Recognition and Organization of Speech and Audio

Dan Ellis
<dpwe@ee.columbia.edu>

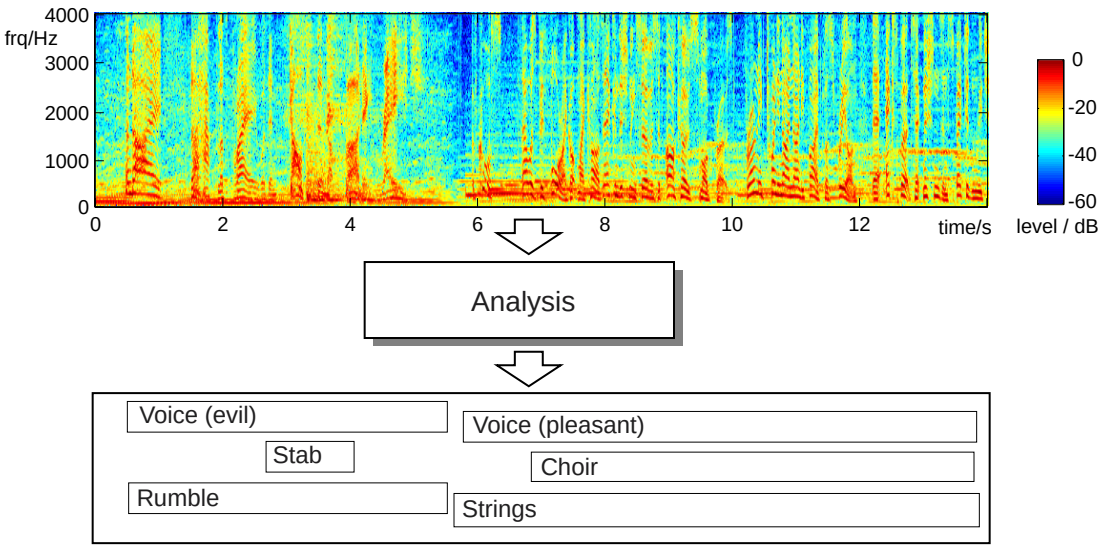
Laboratory for Recognition and Organization of Speech and Audio
(Lab**ROSA**)
Electrical Engineering, Columbia University
http://labrosa.ee.columbia.edu/

