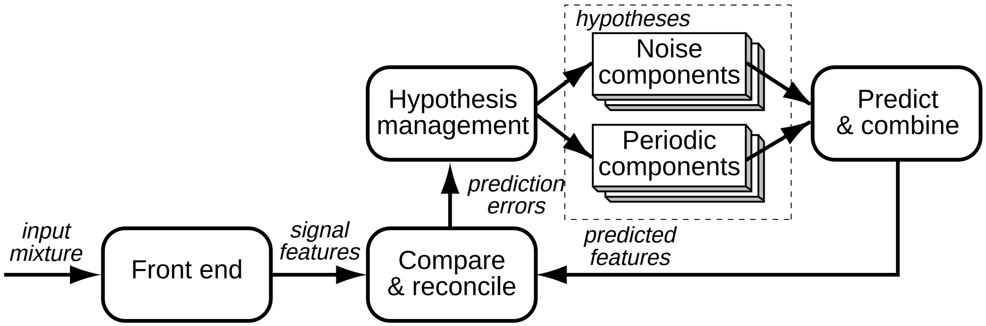
→ **Inference acts at low, preconscious level**

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

- **Data-driven...**



vs. Prediction-driven

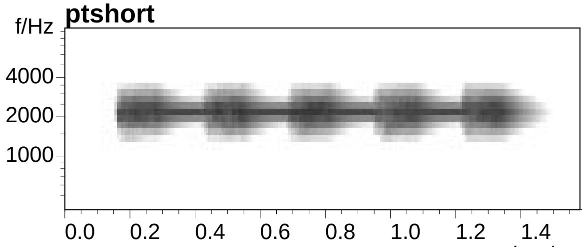


- **PDCASA key features:**
  - 'complete explanation' of all scene energy
  - vocabulary of periodic/noise/transient elements
  - multiple hypotheses
  - explanation hierarchy

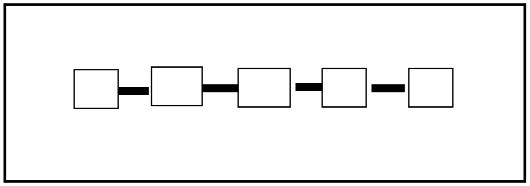
PDCASA for the continuity illusion

- **Subjects hear the tone as continuous**

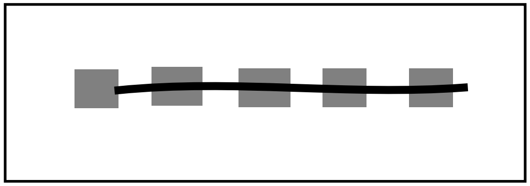
  ... if the noise is a plausible masker



