General Soundtrack Analysis

Dan Ellis
<dpwe@ee.columbia.edu>

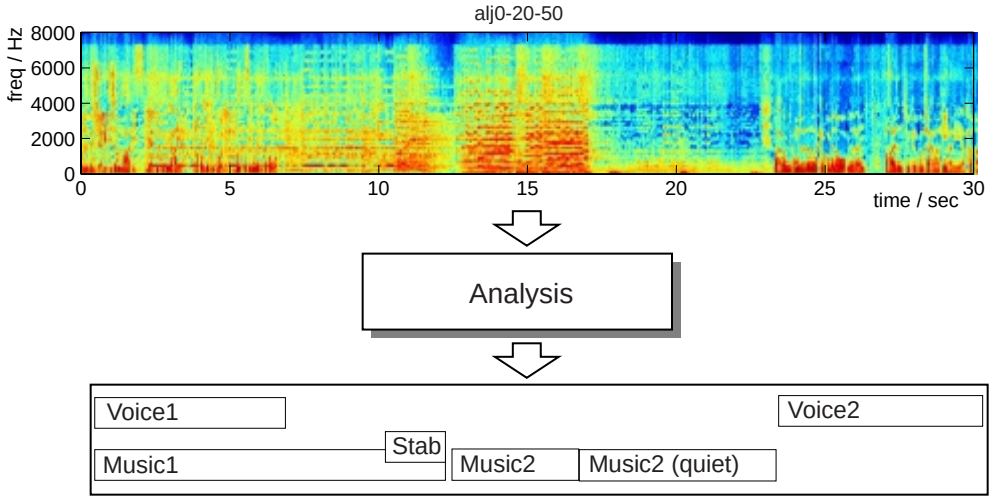
Laboratory for Recognition and Organization of Speech and Audio
(Lab**ROSA**)
Electrical Engineering, Columbia University
http://labrosa.ee.columbia.edu/

Outline

1. **LabROSA introduction**
2. **Broadcast soundtrack monitoring**
3. **Example technologies**
4. **Summary**

Lab ROSA

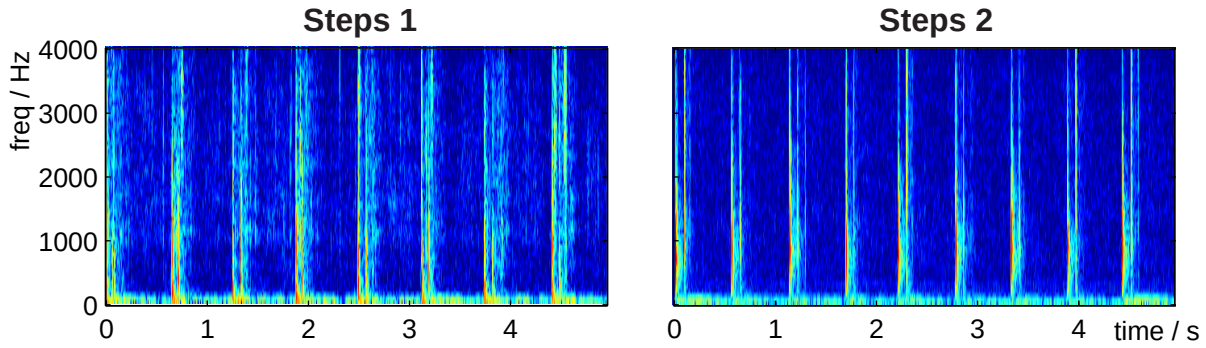
1 LabROSA: Sound Organization



- **Analyzing and describing complex sounds:**
  - continuous sound mixture
    → distinct objects & events
- **Human listeners as the prototype**
  - strong subjective impression when listening
  - ..but hard to 'see' in signal