Outline

1. **Audio Organization**
2. **Speech, music, and other**
3. **Future work**

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

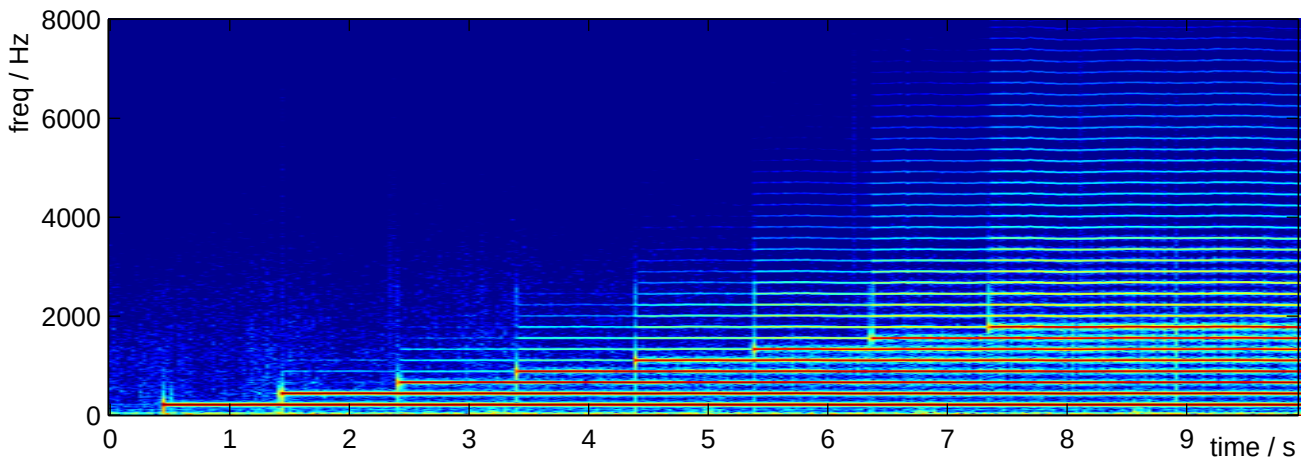
1 Audio Organization



- **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

Human Sound Organization

- **Sound percepts depend on 'grouping cues'**
  - common onset across frequency
  - common periodicity (fundamental)
  - spatial (binaural) cues, familiarity, ...

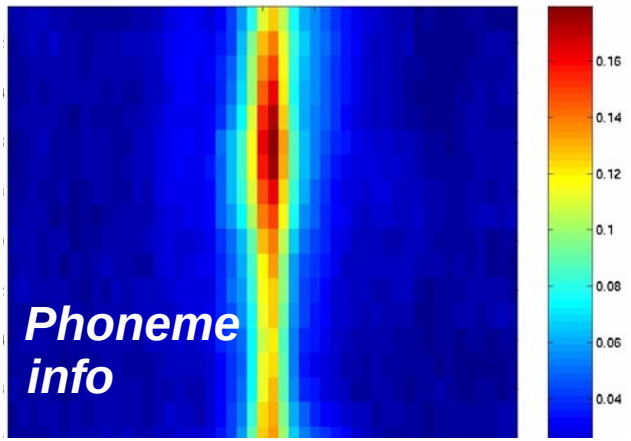


- **Hearing confers evolutionary advantage**
  - optimized to get 'useful' information from sound
- **Auditory perception is *ecologically* grounded**
  - scene analysis is preconscious (→ illusions)

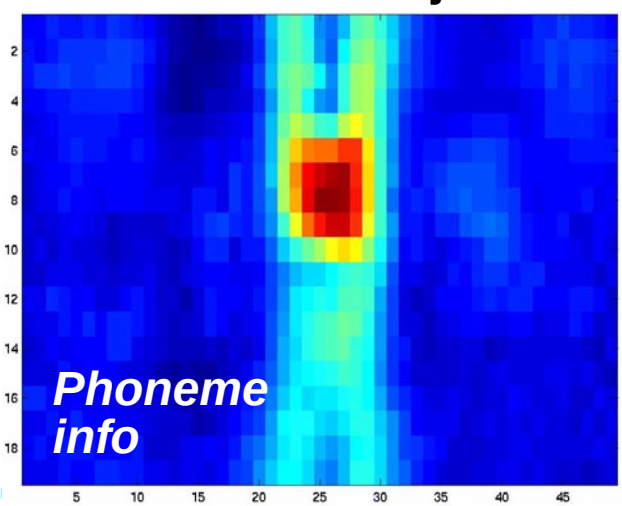
2 Speech & Speaker recognition

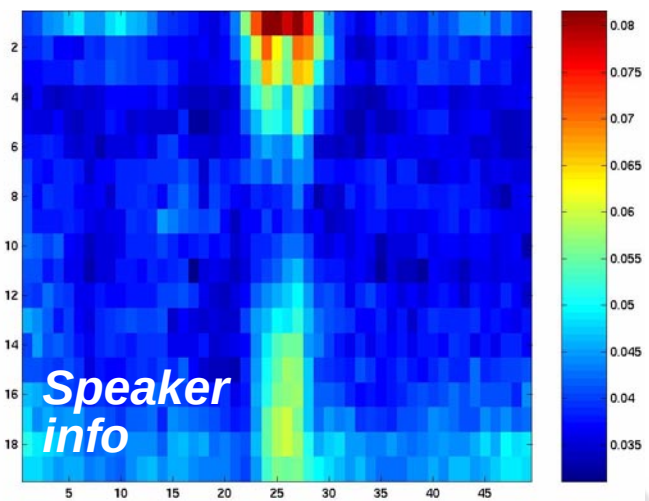
(Patricia Scanlon)

