Recognition & Organization of Speech and Audio

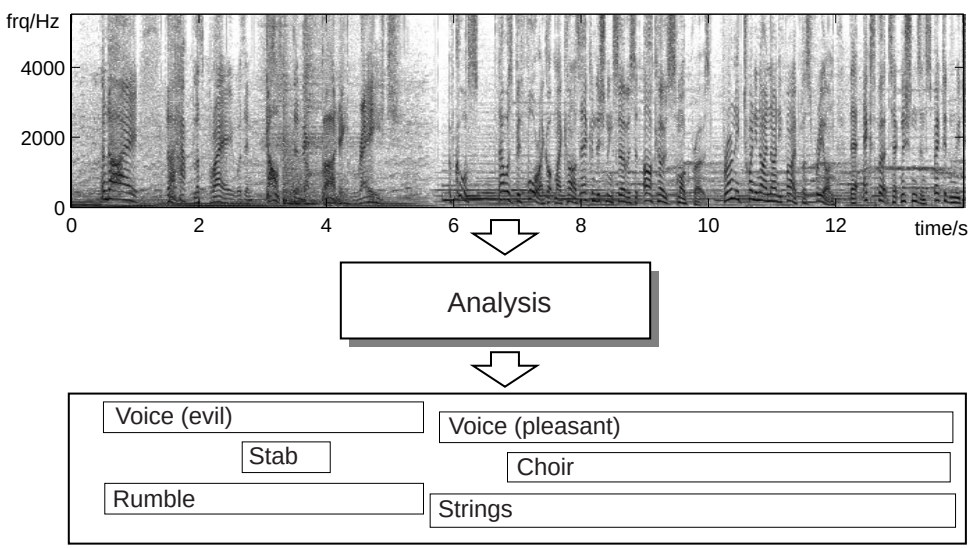
Dan Ellis
Electrical Engineering, Columbia University
<dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/

Outline

1. **Introducing LabROSA**
2. **Robust speech recognition**
3. **General audio analysis**
4. **Summary**

Lab ROSA

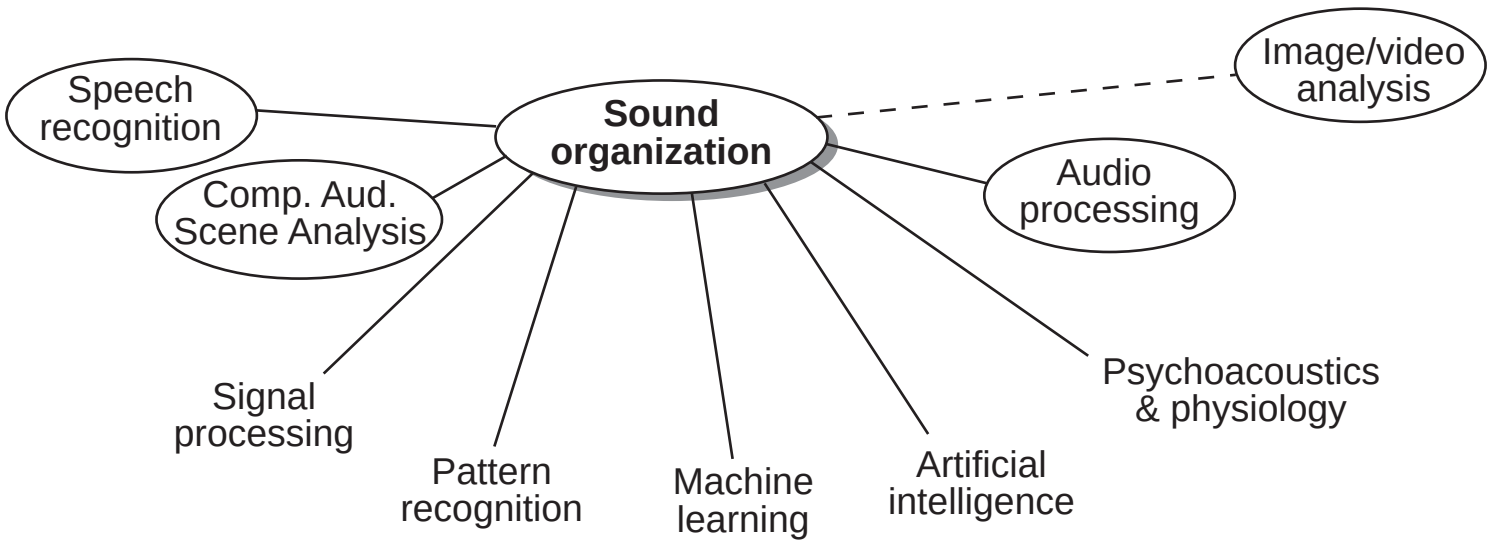
1 Organization of sound mixtures



- **Core operation:**

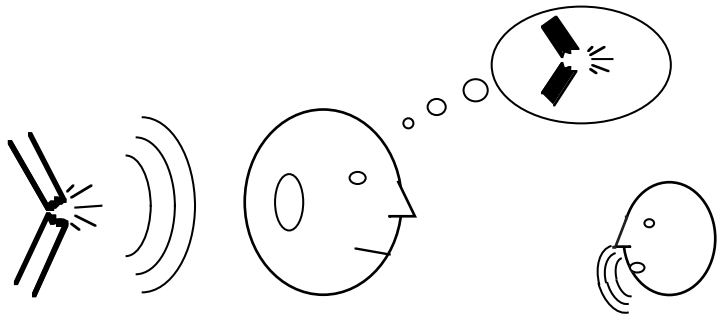
  Converting continuous, scalar signal into discrete, symbolic representation

Positioning sound organization



- **Draws on many techniques**
- **Abuts/overlaps various areas**

About auditory perception



- **Received waveform is a mixture**
  - two sensors, N signals ...
  - need knowledge-based constraints
- **Psychoacoustics:**
  the study of human sound organization
  - ‘auditory scene analysis’ (Bregman’90)
- **Auditory perception is ecologically grounded**
  - scene analysis is preconscious (→ illusions)
  - perceived organization:
    real-world objects + events (transient)
  - subjective *not* canonical (ambiguity)