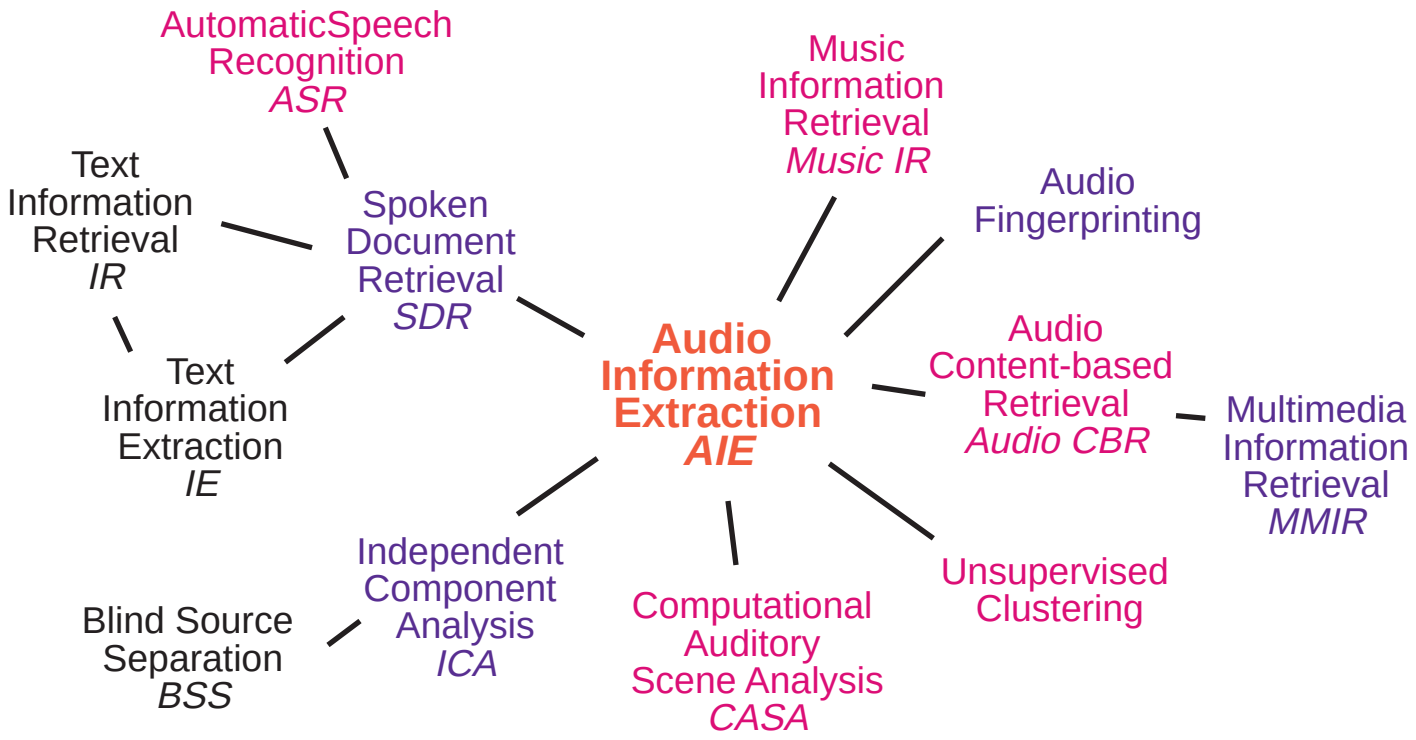
Lab ROSA

The information in sound



- **Hearing confers evolutionary advantage**
  - optimized to get 'useful' information from sound
- **Enormous detail is available in familiar sounds**
  - 'ecological' influence on our sense of hearing

Audio Information Extraction



- **Domain**
  - text ... speech ... music ... general audio
- **Operation**
  - recognize ... index/retrieve ... organize

Lab ROSA

LabROSA Summary

DOMAINS

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

ROSA

- Object-based structure discovery & learning
- Speech recognition
- Speaker description
- Nonspeech recognition
- Scene Analysis
- Audio-visual integration
- Music analysis

APPLICATIONS

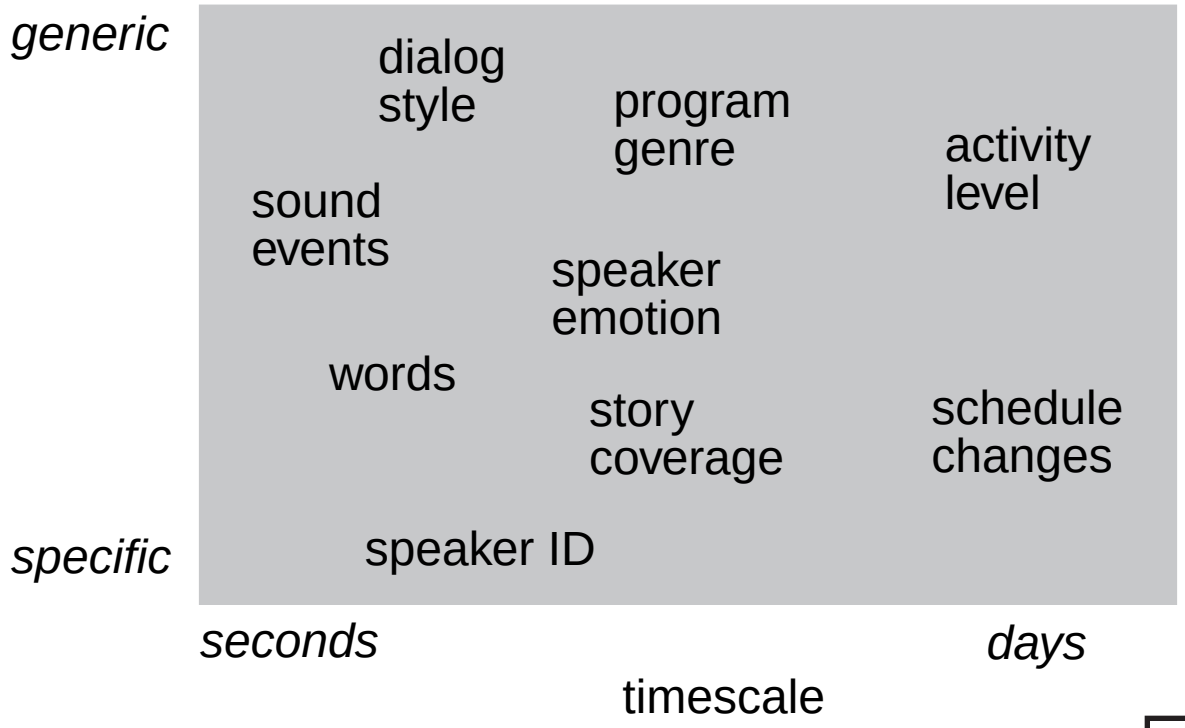
- Multimedia access & search
- Personal media management
- Machine perception & awareness
- Prostheses / human augmentation
- Automatic judgments/recommendation