- **Mutual Information identifies where the information is in time/ frequency:**
  - little temporal structure averaged over all sounds



- **Better with just *vowels:***





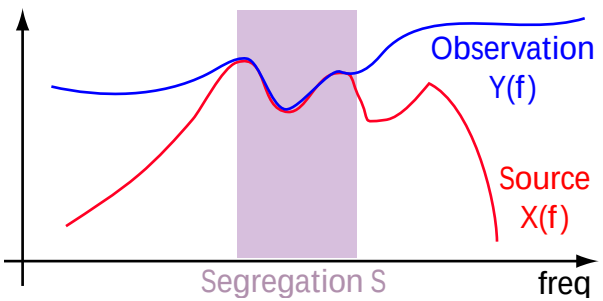
Speech Fragment recognition

(Barker & Cooke/Sheffield)

- **Standard classification chooses between models M to match source features X**

$$M^* = \underset{M}{\operatorname{argmax}}\, P(M|X) = \underset{M}{\operatorname{argmax}}\, P(X|M) \cdot \frac{P(M)}{P(X)}$$

- **Mixtures → observed features Y, segregation S, all related by $P(X|Y,S)$**



- *spectral features* allow clean relationship

- **Joint classification of model and segregation:**

$$P(M,S|Y) = P(M)\int P(X|M) \cdot \frac{P(X|Y,S)}{P(X)} dX \cdot P(S|Y)$$

The Meeting Recorder Project

(CompSci, ICSI, UW, IDIAP, SRI, IBM)

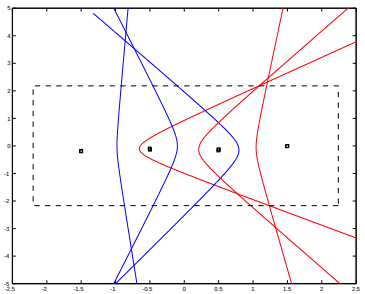
- **Microphones in conventional meetings**
  - for summarization/retrieval/behavior analysis
  - informal, overlapped speech
- **Data collection (ICSI, UW, IDIAP):**

  - 100 hours collected, ongoing transcription
- **NSF 'Mapping Meetings' project**
  - also interest from NIST, DARPA, EU