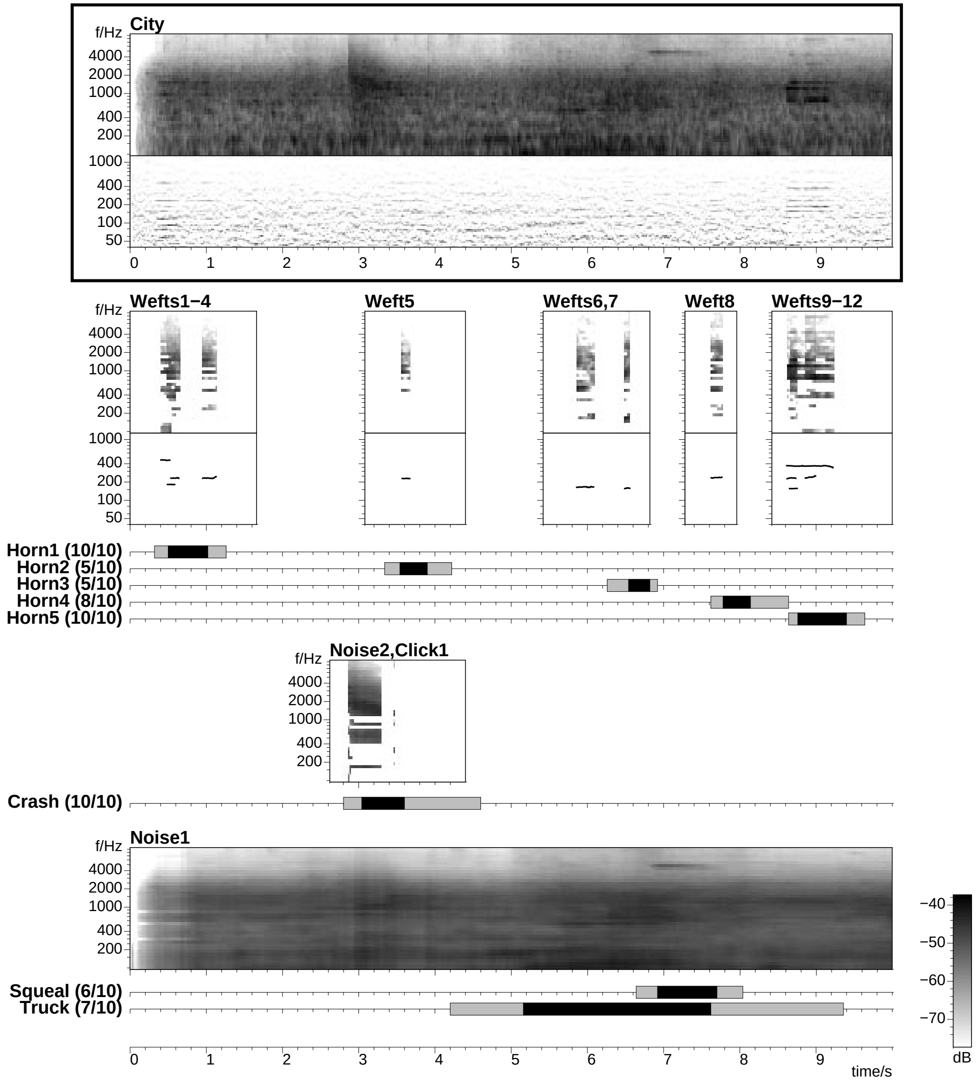
- **Data-driven analysis gives just visible portions:**