Outline

1. LabROSA introduction
2. **Broadcast soundtrack monitoring**
   - Available information
   - Information filtering
   - Intelligent analysis
3. Example technologies
4. Summary

2 Broadcast soundtrack monitoring: What information is available?

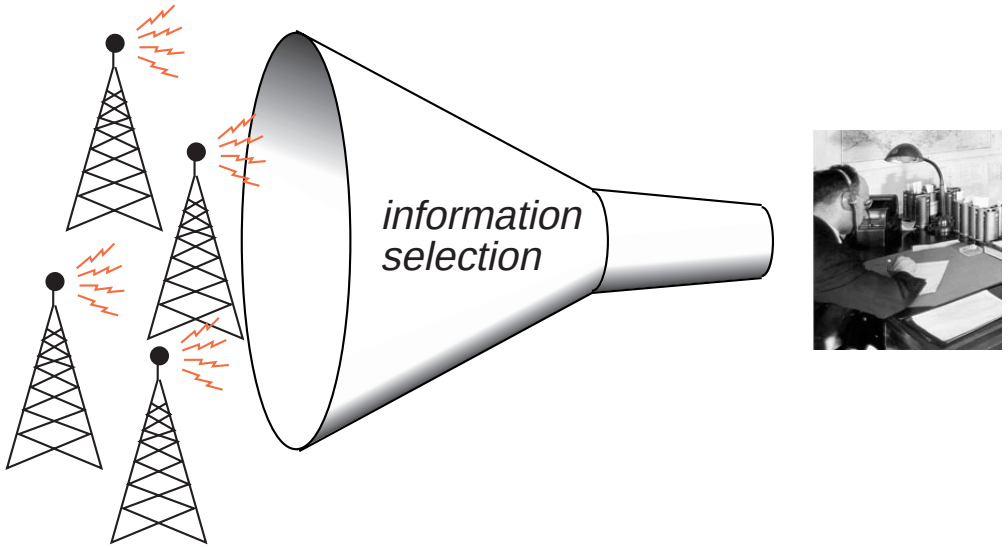
- **Video and soundtrack reinforce, complement**
  - hard things in one domain can be easy in other
- **Information at different scales:**



Lab ROSA

Information filtering

- **Maximizing analyst utility:**

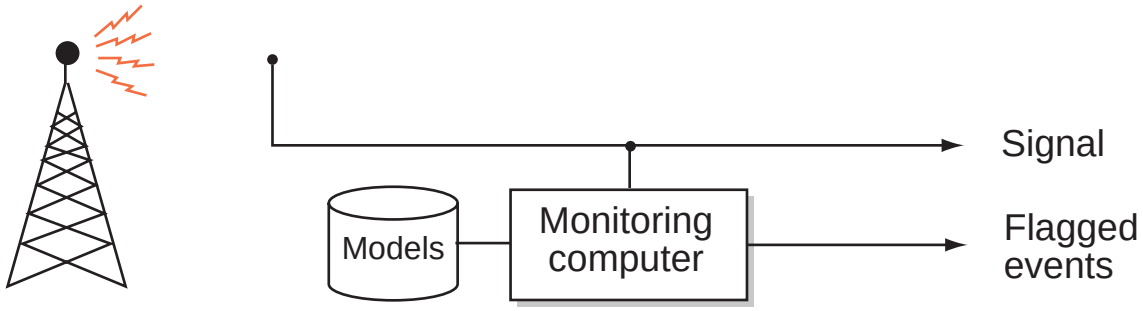


- **Automatic support for 'triage':**
  - segmentation
  - inferring schedules
  - identify repeated segments
  - generic categorization/labeling
  - task-specific flagging

Lab ROSA

Automatic labeling/flagging

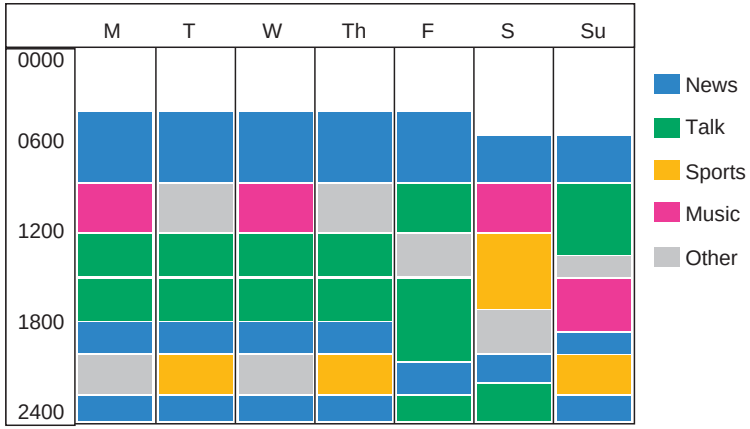
- **Computer as vigilant monitor**



- **Range of tasks**
  - generic ‘unusual’ patterns
  - specific trigger events
- **Issues**
  - false alarms vs. misses
  - combinations of simple detectors for higher-level classes e.g. program type