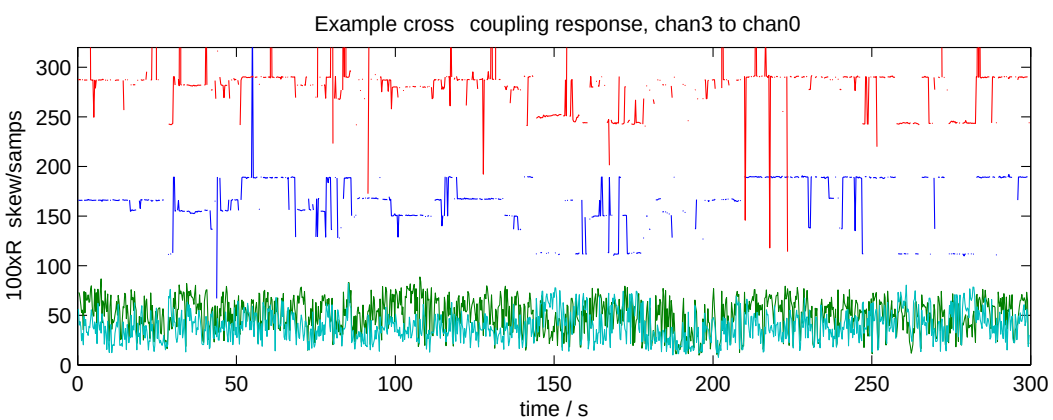
Tabletop mics: Turn detection

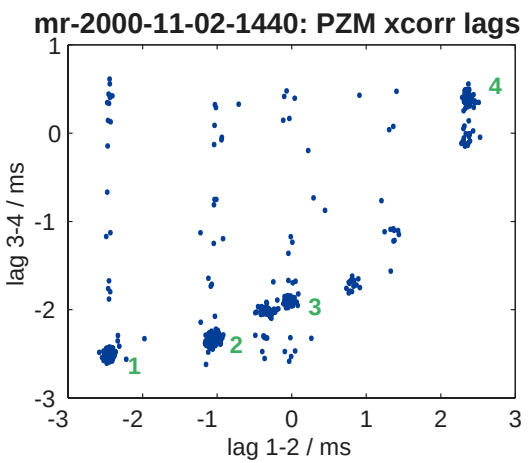
(Huan Wei Hee, Jerry Liu)

- **4 mics ~ 1m separated along center of table**
  - 3 timing differences
  - slight U/D offset to disambiguate



- **Hi-res cross-correlation for timings**
  - use normalized peak value for confidence
  - cluster results



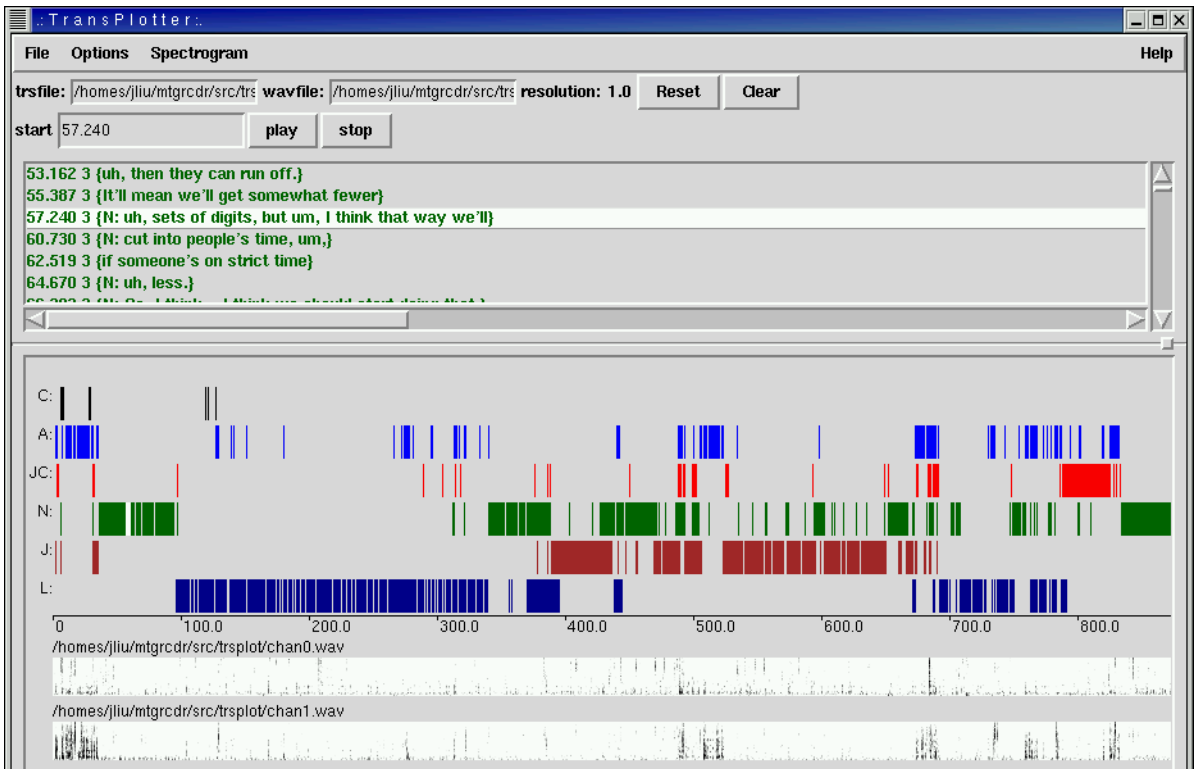


Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Visualization: transPlotter

(Jerry Liu)