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

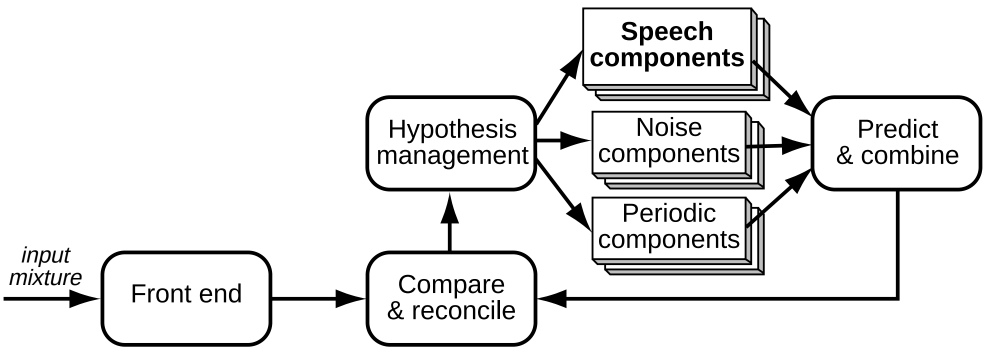


PDCASA analysis of a complex scene



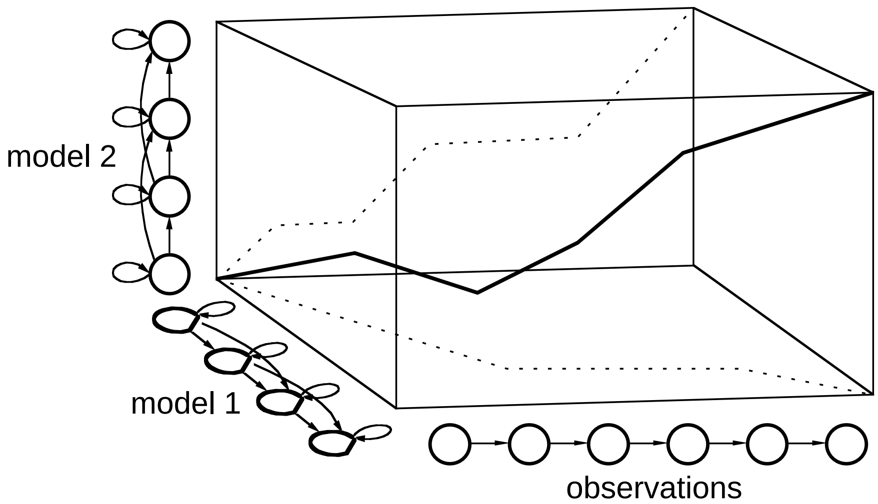
CASA for speech recognition

- **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
  - RESPITE project: modifications to recognizer
- **Prediction-driven: speech as component**
  - same ‘reconciliation’ of speech hypotheses
  - need to express ‘predictions’ in signal domain



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

Outstanding issues in CASA

- **What is the architecture?**
  - data-driven versus prediction-driven
  - representations at different levels
  - hypothesis search
- **How to combine different cues?**
  - priority of different cues
  - resolving conflicting cues
  - bottom-up versus top-down
- **How to exploit training data?**
  - .. the big lesson from speech recognition
- **Evaluation**
  - .. a more subtle lesson

Outline

1. Sound content analysis
2. Speech recognition
3. Auditory scene analysis
4. **Audio content indexing**
    - Spoken document retrieval
    - Handling nonspeech audio
    - Object-based analysis and retrieval
    - Audio-video content organization
5. Conclusions

4

Audio content indexing: Spoken document retrieval (SDR)

- **Idea: speech recognition transcripts as indexes**
- **Best broadcast news systems are not great**
  - 15-30% WER on real broadcasts
- **Word errors vary in their impact:**

F0: THE VERY EARLY RETURNS OF THE NICARAGUAN PRESIDENTIAL ELECTION SEEMED TO FADE BEFORE THE LOCAL MAYOR ON A LOT OF LAW

F4: AT THIS STAGE OF THE ACCOUNTING FOR SEVENTY SCOTCH ONE LEADER DANIEL ORTEGA IS IN SECOND PLACE THERE WERE TWENTY THREE PRESIDENTIAL CANDIDATES OF THE ELECTION