Schedule summarization

- **Learn the patterns of a particular broadcaster:**
  - 24/7 monitoring
  - automatic segmentation into programs
  - automatic classification of genres



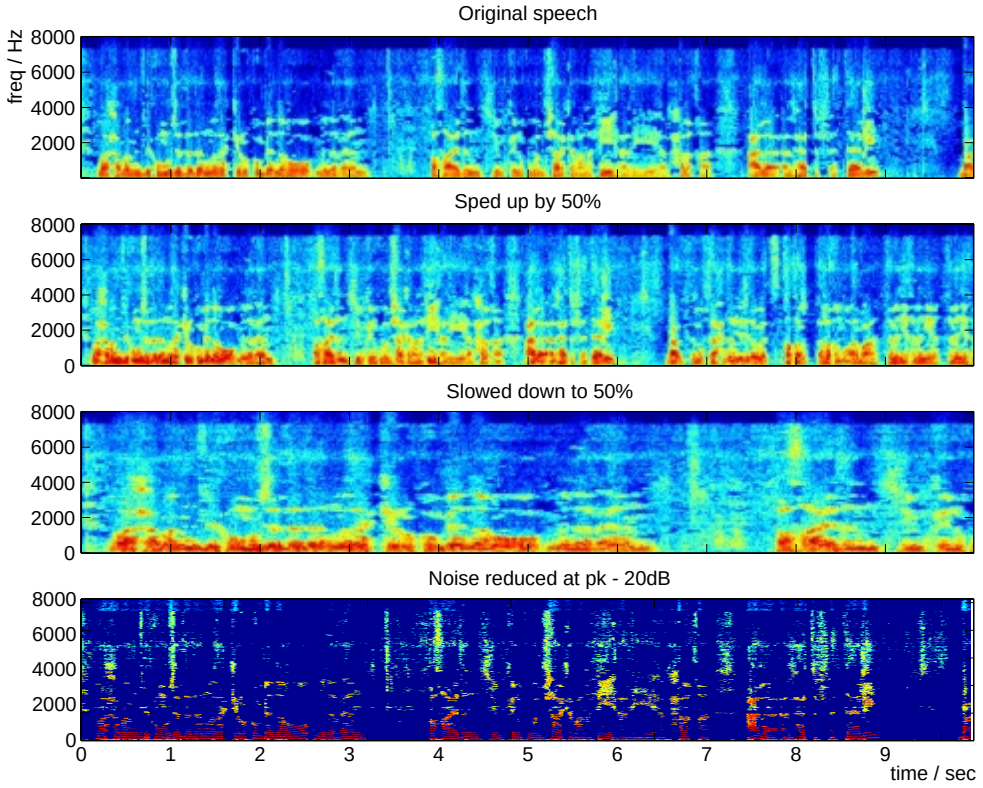
- **Support information extraction**
  - program types, repetitions, summarization
  - schedule changes → activity?

Outline

1. LabROSA introduction
2. Broadcast soundtrack monitoring
3. **Example technologies**
   - Sound enhancement
   - Locating repetitions
   - Labeling soundtracks
4. Summary

3 Example technologies: Sound enhancement

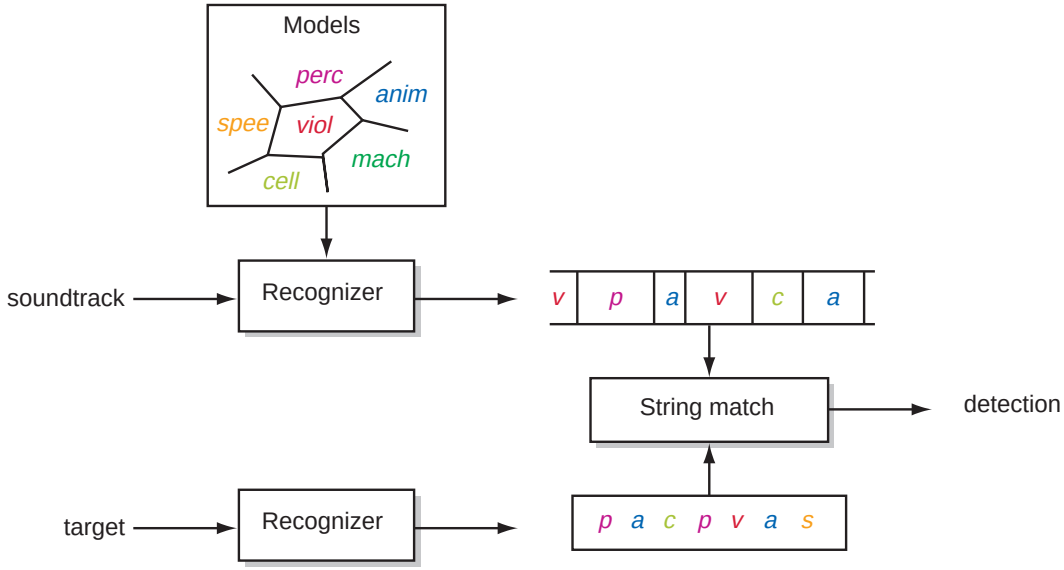
- **Time scale modification**
  - speed up for rapid skimming
  - slow down for close listening
- **Noise reduction**