- **Speaker turn *patterns* are informative**

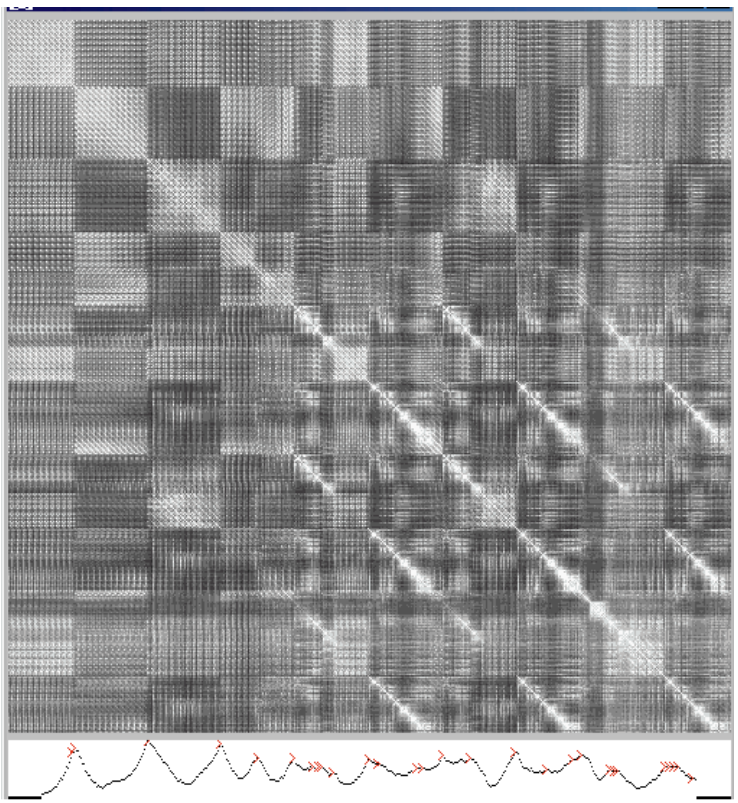


- **Browser for 'high-level' view, quick examination**
  - snack, iwidgets based

Music analysis: Structure recovery

(Rob Turetsky, Alex Sheh)