Key themes for LabROSA

http://www.ee.columbia.edu/~dpwe/LabROSA/

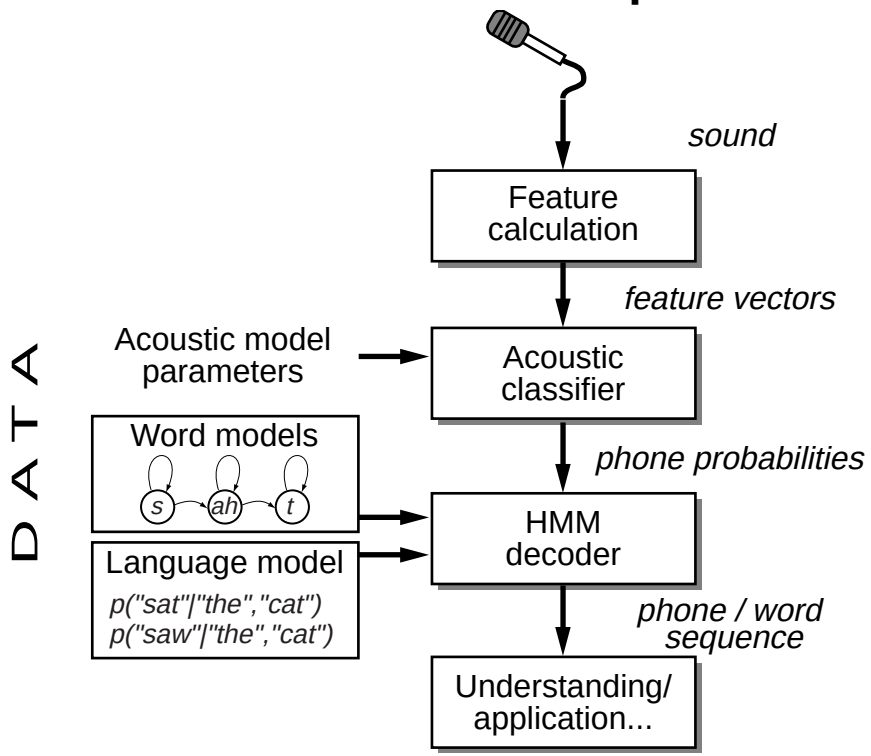
- **Sound organization: construct hierarchy**
  - at an instant (sources)
  - along time (segmentation)
- **Scene analysis**
  - find attributes according to objects
  - use attributes to form objects
  - ... plus constraints of knowledge
- **Exploiting large data sets (the ASR lesson)**
  - supervised/labeled: pattern recognition
  - unsupervised: structure discovery, clustering
- **Special cases:**
  - speech recognition
  - other source-specific recognizers
- **... within a 'complete explanation'**

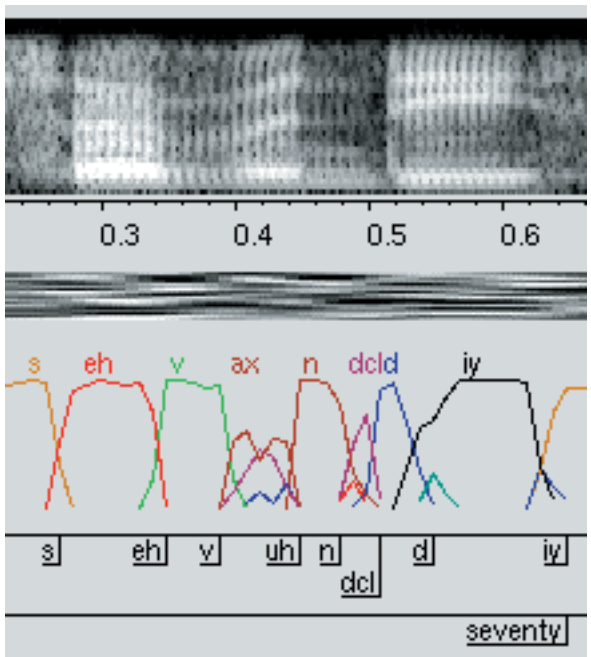
Outline

1. Introducing LabROSA
2. **Robust speech recognition**
   - ASR overview
   - Tandem modeling
   - Missing data and multisource decoding
3. General audio analysis
4. Summary

Automatic Speech Recognition (ASR)

- **Standard speech recognition structure:**



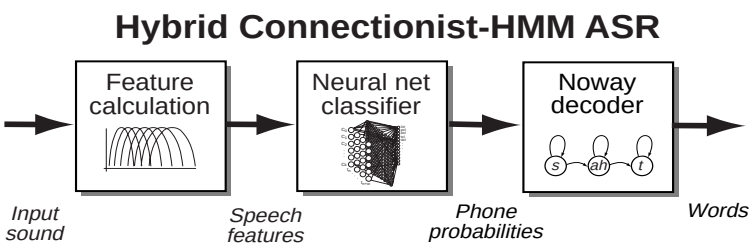


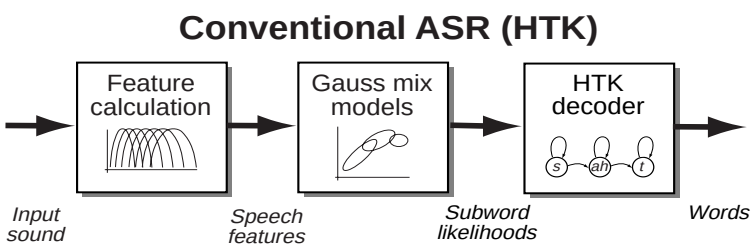
- **'State of the art' word-error rates (WERs):**
  - 2% (dictation) - 30% (telephone conversations)
- **Can use multiple streams...**

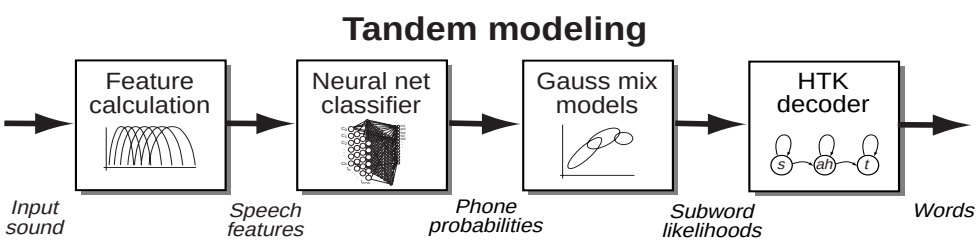
Tandem speech recognition

(with Hermansky, Sharma & Sivadas/OGI, Singh/CMU, ICSI)

- **Neural net estimates phone posteriors; but Gaussian mixtures model finer detail**
- **Combine them!**





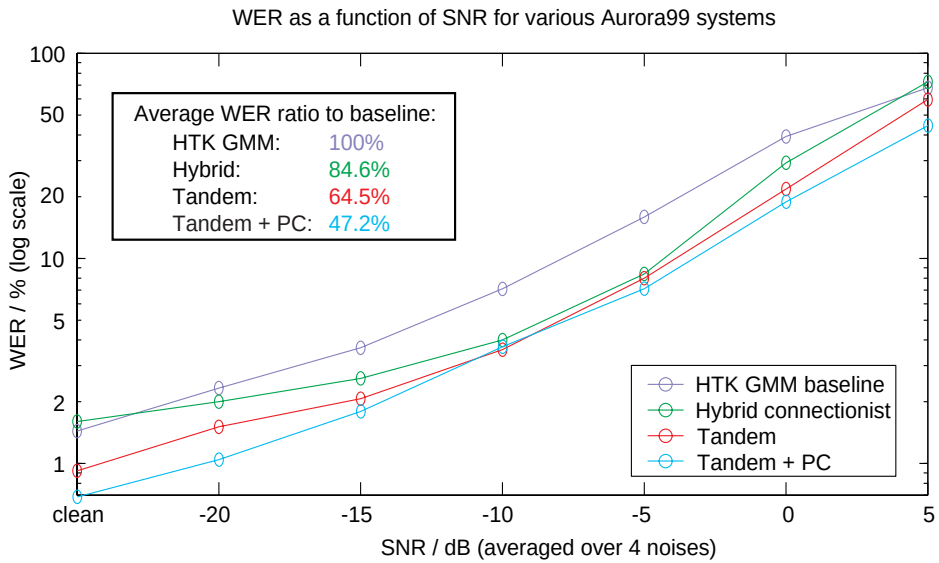


- **Train net, then train GMM on net output**
  - GMM is ignorant of net output ‘meaning’

Lab ROSA

Tandem system results

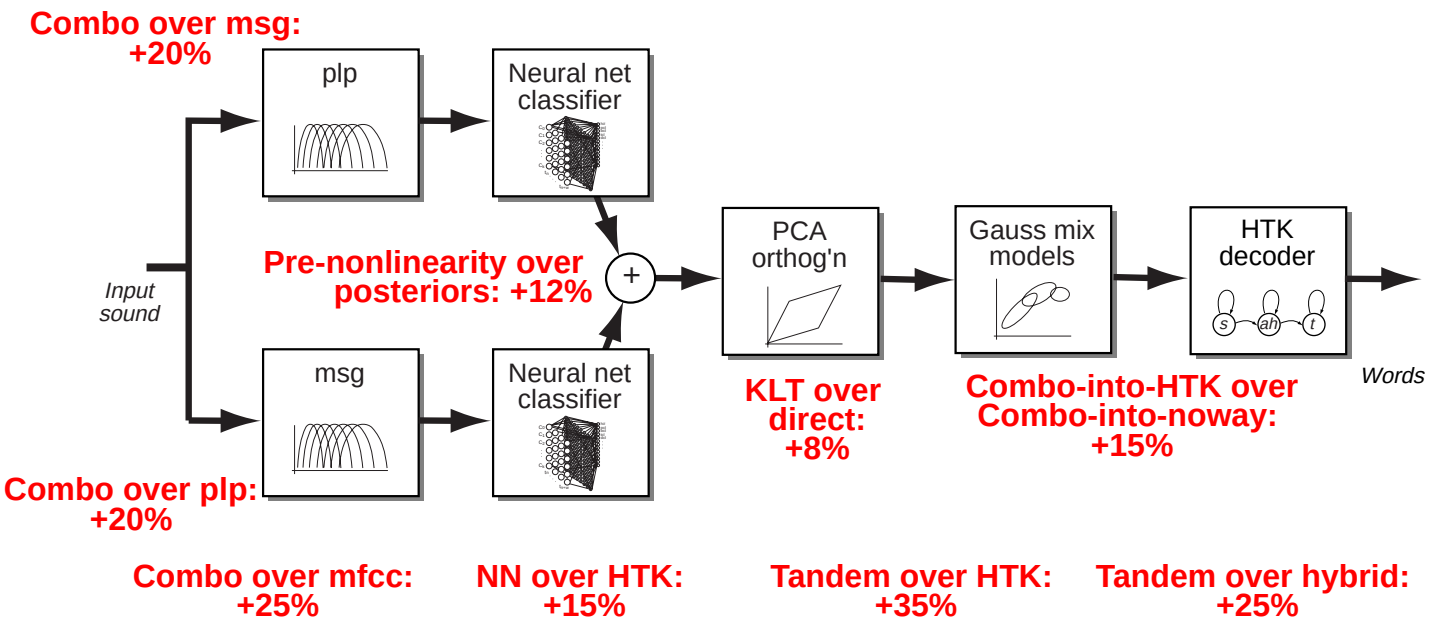
- **It works very well (‘Aurora’ noisy digits):**



System-features	*Avg. WER 20-0 dB*	*Baseline WER ratio*
HTK-mfcc	13.7%	100%
Neural net-mfcc	9.3%	84.5%
Tandem-mfcc	7.4%	64.5%
Tandem-msg+plp	**6.4%**	**47.2%**

Relative contributions

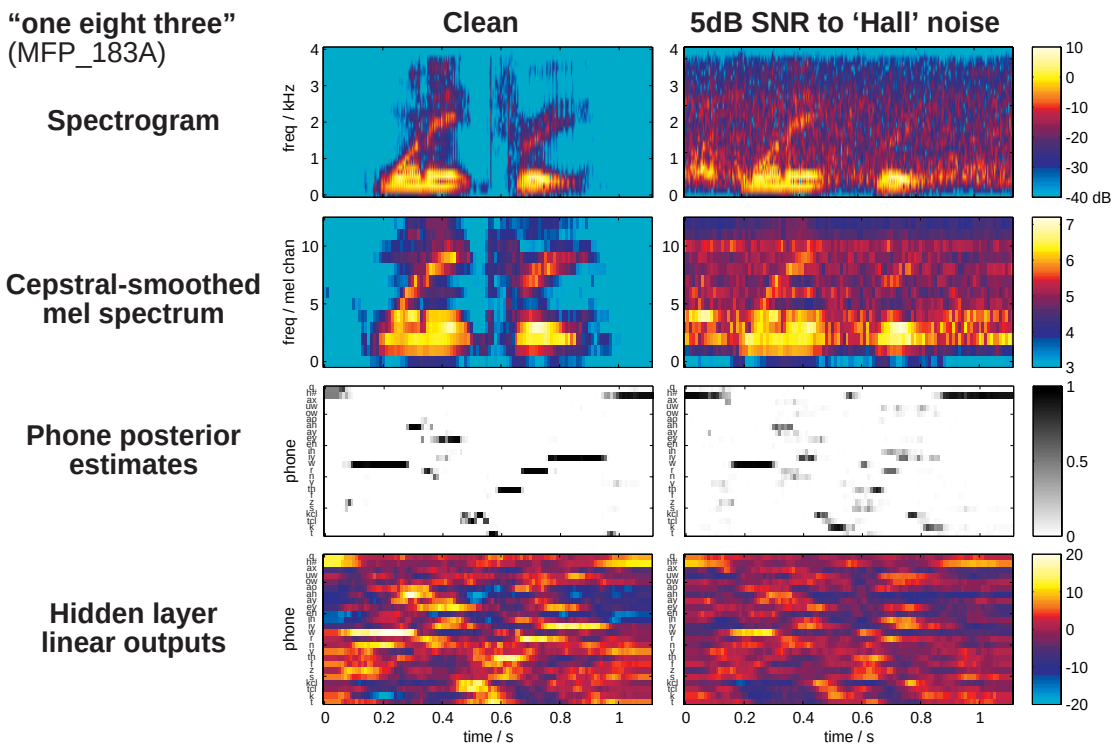
- **Approx relative impact on baseline WER ratio for different component:**



Tandem combo over HTK mfcc baseline: +53%

Inside Tandem systems: What's going on?

- **Visualizations of the net outputs**

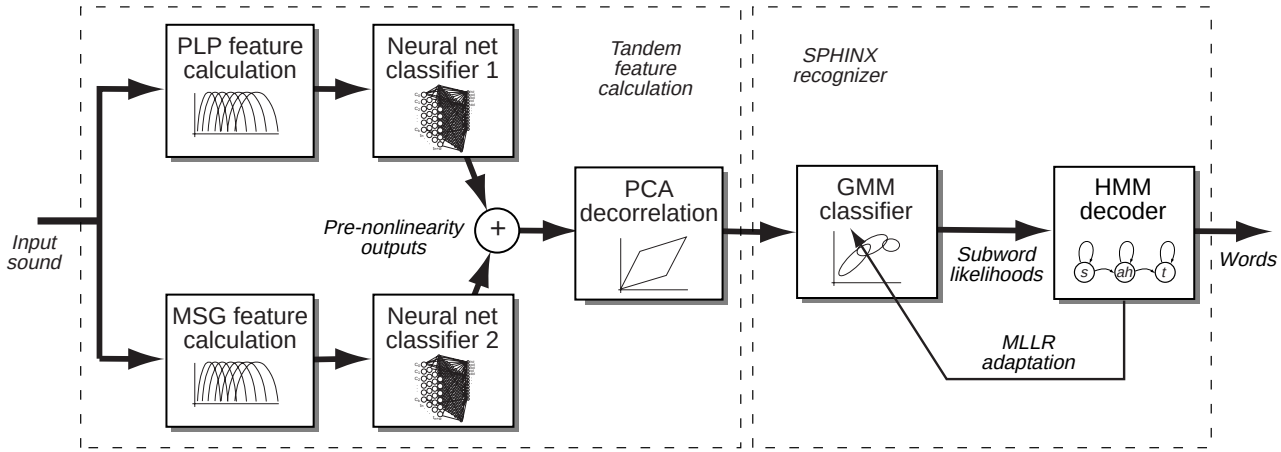


- **Neural net normalizes away noise**

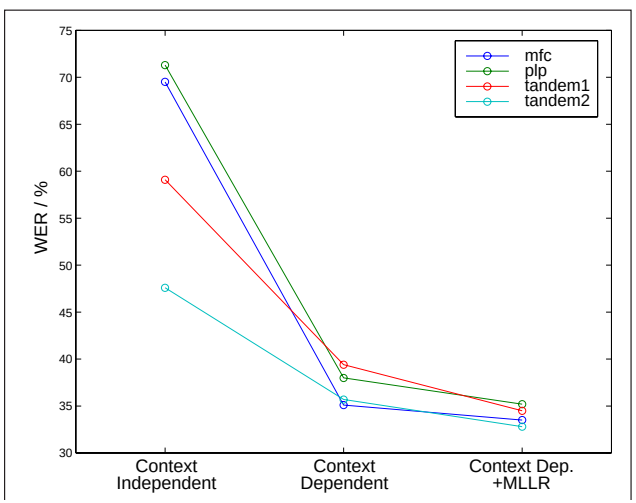
Lab ROSA

Tandem for large vocabulary recognition

- **CI Tandem front end + CD LVCSR back end**



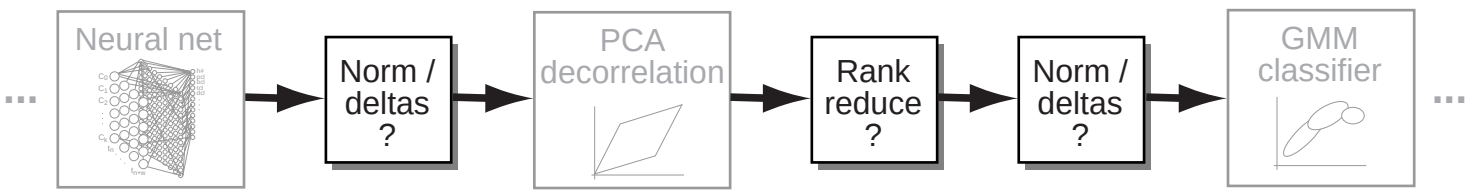
- **Tandem benefits reduced:**



‘Tandem-domain’ processing

(with Manuel Reyes)

- **Can we improve the ‘tandem’ features with conventional processing (deltas, normalization)?**



- **Somewhat..**

Processing	*Avg. WER 20-0 dB*	*Baseline WER ratio*
Tandem PLP mis-match baseline (24 els)	11.1%	70.3%
Rank reduce @ 18 els	11.8%	77.1%
Delta → PCA	9.7%	60.8%
PCA → Norm	9.0%	58.8%
Delta → PCA → Norm	**8.3%**	**53.6%**

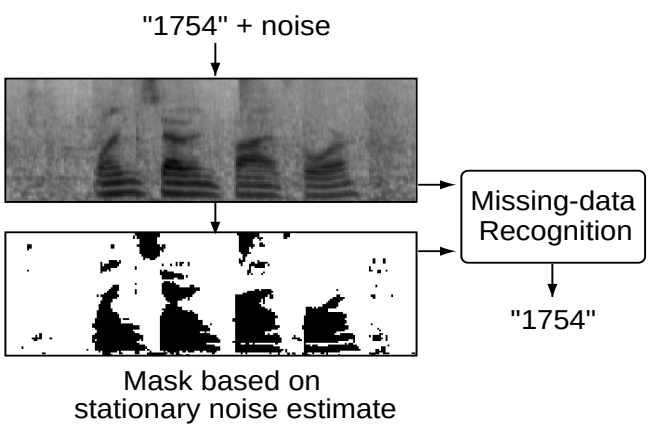
Missing data recognition

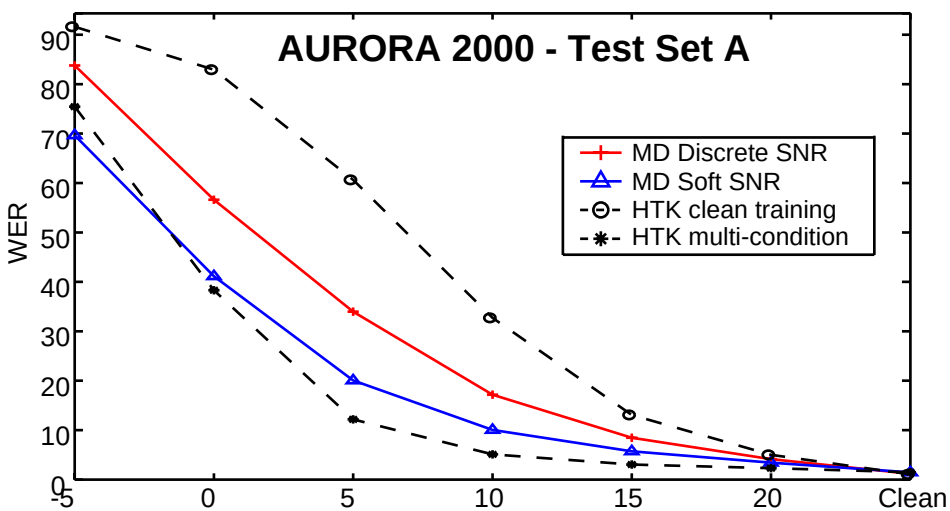
(with Cooke, Green, Barker @ Sheffield)

- **Noisy training seems to miss the point**
  - rather have single ‘clean’ models
- **Use missing feature theory...**
  - integrate over missing data dimensions x_m