Lab ROSA

Repetition detection

- **Matching signals is expensive**
  - but matching *strings* is fast
- **Represent audio as simple string**
  - recognition into arbitrary classes
  - match detection becomes string matching

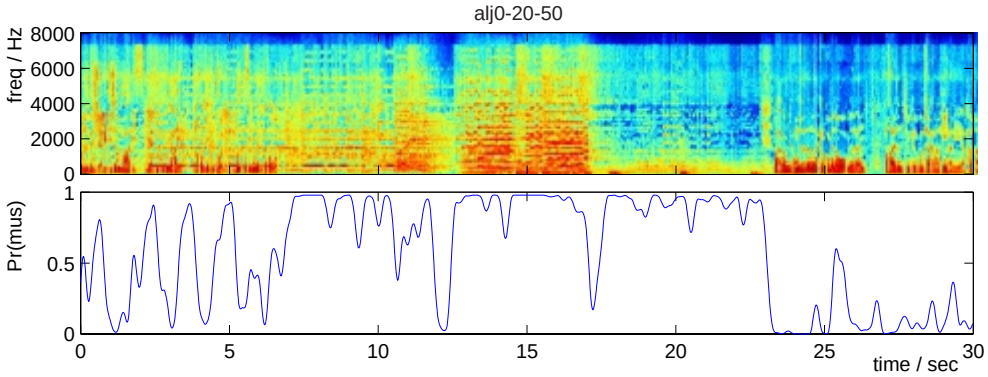


- **Monitor for many ‘signatures’**

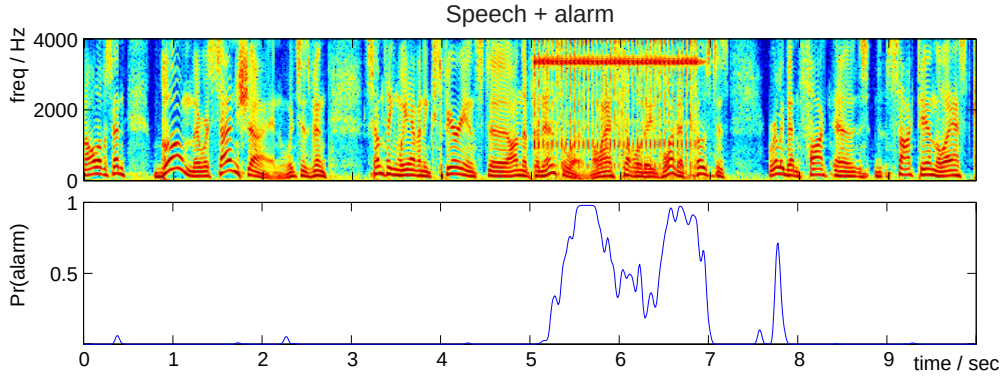
Lab ROSA

Soundtrack labeling

- **Broad class: speech-music**



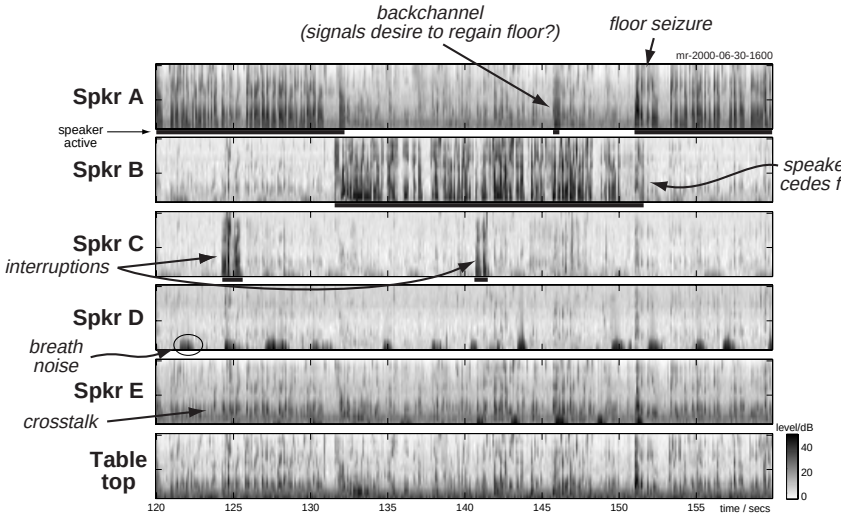
- **Specific events: alarms**

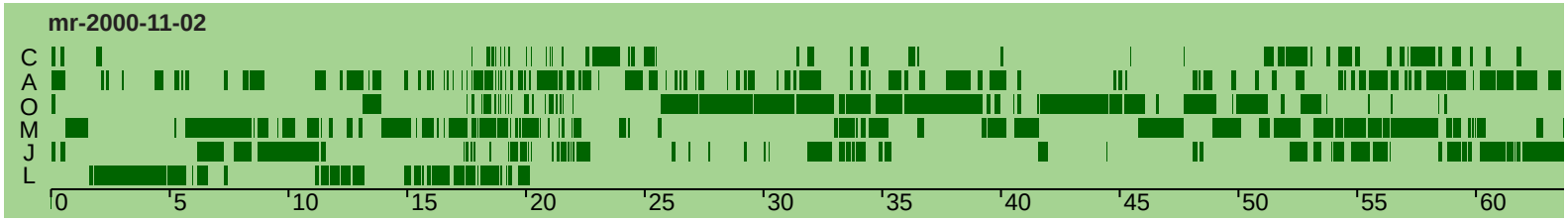


Lab ROSA

Speaking style

- **Recognizing information other than words**
- **Meeting Recorder project: Locate overlaps**





- **Speaker emotion**