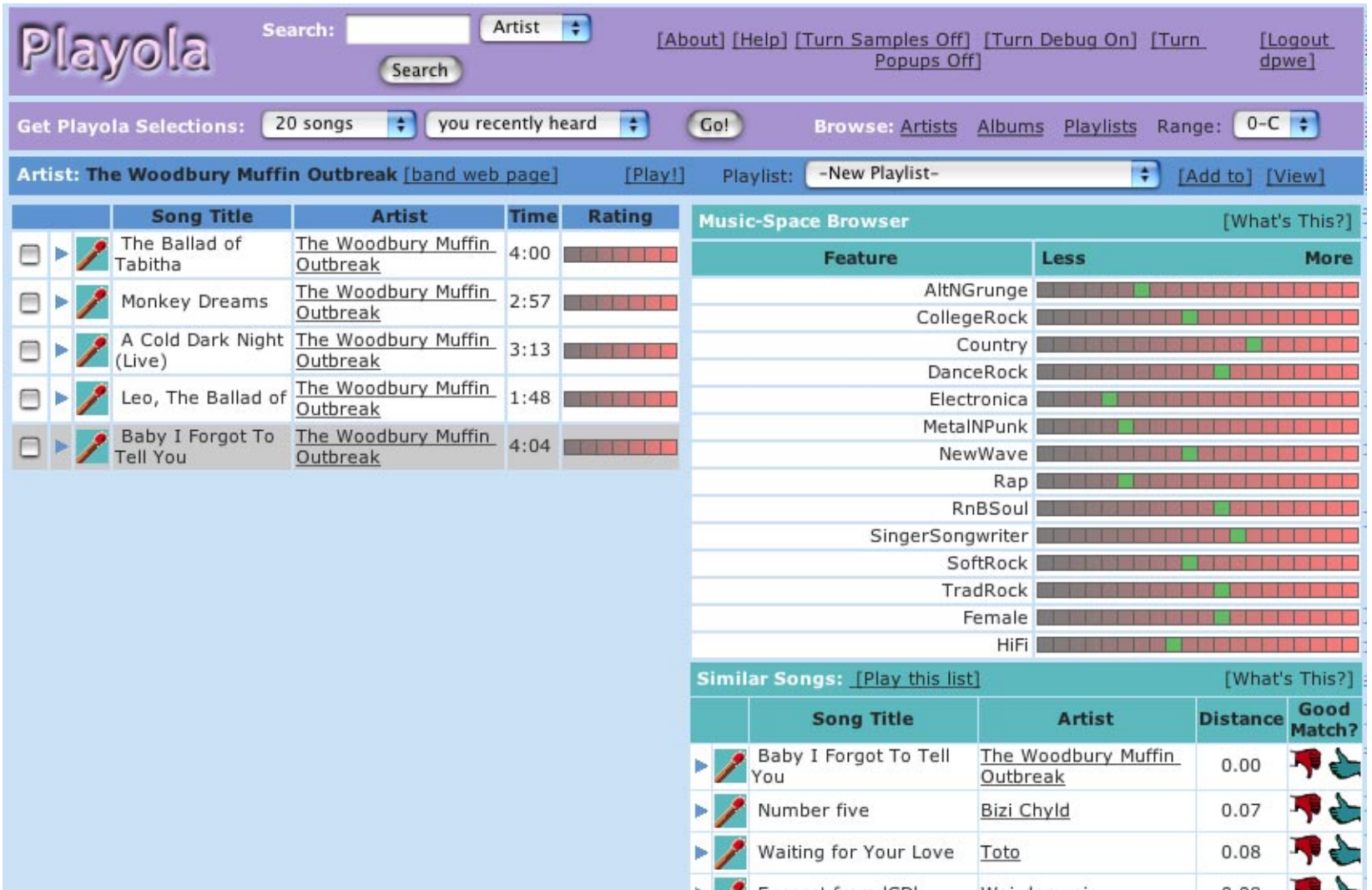
- **Identify & match segment-level structure in music**



- Foote similarity matrices point to repeated sections
- Chord segmentation using ASR ‘hidden state’ model (train with chord transcriptions)
- Note transcription/timing by aligning audio to resynthesized MIDI versions

Music Similarity & Recommendation

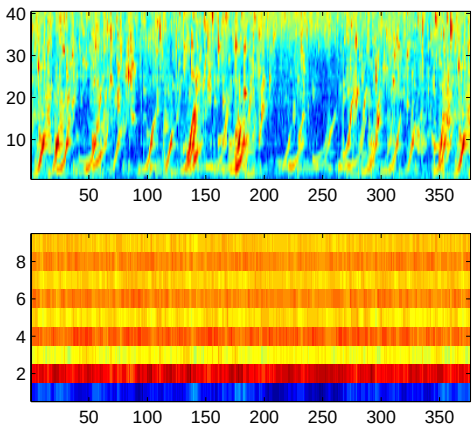
(Adam Berenzweig, Lawrence/NECI)

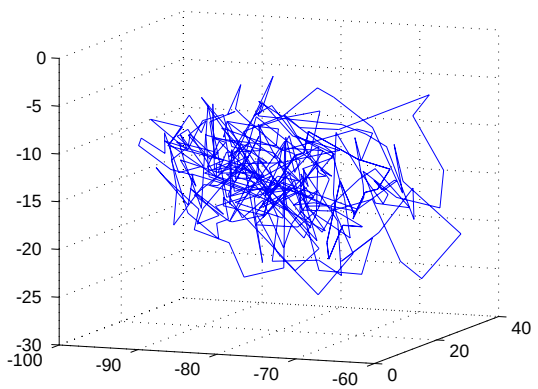


Sound texture modeling

(Marios Athineos)