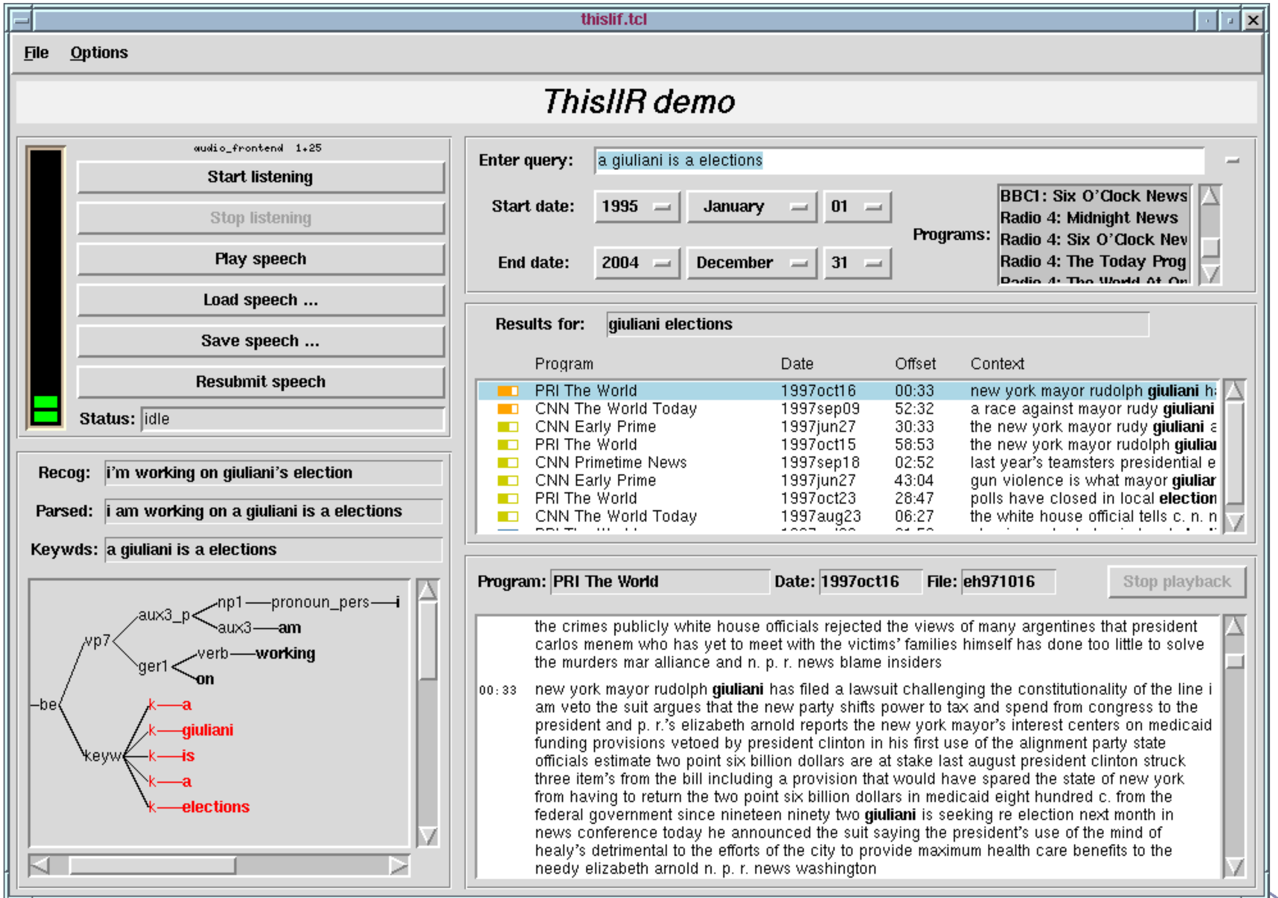
F5: THE LABOR MIGHT DO WELL TO REMEMBER THE LOST A MAJOR EPISODE OF TRANSATLANTIC CONNECT TO A CORPORATION IN BOTH CONSERVATIVE PARTY OFFICIALS FROM BRITAIN GOING TO WASHINGTON THEY WENT TO WOOD BUYS GEORGE BUSH ON HOW TO WIN A SECOND TO NONE IN LONDON THIS IS STEPHEN BEARD FOR MARKETPLACE

- **Good enough for information retrieval (IR)**
  - e.g. TREC-8 average precision:
    reference transcript ~ 0.5
    30% WER ~ 0.4

Thematic Indexing of Spoken Language

(with Sheffield, Cambridge, BBC)

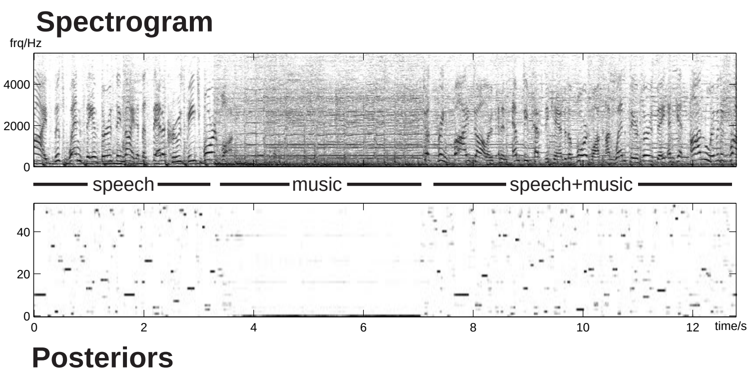
- **SDR for BBC broadcast news archive**
  - 1000+ hr archive, automatically updated

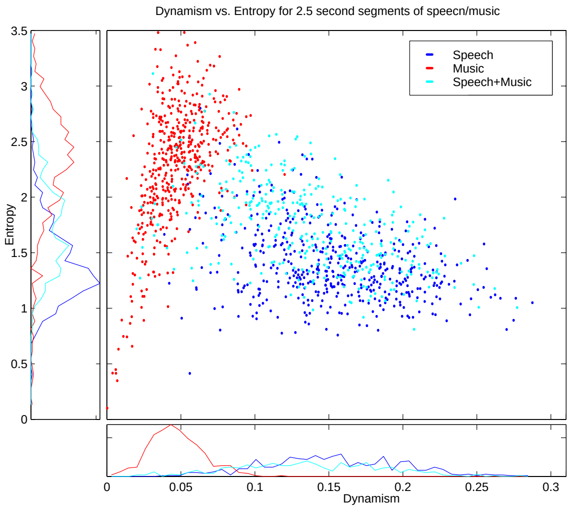


Speech and nonspeech

(with Gethin Williams)

- **ASR run over entire soundtracks?**
  - for nonspeech, result is nonsense
- **Watch behavior of speech acoustic model:**
  - average per-frame entropy
  - 'dynamism' - mean-squared 1st-order difference

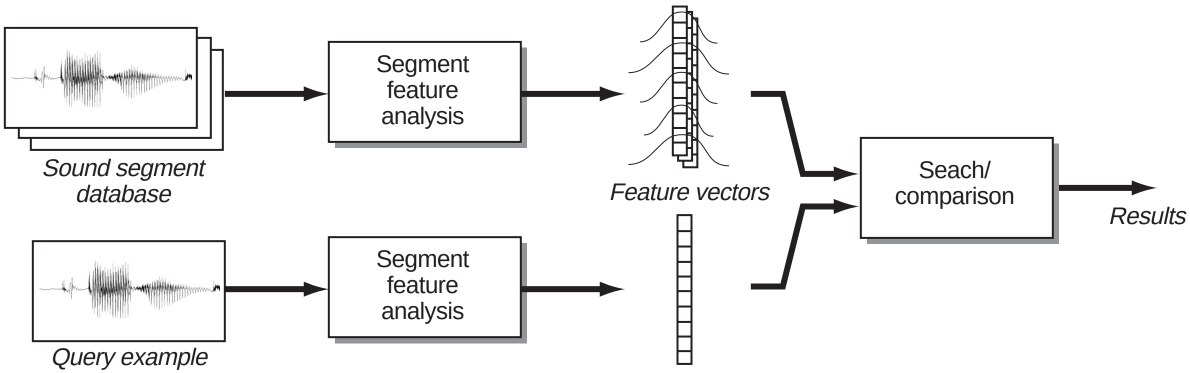




- **1.3% error on 2.5 second speech-music testset**

Element-based audio indexing

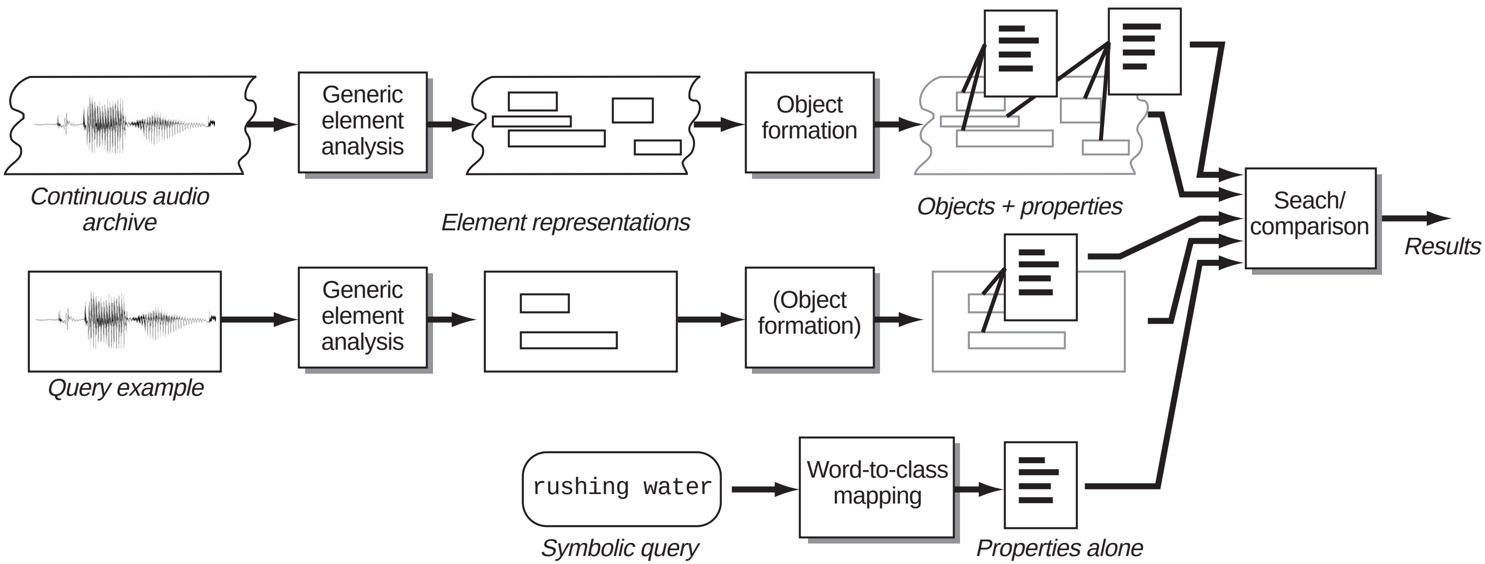
- **Search for nonspeech audio databases**
  - e.g. Muscle Fish 'SoundFisher' for SFX libraries
- **Segment-level features**



  - well-performing features:
    spectral centroid, dynamics, tonality ...
- **Each segment is an object**
  - not applicable to continuous recordings

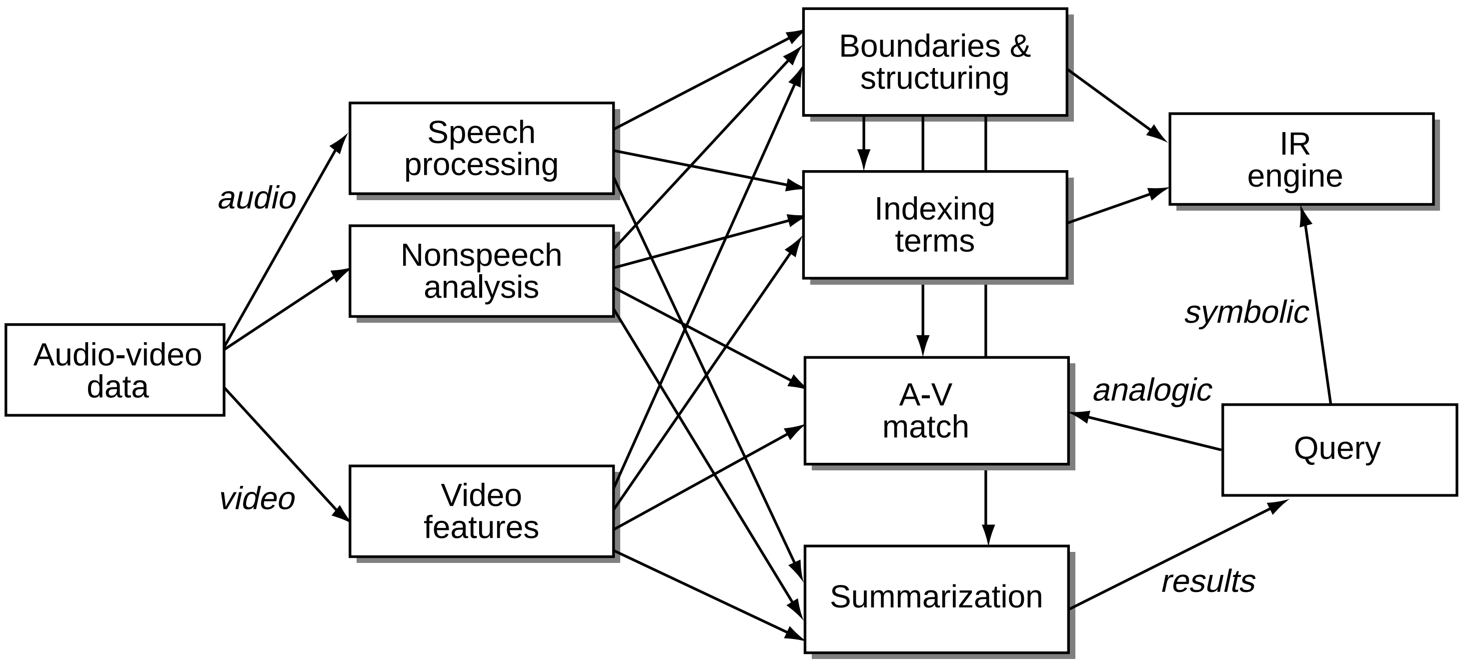
Object-based audio indexing

- **Using 'generic sound elements'**
  - decompose sound into elements; match subsets
  - how to generalize?
  - how to use segment-style features?
- **Form into objects for higher-order properties**
  - CASA-type object formation (onset, harmonicity)



Audio-video organization & retrieval

- **How it might work...**



AV indexing components

- **Recovering broad temporal structure**
  - speaker turns ; speech & music ; repetition
  - characteristic of genres e.g. news shows
  - indexible attributes in themselves
- **Posing queries:**
  - term-based
  - proximity to examples
  - dynamic audio-visual sketches?
- **How to define index/query terms?**
  - different kinds of terms: literal versus thematic
  - machine learning of event classes
- **Summarization**
  - for displaying ‘hits’: impacts usability
  - text / image / video / sound
  - tricks e.g. to find most salient words

Open issues in audio indexing

- **Information from speech**
  - multiple, confidence-tagged results? (not WER)
  - prosodics; emphasis; speaking style
  - speaker tracking, identity, character
- **Information from nonspeech**
  - how to define objects
  - how to match symbolic search terms
- **Integrating audio and video**
  - combining information for search elements
  - forms of query
- **Related applications**
  - ‘structured content’ encoders (e.g. MPEG4SA)
  - semantic hearing aids ; robot monitors

Outline

1. Sound content analysis
2. Speech recognition
3. Auditory scene analysis
4. Audio content indexing
5. **Conclusions**

5

Conclusions: The state of sound content analysis

- **Speech recognition:**
  - focussed application, practical results
  - powerful statistical pattern recognition tools
  - able to exploit large training sets
- **Computational Auditory Scene Analysis:**
  - real-world sounds are mixtures
  - discover advanced ecological constraints
  - results still rather preliminary
- **Content-based retrieval:**
  - compelling problem; forgiving application
  - leveraging audio-visual correlations
  - fertile ground for research