  - depends on good baseline

Lab ROSA

Outline

1. LabROSA introduction
2. Broadcast soundtrack monitoring
3. Example technologies
4. **Summary**

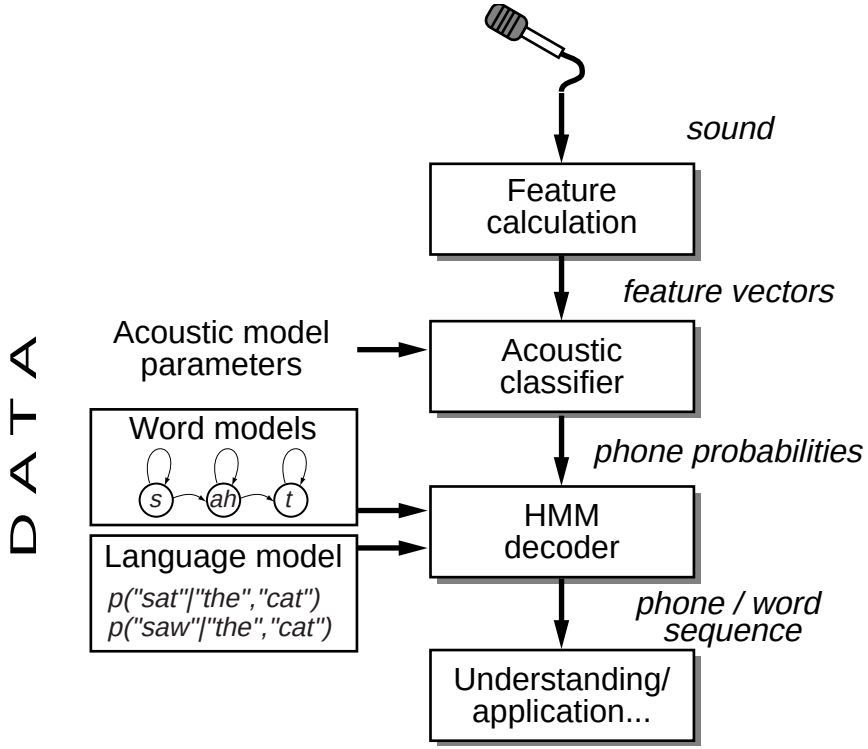
4 Summary: Soundtrack analysis

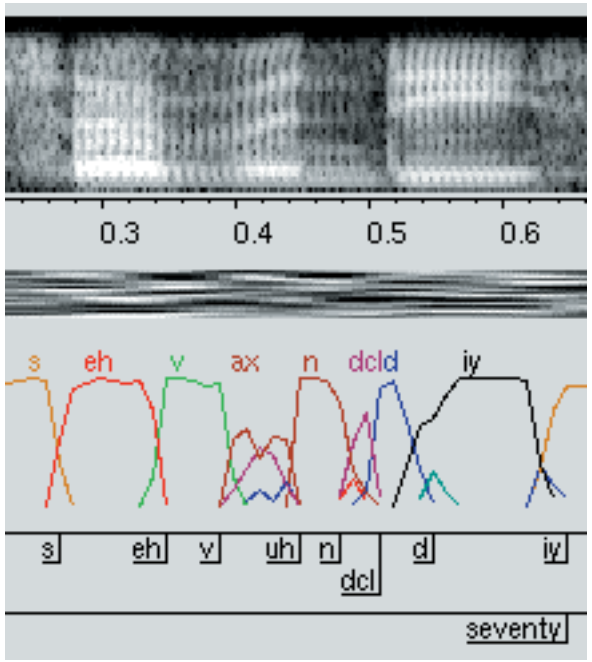
- **Soundtrack carries information**
  - useful and detailed
  - complementary to image
  - rapid processing
- **Open source monitoring**
  - Need to find the interesting bits
  - Short-time: specific detectors
  - Long-time: schedule, genre classification
- **Current techniques**
  - Signal enhancement
  - Detectors for speech, music, alarms etc.
  - Classification of interaction style, emotion

Extra slides

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...**

Lab ROSA

The Meeting Recorder project

(with ICSI, UW, SRI, IBM)

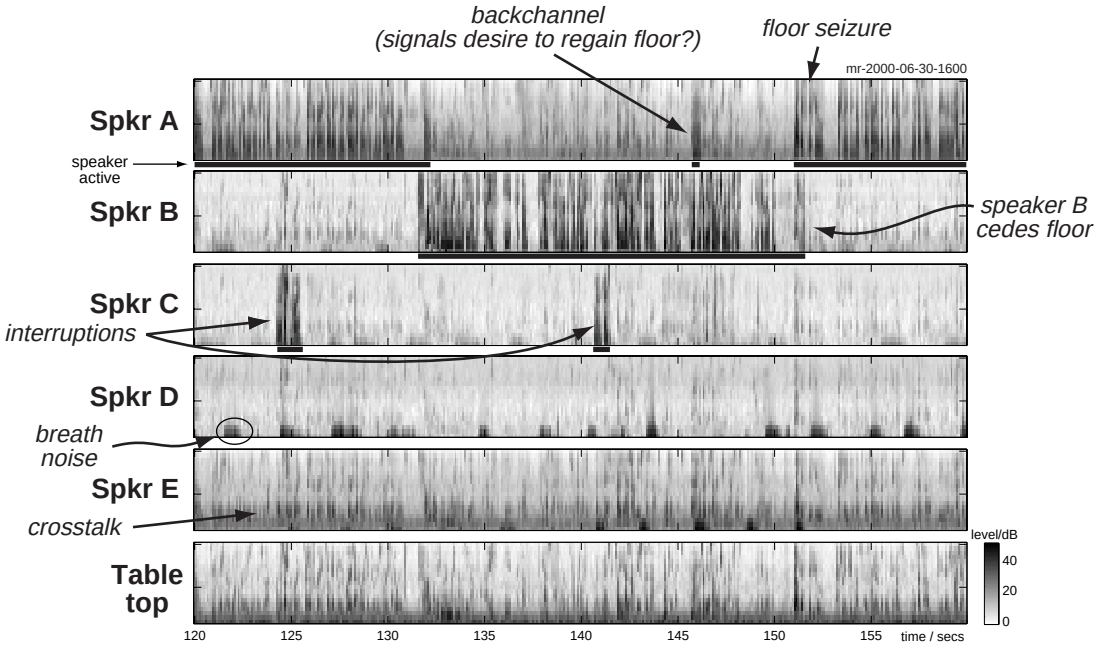
- **Microphones in conventional meetings**
  - for summarization/retrieval/behavior analysis
  - informal, overlapped speech
- **Data collection (ICSI, UW, ...):**

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

Lab ROSA

Crosstalk cancellation

- **Baseline speaker activity detection is hard:**

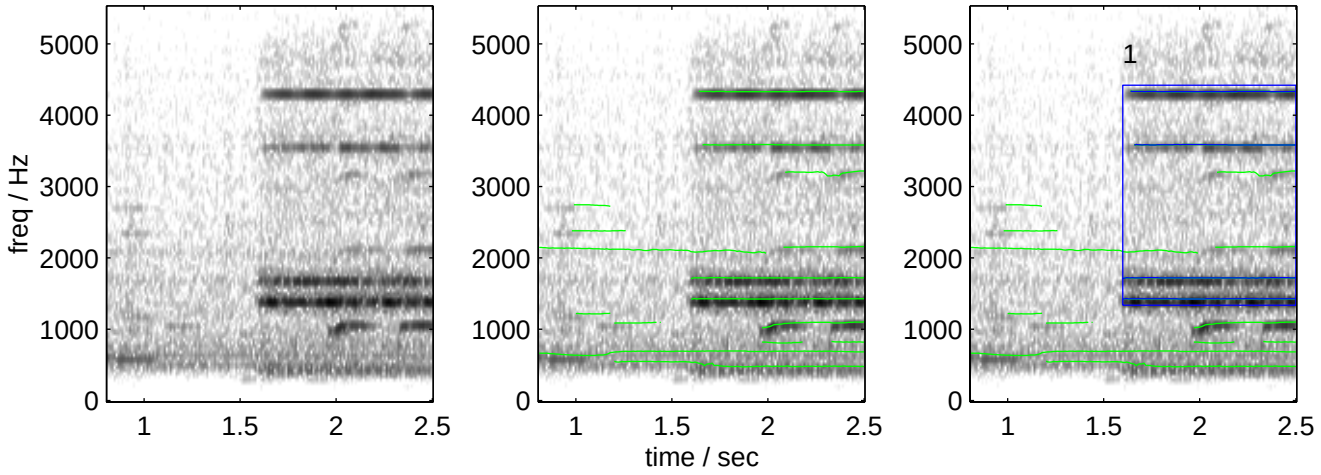


- **Noisy crosstalk model:** $\mathbf{m} = \mathbf{C} \cdot \mathbf{s} + \mathbf{n}$
- **Estimate subband C_{Aa} from A's peak energy**
  - ... including pure delay (10 ms frames)
  - ... then linear inversion