- **Sound textures are important but neglected**
  - fire, rain, paper: no clear pitch, onsets, shape
  - typically modeled with filtered white noise
- **LPC on *spectrum* captures temporal structure:**
  - better parameterization of texture structure?



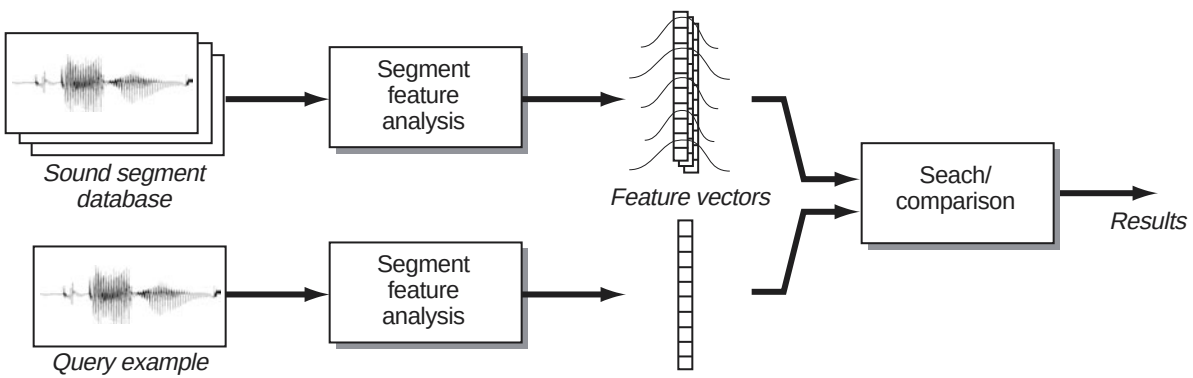


- generate variations?

Audio Information Retrieval

(Manuel Reyes)

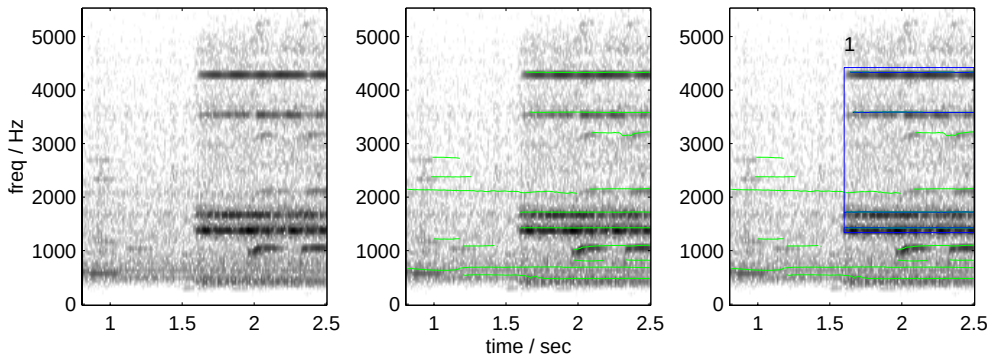
- **Searching in a database of audio**
  - speech .. use ASR
  - text annotations .. search them
  - sound effects library?
- **e.g. Muscle Fish "SoundFisher" browser**
  - define multiple 'perceptual' feature dimensions
  - search by proximity in (weighted) feature space