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

  - trick is finding good/bad data mask
  - soft classification improves

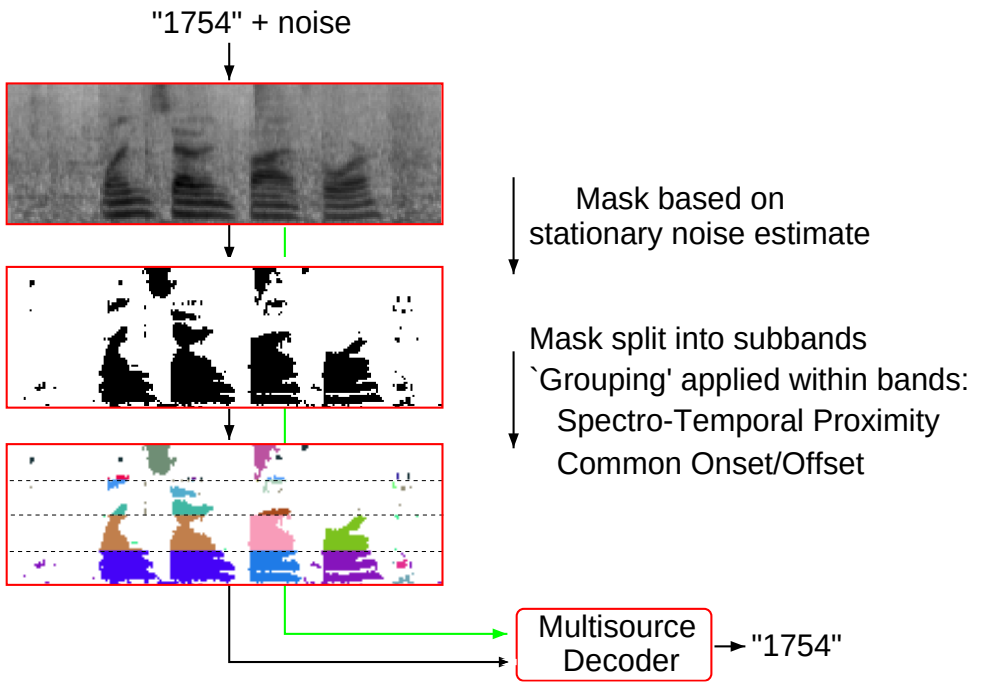




Multi-source decoding

(Jon Barker @ Sheffield)

- **Search of sound-fragment interpretations**



- **CASA for masks/fragments**
  - larger fragments → quicker search

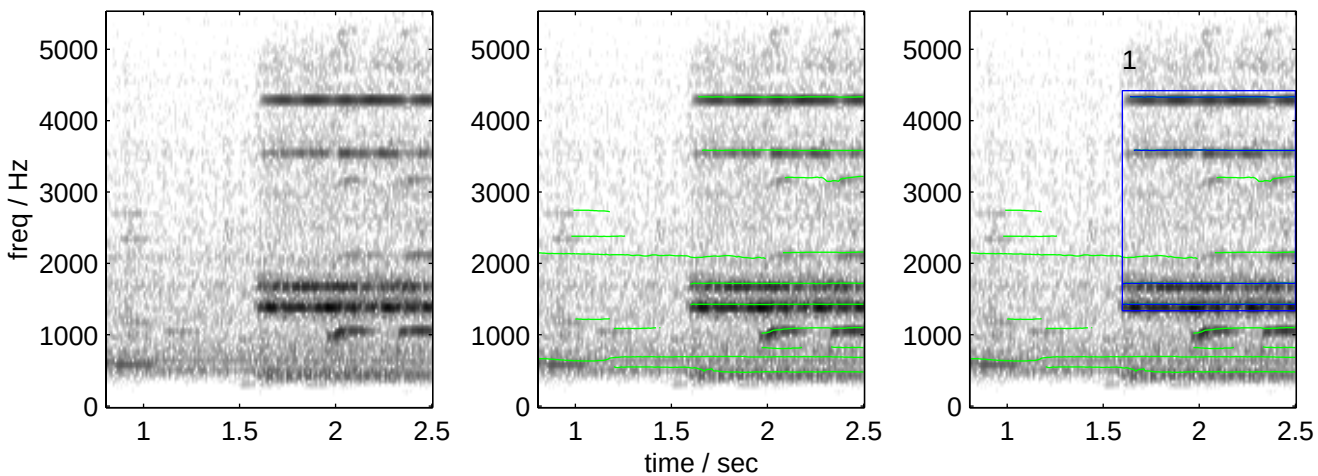
Lab ROSA

Outline

1. Introducing LabROSA
2. Robust speech recognition
3. **General audio analysis**
   - Alarm sound detection
   - Computational Auditory Scene Analysis
   - Music analysis
   - The Meeting Recorder project
4. Summary

Alarm sound detection

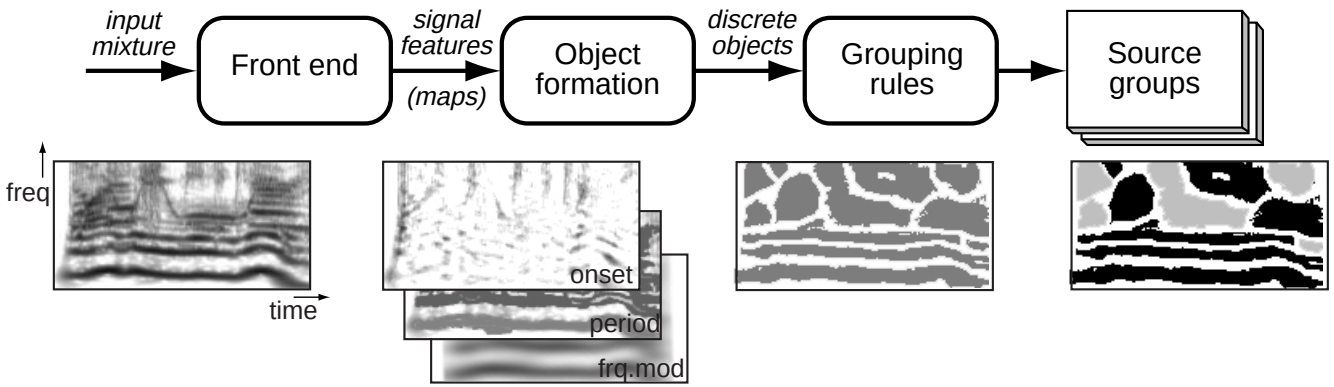
- **Alarm sounds have particular structure**
  - people ‘know them when they hear them’
  - build a generic detector?
- **Isolate alarms in sound mixtures**



  - representation of energy in time-frequency
  - formation of atomic elements
  - grouping by common properties (onset &c.)
  - classify by attributes...

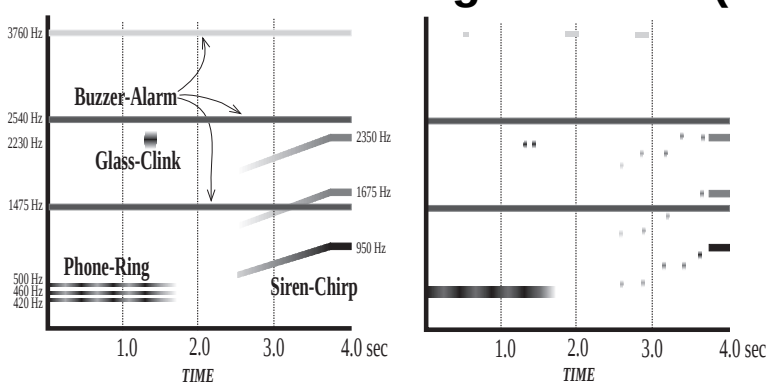
Computational Auditory Scene Analysis (CASA)

- **Implement psychoacoustic theory? (Brown'92)**



  - what are the features? how are they used?

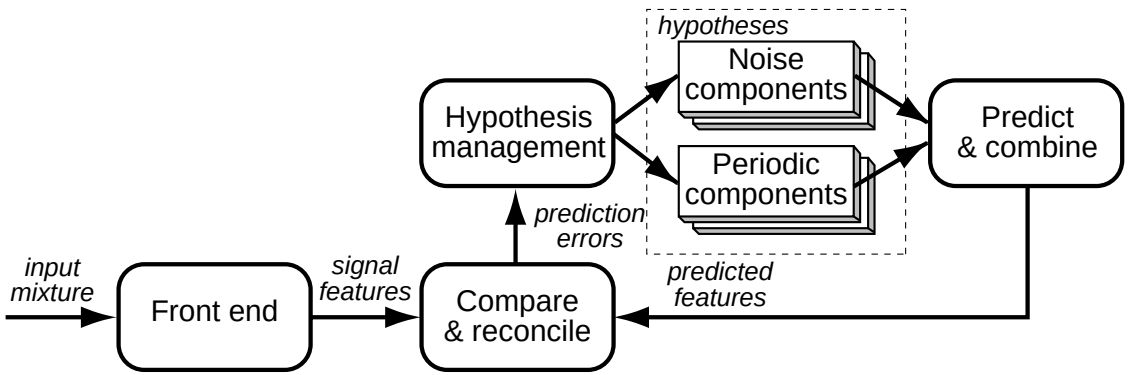
- **Additional 'knowledge' needed (Klassner'96)**



Lab ROSA

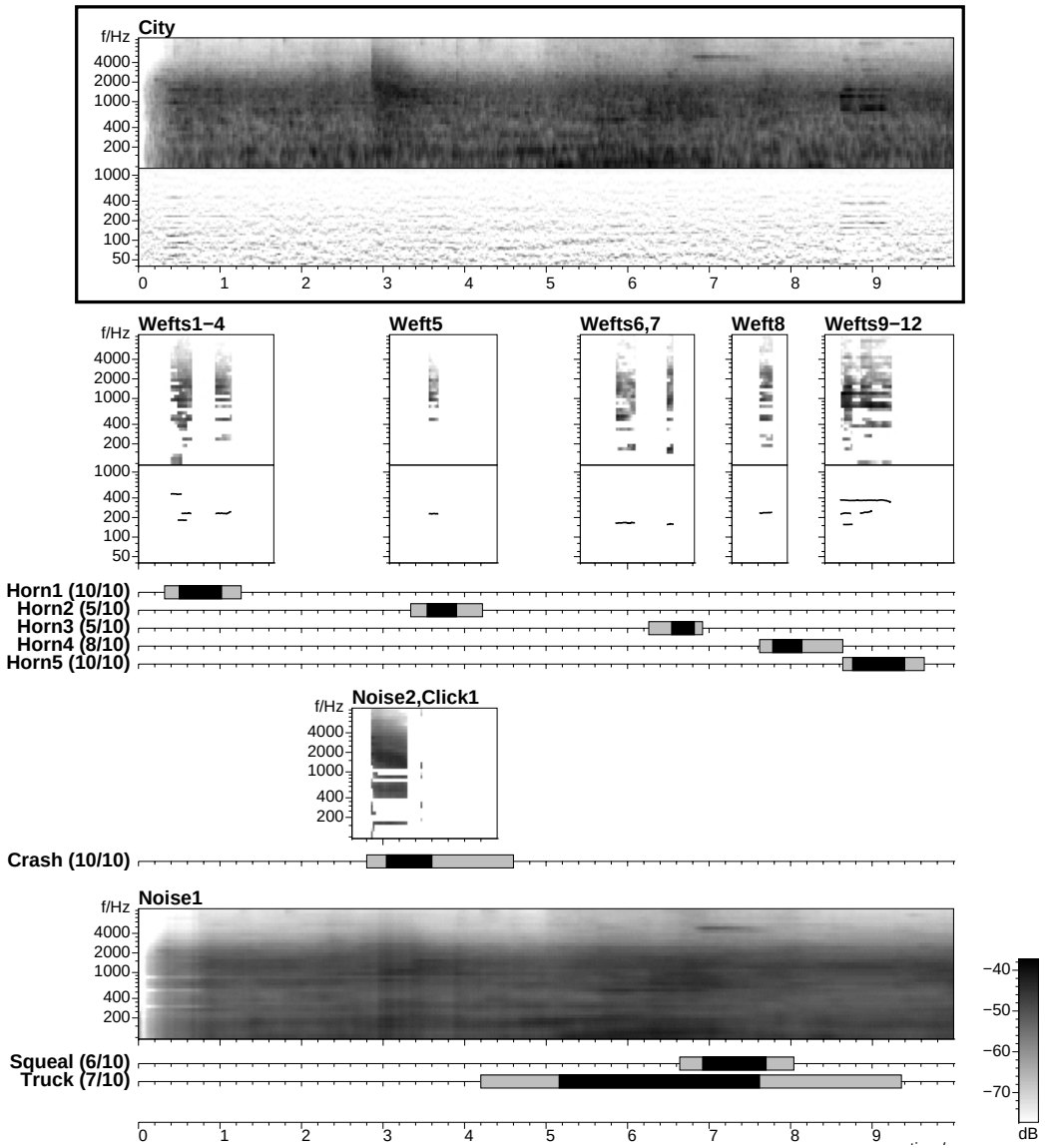
Prediction-driven CASA

- **Data-driven (bottom-up) fails for noisy, ambiguous sounds (most mixtures!)**
- **Need top-down constraints:**



- fit vocabulary of generic elements to sound ... bottom of a hierarchy?
- account for entire scene
- driven by prediction failures
- pursue alternative hypotheses

PDCASA example

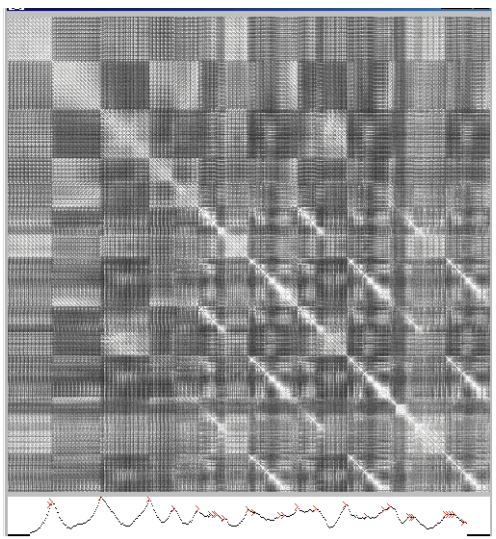


Lab ROSA

Music analysis: Structure recovery

(with Rob Turetsky)

- **Structure recovery by similarity matrices (after Foote)**

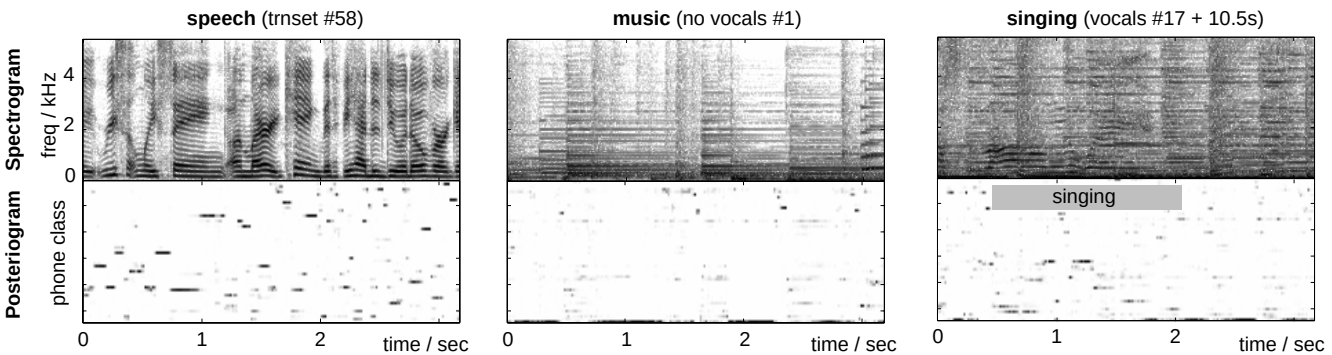


  - similarity distance measure?
  - segmentation & repetition structure
  - interpretation at different scales: notes, phrases, movements
  - incorporating musical knowledge: ‘theme similarity’

Music analysis: Lyrics extraction

(with Adam Berenzweig)

- **Vocal content is highly salient, useful for retrieval**
- **Can we find the singing? Use an ASR classifier:**



- **Frame error rate ~20% for segmentation based on posterior-feature statistics**
- **Lyric segmentation + transcribed lyrics → training data for lyrics ASR...**

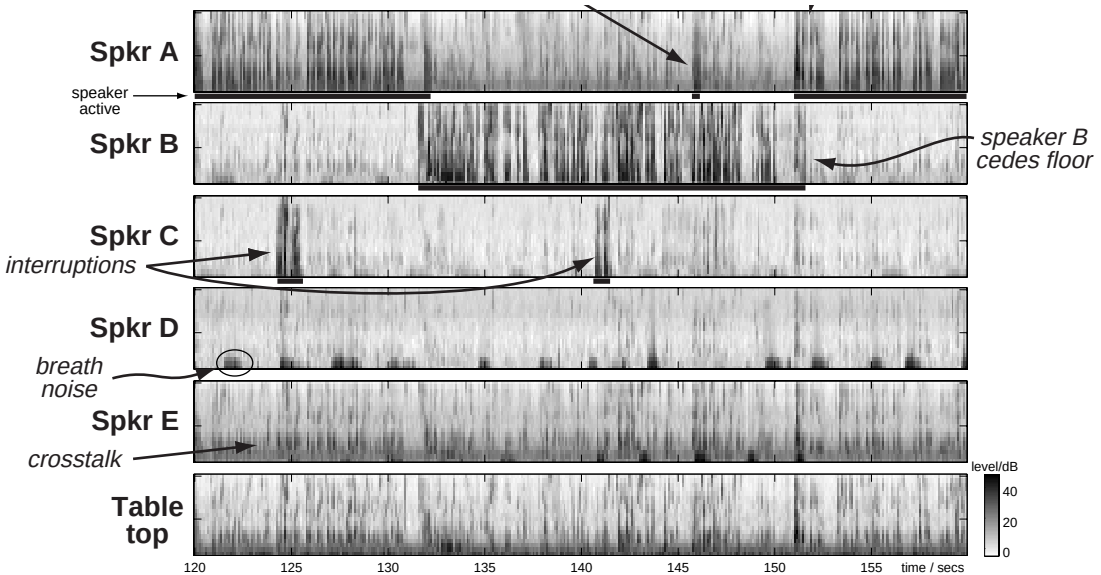
The Meeting Recorder project

(with ICSI, UW, SRI, IBM)

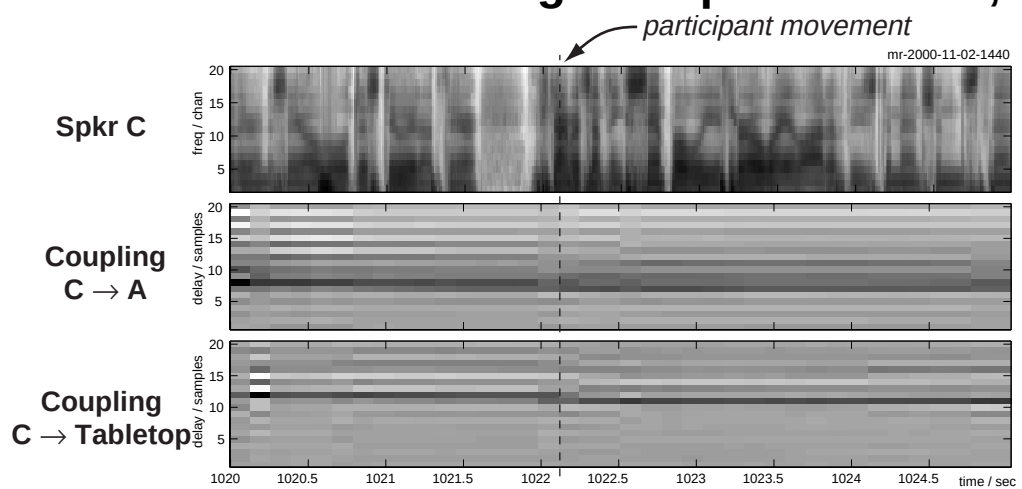
- **Microphones in conventional meetings**
  - for summarization/retrieval/behavior analysis
  - informal, overlapped speech
- **Data collection (ICSI, UW, ...):**

  - 100 hours collected, ongoing transcription
  - headsets + tabletop + ‘PDA’

Meeting recorder: Difficult data



- **Cross-correlation gives speaker turns, motion**



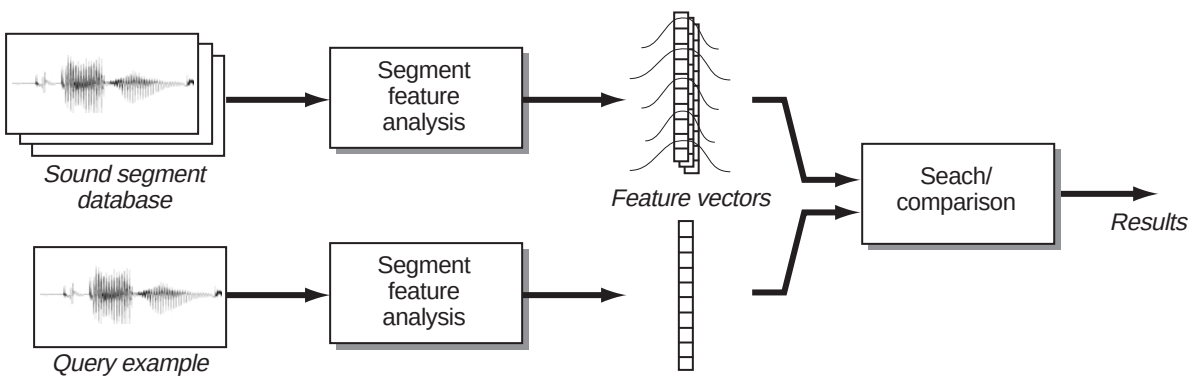
Lab ROSA

Outline

1. Introducing LabROSA
2. Robust speech recognition
3. General audio analysis
4. **Summary**
   - Some future project ideas
   - LabROSA Summary

Future: Audio Information Retrieval

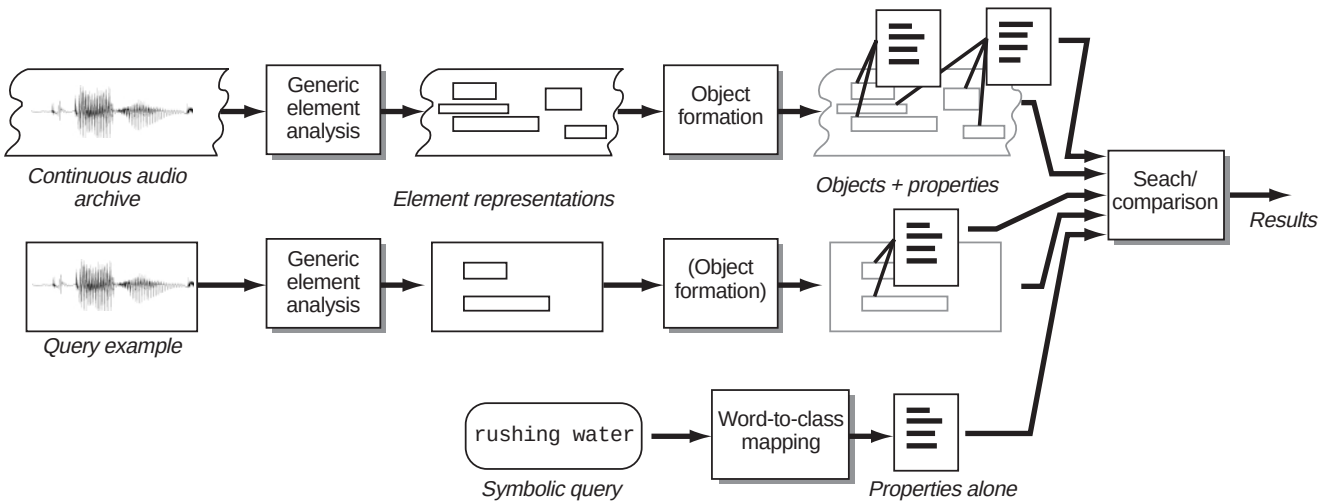
- **Searching in a database of audio**
  - speech .. use ASR
  - text annotations .. search them
  - sound effects library?
- **e.g. Muscle Fish “SoundFisher” browser**
  - define multiple ‘perceptual’ feature dimensions
  - search by proximity in (weighted) feature space



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

CASA for audio retrieval

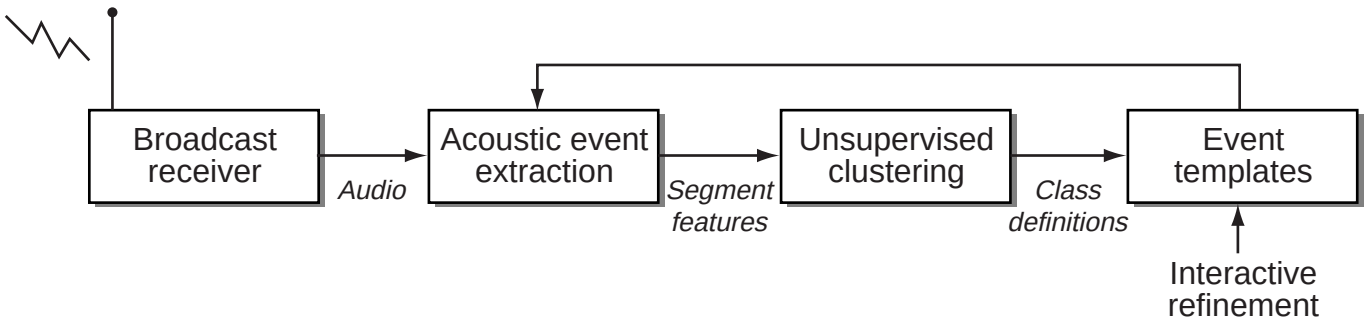
- **When audio material contains mixtures, global features are insufficient**
- **Retrieval based on element/object analysis:**



- features are calculated over grouped subsets

Future: ‘Machine listener’

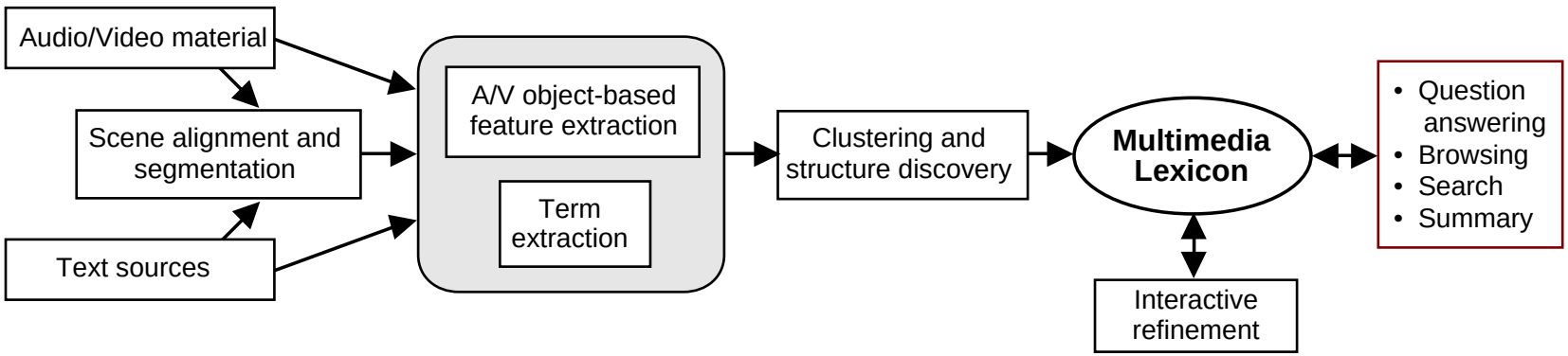
- **Goal: Unsupervised structure discovery**



- **What can you do with a large unlabeled training set (e.g. broadcast)?**
  - bootstrap learning: look for common patterns
  - have to learn generalizations in parallel: e.g. self-organizing maps, EM HMMs
  - post-filtering by humans may find ‘meaning’ in clusters

Audio-video-text content analysis

(with Shih-Fu Chang, Kathleen McKeown)



- **Audio and video provide complementary info**
  - correlate object features to define templates?
- **Associated text annotations provide a very small amount of labeling**
  - .. but for a very large number of examples – sufficient to obtain purchase?
  - build a ‘multimedia lexicon’ for question-answering

Summary: Applications for sound organization

What do people do with their ears?

- **Human-computer interface**
  - .. includes knowing when (& why) you've failed
- **Robots**
  - intelligence requires perceptual awareness
  - Sony's AIBO: dog-hearing
- **Archive indexing & retrieval**
  - pure audio archives
  - true multimedia content analysis
- **Content 'understanding'**
  - intelligent classification & summarization
- **Autonomous monitoring**
- **'Structure discovery' algorithms**

LabROSA Summary

DOMAINS

- Broadcast
- Movies
- Lectures
- Meetings
- Personal recordings
- Location monitoring

ROSA

- Object-based structure discovery & learning

- Speech recognition
- Speech characterization
- Nonspeech recognition
- Scene analysis
- Audio-visual integration
- Music analysis

APPLICATIONS

- Structuring
- Search
- Summarization
- Awareness
- Understanding