Lab ROSA

Alarm sound detection

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

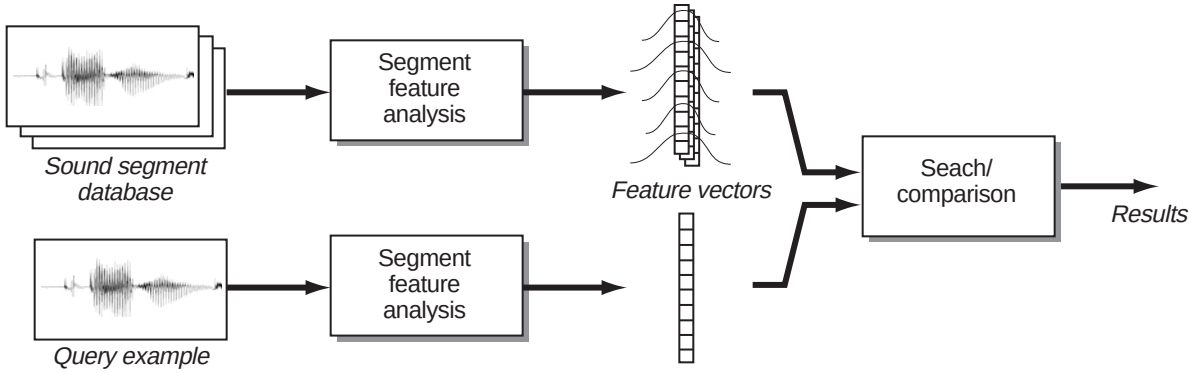


  - representation of energy in time-frequency
  - formation of atomic elements
  - grouping by common properties (onset &c.)
  - classify by attributes...
- **Key: recognize *despite* background**

Audio Information Retrieval

(with Manuel Reyes)

- **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

Lab ROSA

Audio Retrieval: Results

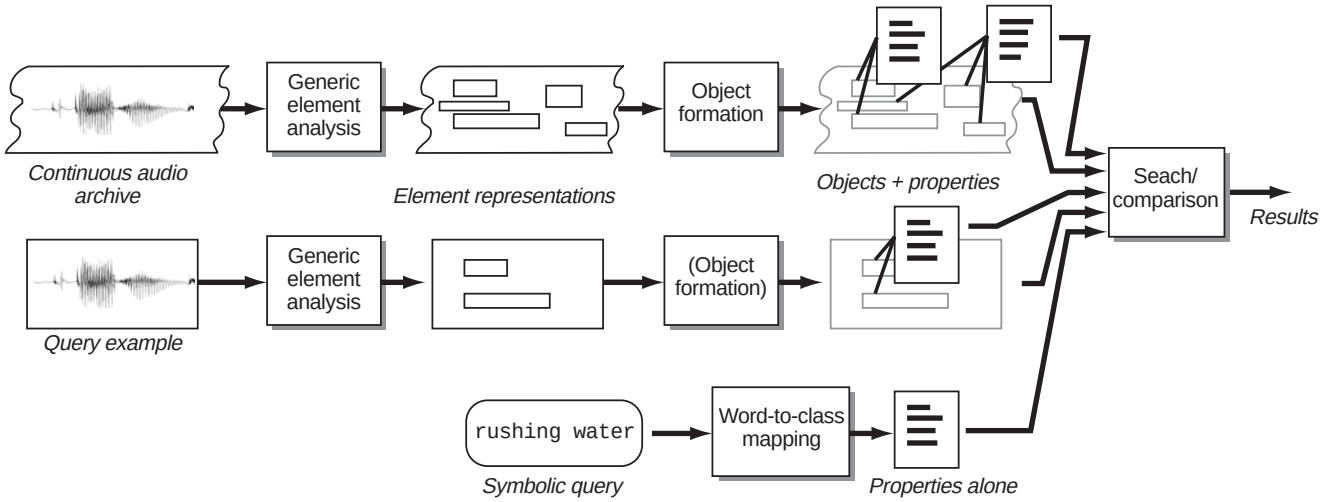
- **Musclefish corpus**
  - most commonly reported set
- **Features**
  - mfcc, brightness, bandwidth, pitch ...
  - no temporal sequence structure
- **Results:**
  - 208 examples, 16 classes, 84% correct
  - confusions:

	Instr	*Spch*	*Env*	*Anim*	*Mech*
Musical instrs.	136 (**14**)				
Speech		17 (**7**)			**2**
Eviron.		**2**	6 (**1**)		
Animals	**2**		**2**	1 (**0**)	
Mechanical	**1**				15 (**2**)

Lab ROSA

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