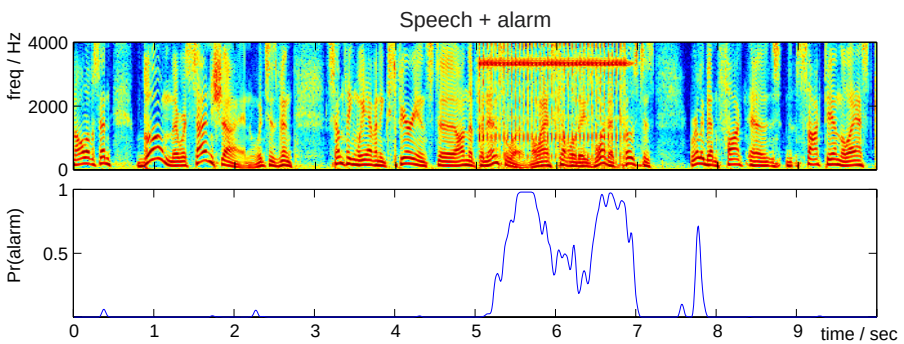
  - features are 'global' for each soundfile, no attempt to separate mixtures

Alarm sound detection

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



  - sinusoid peaks have invariant properties



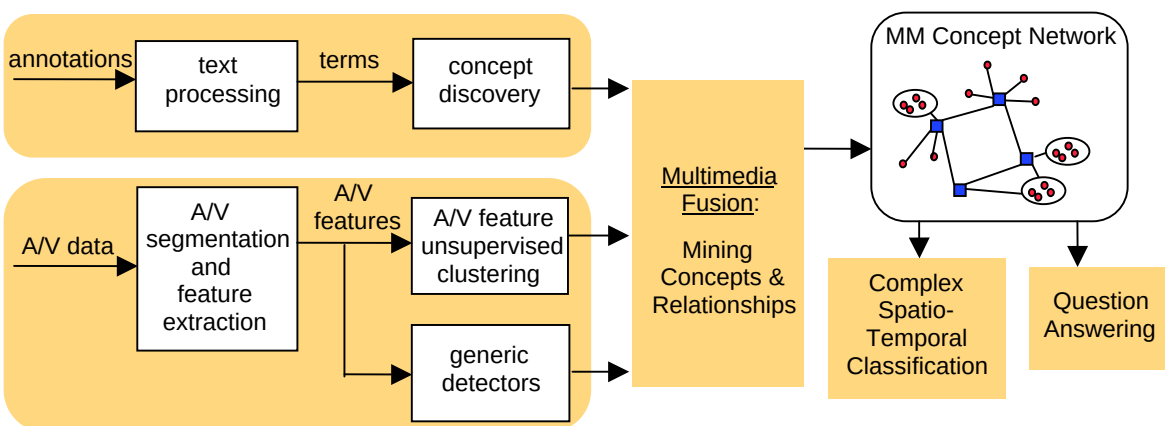
  - cepstral coefficients are easy to model

2

Future work:
Automatic audio-video analysis

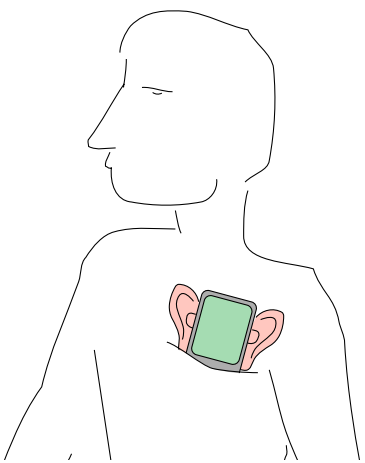
(Shih-Fu Chang, Kathy McKeown)

- **Documentary archive management**
  - huge ratio of raw-to-finished material
  - costly manual logging
- **Problem: term ↔ signal mapping**
  - training corpus of past annotations
  - interactive semi-automatic learning



The ‘Machine listener’

- **Goal: An auditory system for machines**
  - use same environmental information as people
- **Signal understanding**
  - monitor for particular sounds
  - real-time description
- **Scenarios**



  - personal listener → summary of your day
  - future prosthetic hearing device
  - autonomous robots

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

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio