Sound, Mixtures, and Learning: LabROSA overview

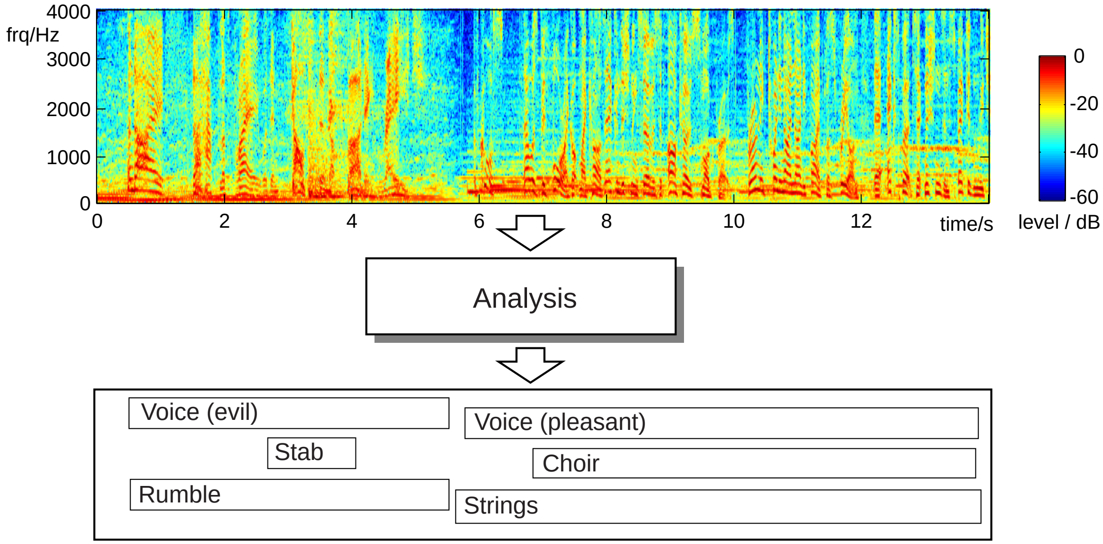
Dan Ellis
<dpwe@ee.columbia.edu>

Laboratory for Recognition and Organization of Speech and Audio
Electrical Engineering Dept., Columbia University, New York
http://labrosa.ee.columbia.edu/

Outline

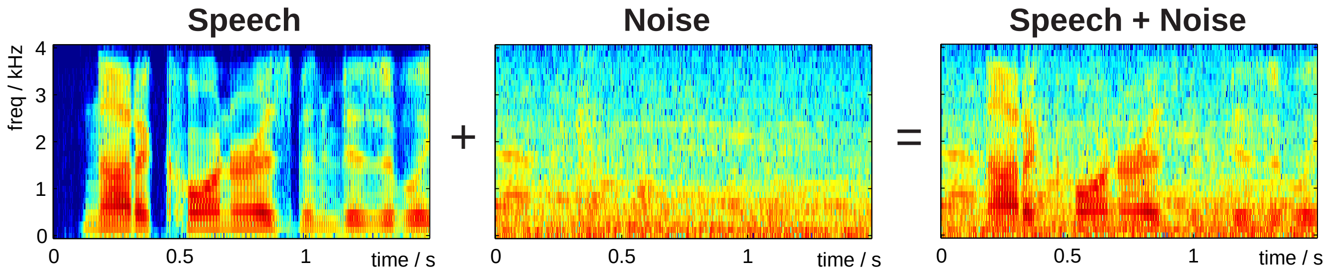
1. **Auditory Scene Analysis**
2. **Speech Recognition & Mixtures**
3. **Music Analysis & Similarity**
4. **General Sound Organization**
5. **Future Work**

1 Auditory Scene Analysis



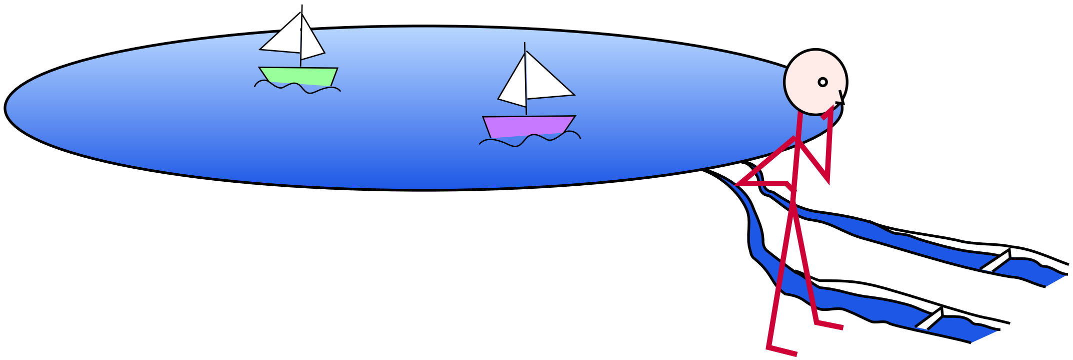
- ***Auditory Scene Analysis*: describing a complex sound in terms of high-level sources/events**
  - ... like listeners do
- **Hearing is *ecologically* grounded**
  - reflects ‘natural scene’ properties
  - subjective, not absolute

Sound, mixtures, and learning



- **Sound**
  - carries useful information about the world
  - complements vision
- **Mixtures**
  - .. are the rule, not the exception
  - medium is 'transparent', sources are many
  - must be handled!
- **Learning**
  - the 'speech recognition' lesson:
    let the data do the work
  - like listeners

The problem with recognizing mixtures

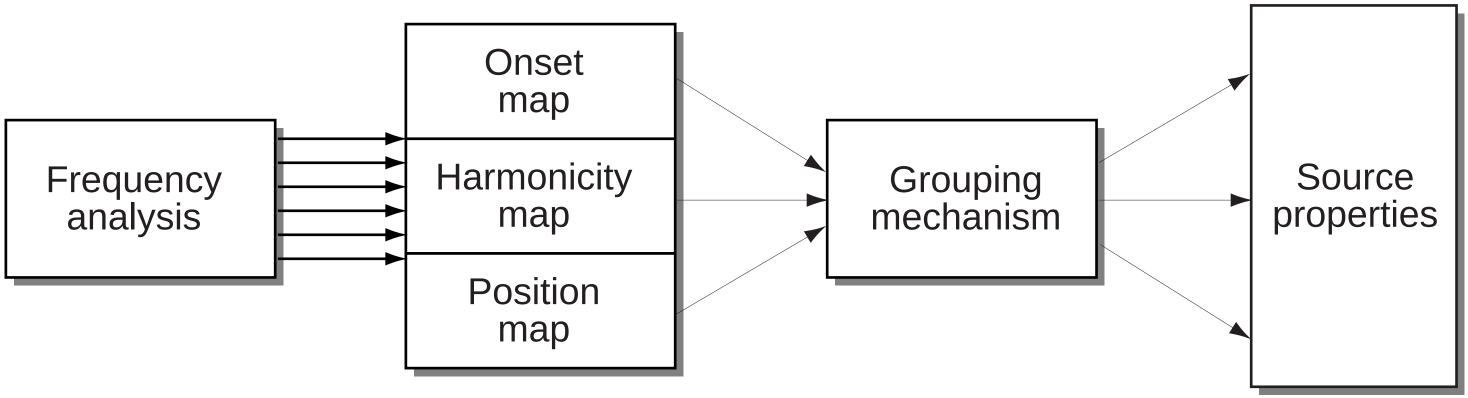


"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- **Received waveform is a mixture**
  - two sensors, N signals ... *underconstrained*
- **Disentangling mixtures as the primary goal?**
  - perfect solution is not possible
  - need experience-based *constraints*

Human Auditory Scene Analysis
(Bregman 1990)

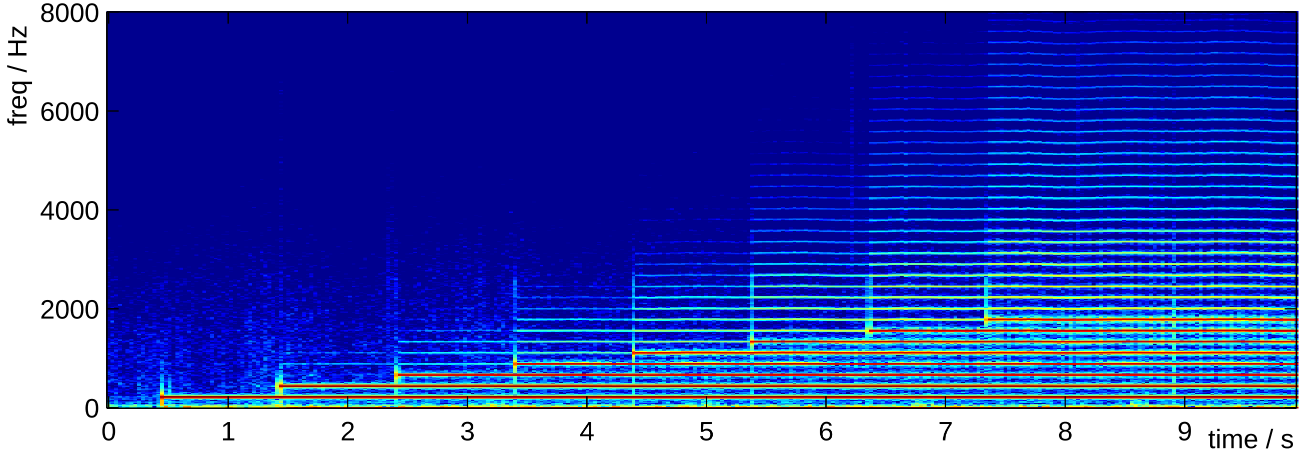
- **How do people analyze sound mixtures?**
  - break mixture into small *elements* (in time-freq)
  - elements are *grouped* in to sources using *cues*
  - sources have aggregate *attributes*
- **Grouping 'rules' (Darwin, Carlyon, ...):**
  - cues: common onset/offset/modulation, harmonicity, spatial location, ...



(after Darwin, 1996)

Cues to simultaneous grouping

- **Elements + attributes**

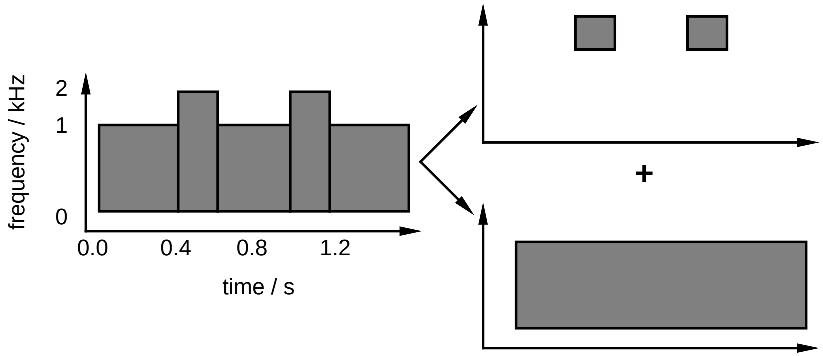


- **Common onset**
  - simultaneous energy has common source
- **Periodicity**
  - energy in different bands with same cycle
- **Other cues**
  - spatial (ITD/IID), familiarity, ...

The effect of context

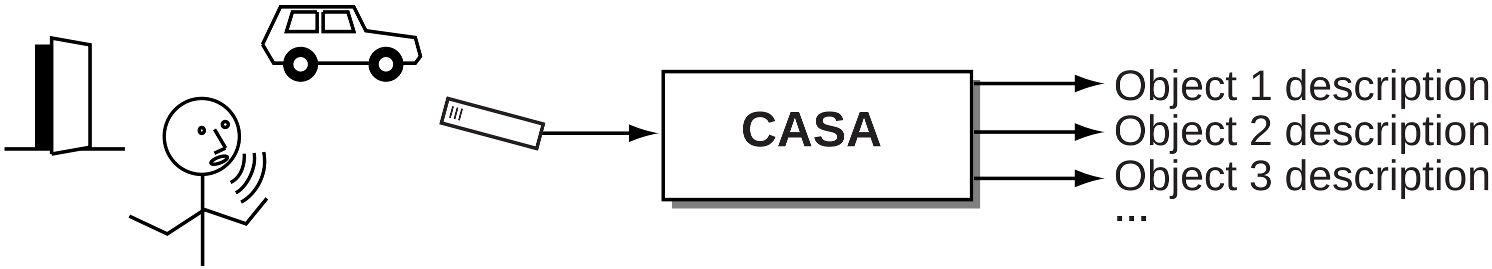
- **Context can create an 'expectation': i.e. a bias towards a particular interpretation**
- **e.g. Bregman's "old-plus-new" principle:**

  A change in a signal will be interpreted as an *added* source whenever possible



  - a different division of the same energy depending on what preceded it

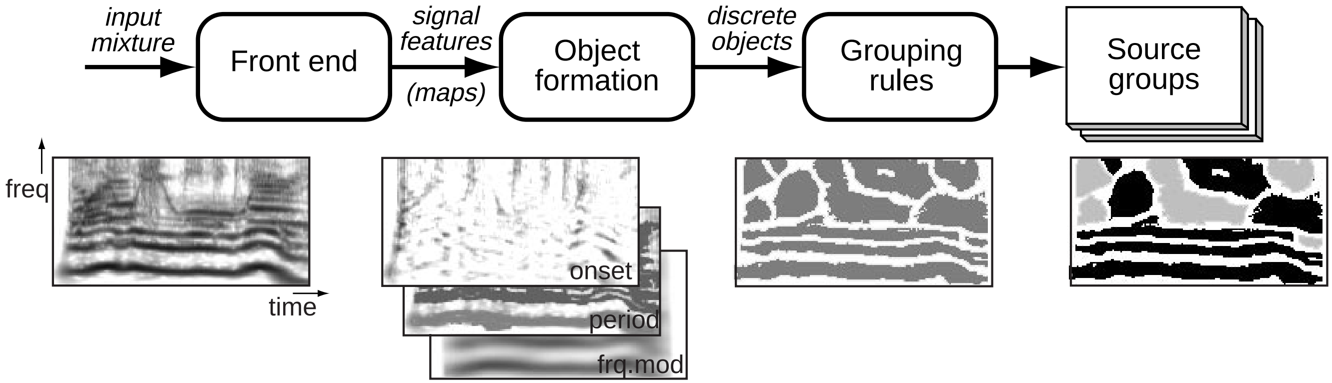
Computational Auditory Scene Analysis (CASA)



- **Goal: Automatic sound organization ; Systems to ‘pick out’ sounds in a mixture**
  - ... like people do
- **E.g. voice against a noisy background**
  - to improve speech recognition
- **Approach:**
  - psychoacoustics describes grouping ‘rules’
  - ... just implement them?

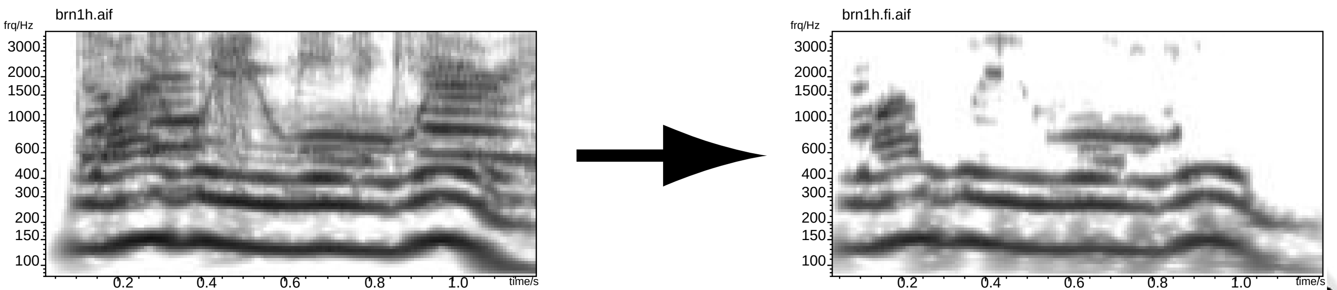
The Representational Approach
(Brown & Cooke 1993)

- **Implement psychoacoustic theory**



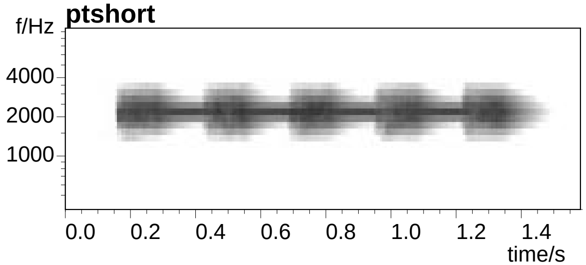
  - ‘bottom-up’ processing
  - uses common onset & periodicity cues

- **Able to extract voiced speech:**

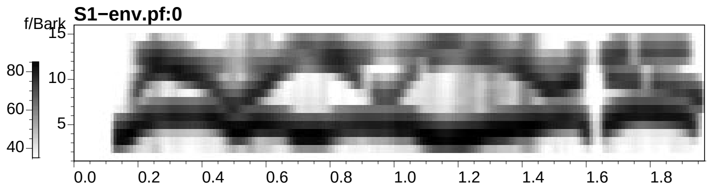


Restoration in sound perception

- **Auditory ‘illusions’ = hearing what’s not there**
- **The continuity illusion**



- **SWS**



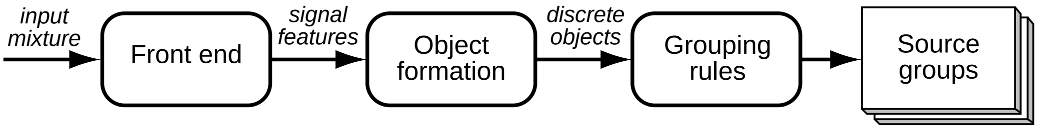
  - duplex perception

- **How to model in CASA?**

Adding top-down constraints

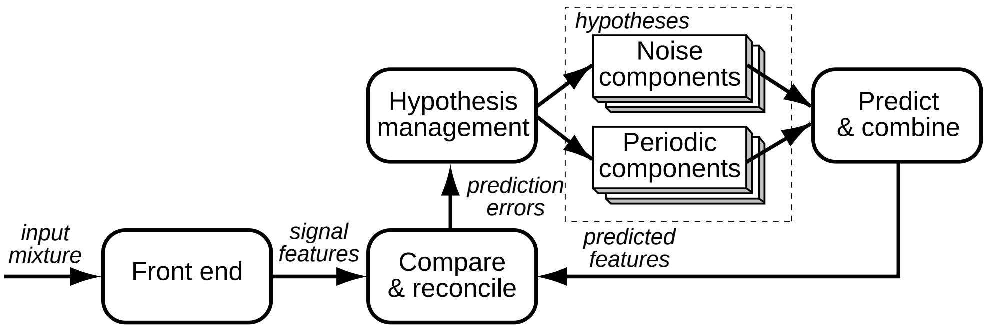
Perception is not *direct*
but a *search* for *plausible hypotheses*

- **Data-driven (bottom-up)...**



  - objects irresistibly appear

vs. Prediction-driven (top-down)

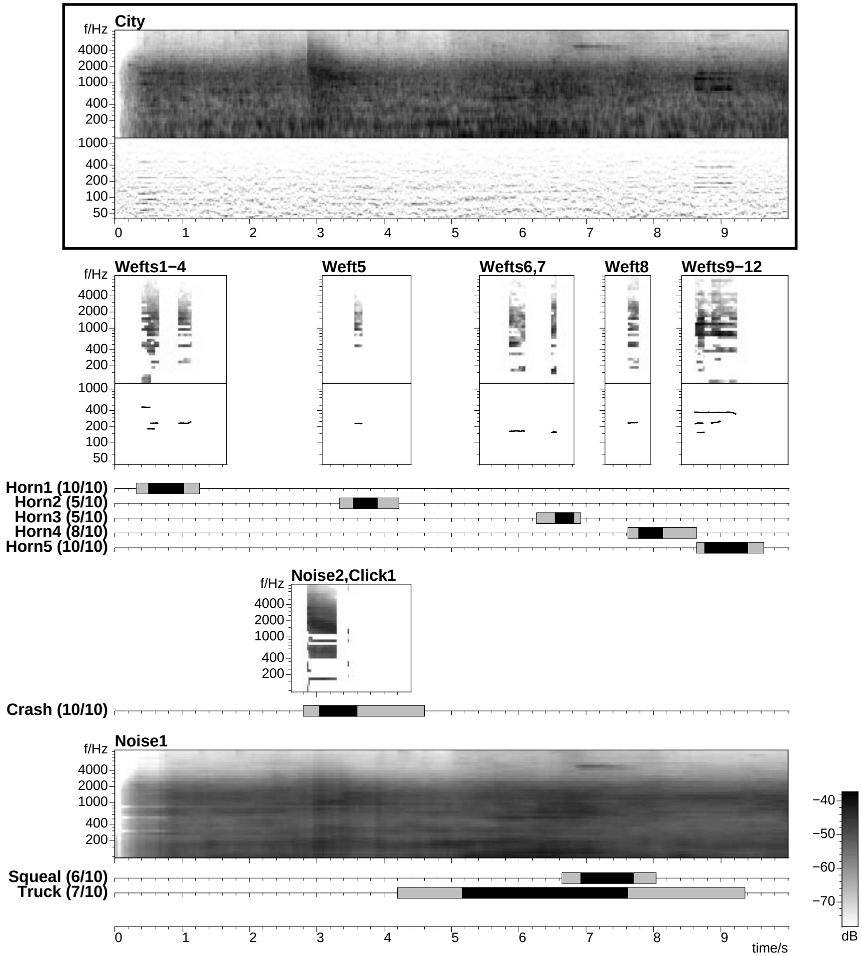


  - match observations
    with parameters of a world-model
  - need world-model constraints...

Prediction-Driven CASA

(Ellis 1996)

- **Explain a complex sound with basic elements**



Approaches to sound mixture recognition

- **Recognize combined signal**
  - ‘multicondition training’
  - combinatorics..
- **Separate signals**
  - e.g. CASA, ICA
  - nice, if you can do it
- **Segregate features into fragments**
  - then missing-data recognition

Aside: Evaluation

- **Evaluation is a big problem for CASA**
  - what is the goal, really?
  - what is a good test domain?
  - how do you measure performance?
- **SNR improvement**
  - not easy given only before-after signals: correspondence problem
  - can do with fixed filtering mask; rewards removing signal as well as noise
- **ASR improvement**
  - recognizers typically very sensitive to artefacts
- **'Real' task?**
  - mixture corpus with specific sound events...

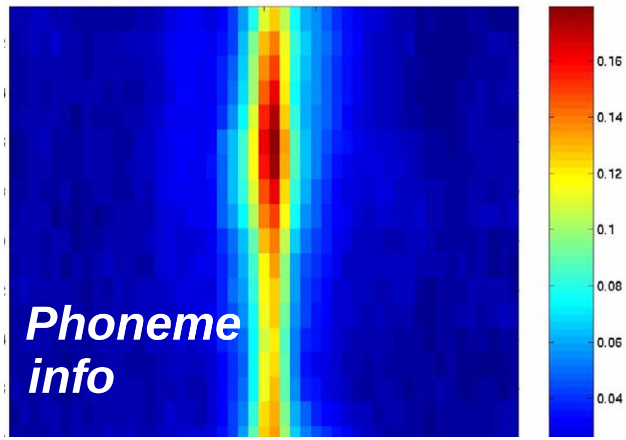
Outline

1. Auditory Scene Analysis
2. **Speech Recognition & Mixtures**
   - the information in speech
   - Meeting Recorder project
   - speech fragment decoding
3. Music Analysis & Similarity
4. General Sound Organization
5. Future Work

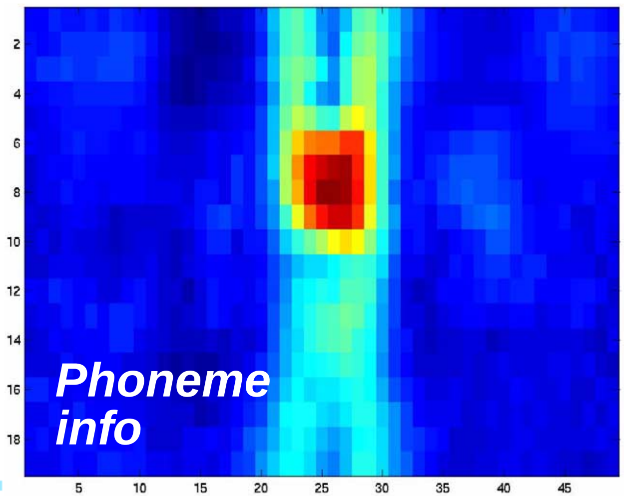
2 The information in speech

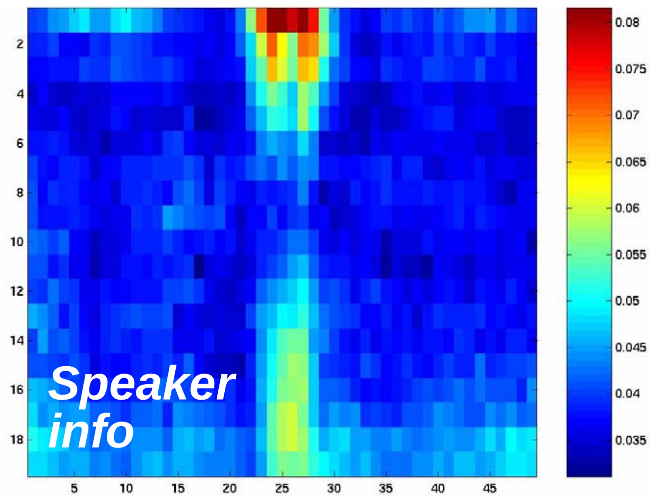
(Patricia Scanlon)

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



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

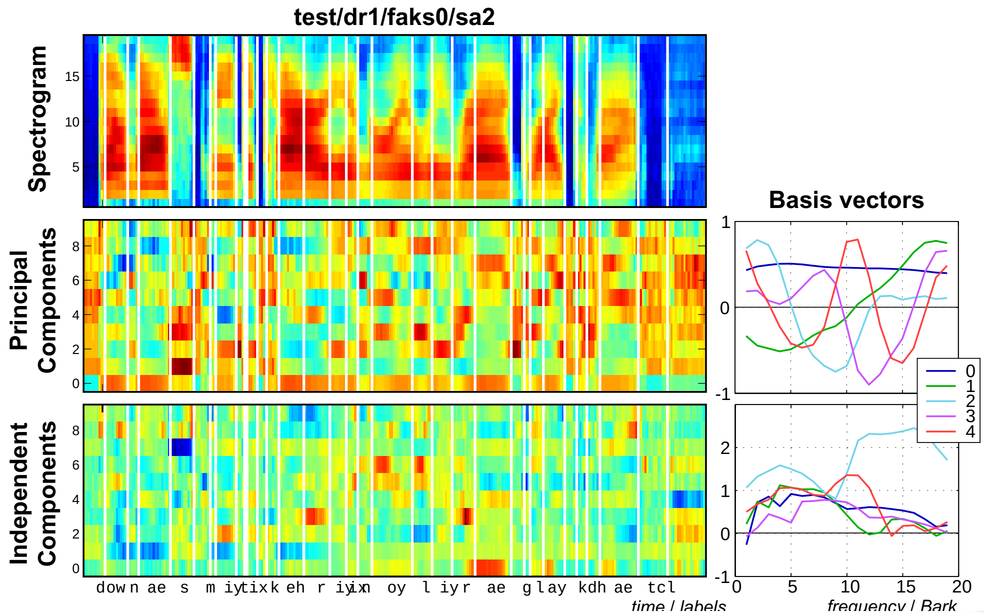




The best subword units?

(Eric Fosler)

- **Speech recognizers typically use phonemes**
  - inherited from linguistics
- **Alternative approach is ‘articulatory features’**
  - orthogonal attributes defining subwords
- **Can we infer a feature set from the data**
  - using e.g. Independent Component Analysis



The Meeting Recorder Project

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

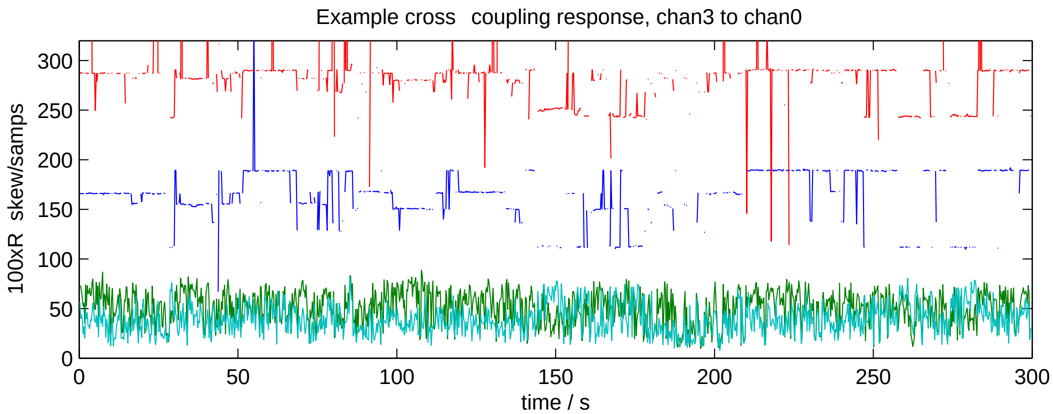
- **Microphones in conventional meetings**
  - for summarization/retrieval/behavior analysis
  - informal, overlapped speech
- **Data collection (ICSI, UW, IDIAP):**

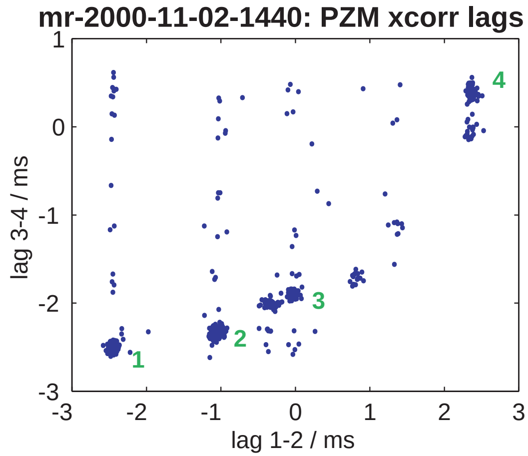
  - 100 hours collected, ongoing transcription
- **NSF 'Mapping Meetings' project**
  - also interest from NIST, DARPA, EU

Speaker Turn detection

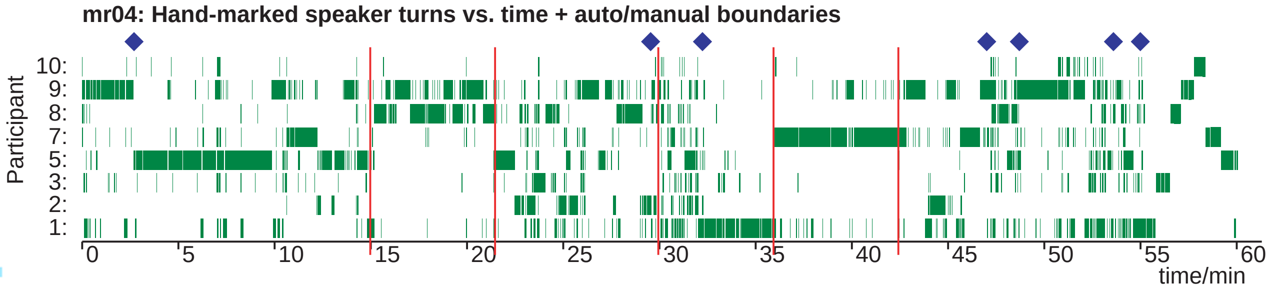
(Huan Wei Hee, Jerry Liu)

- **Acoustic:**
  Triangulate tabletop mic timing differences
  - use normalized peak value for confidence





- **Behavioral: Look for patterns of speaker turns**



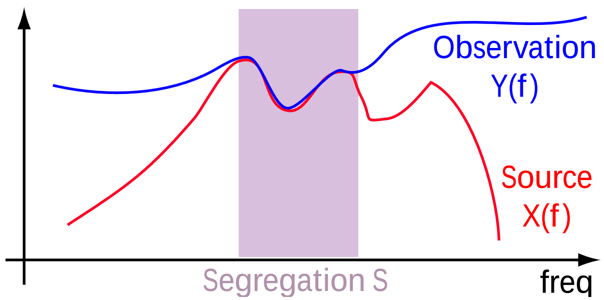
Speech Fragment recognition

(Barker & Cooke/Sheffield)

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

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

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



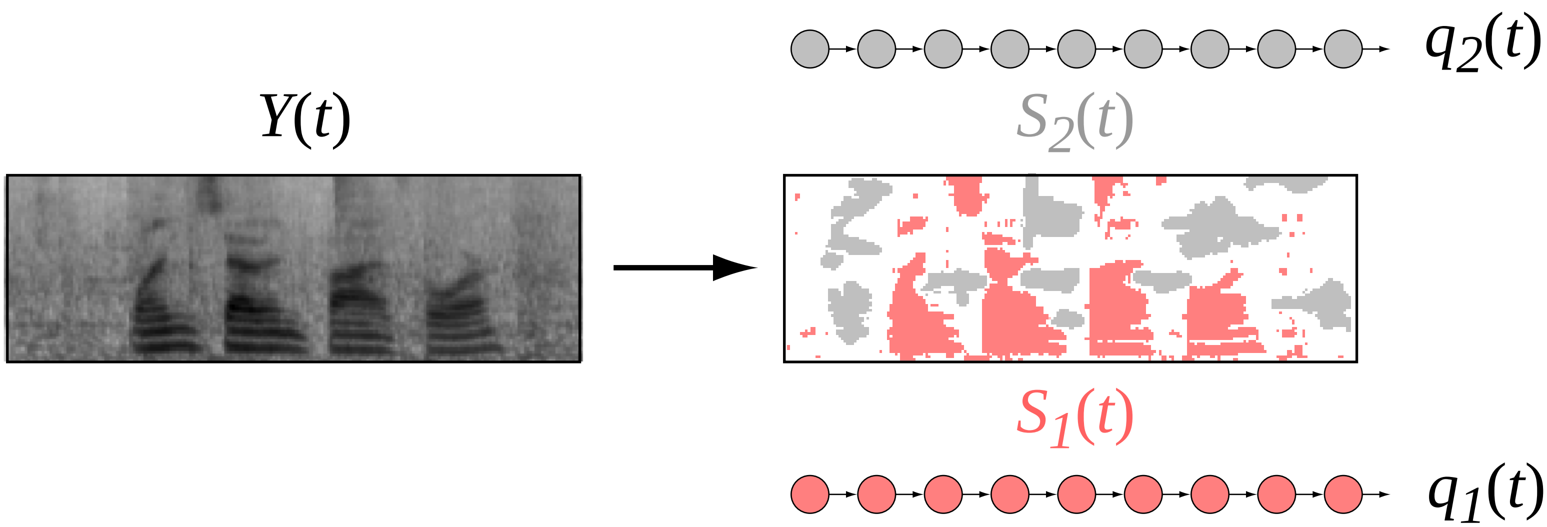
  - *spectral features* allow clean relationship

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

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

Multi-source decoding

- **Search for more than one source**



- **Mutually-dependent data masks**
- **Use e.g. CASA features to propose masks**
  - locally coherent regions
- **Theoretical vs. practical limits**

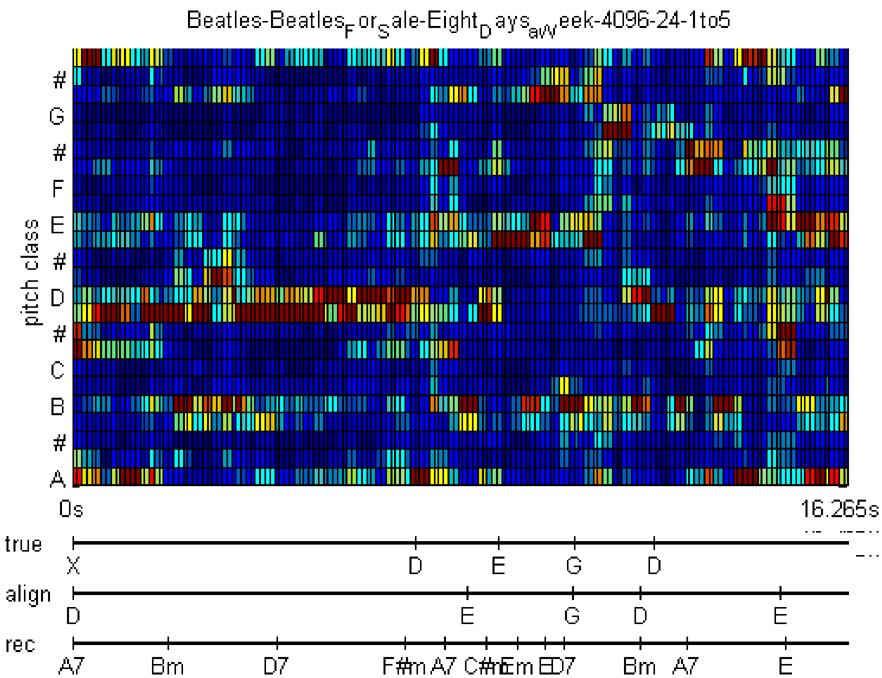
Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. **Music Analysis & Similarity**
   - musical structure analysis
   - similarity browsing
4. General Sound Organization
5. Future Work

Music Structure Analysis

(Alex Sheh)

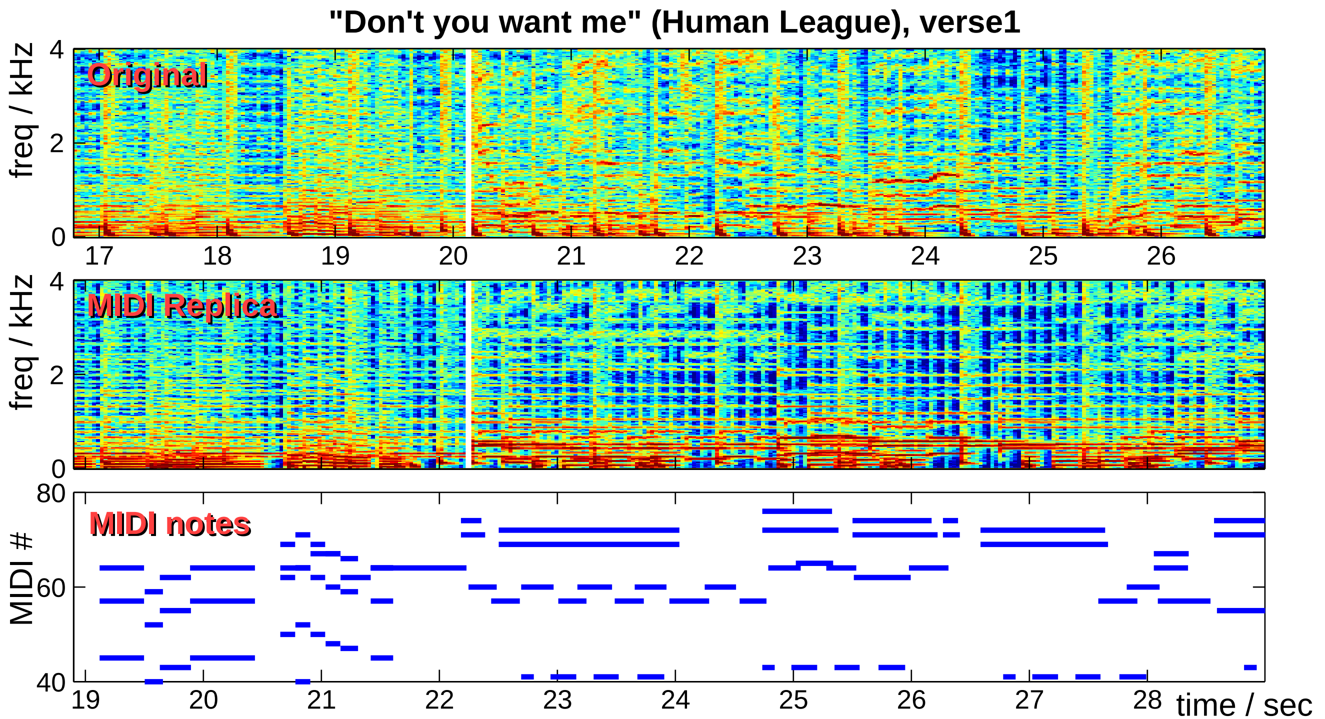
- **Fine-level information from music**
  - for searching
  - for modeling/statistics
- **e.g. Chord sequences via PCPs :**



Ground truth for Music Recordings

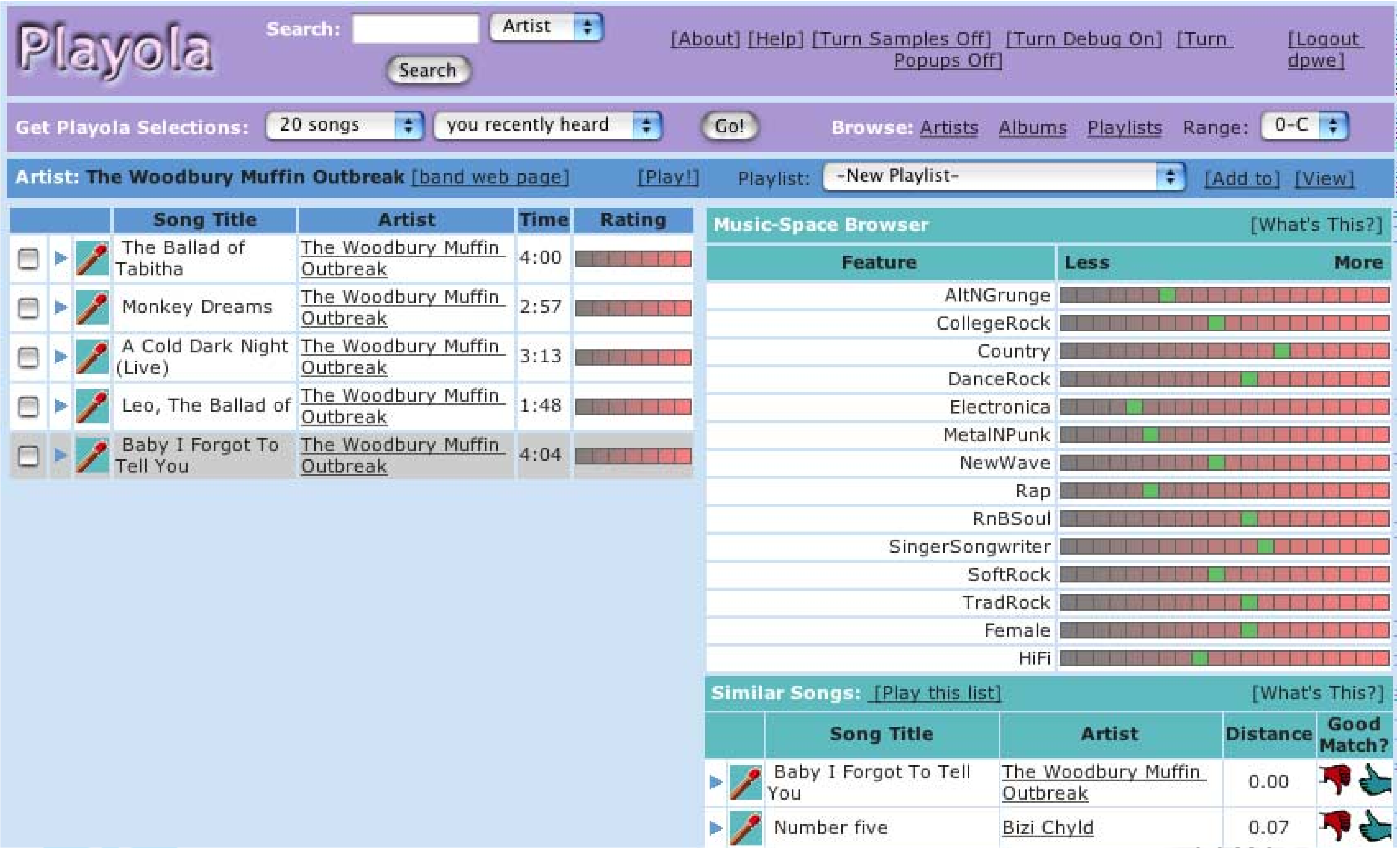
(Rob Turetsky)

- **Machine Learning algorithms need labels**
  - but real recordings don't have labels
- **MIDI 'replicas' exist**
- **Alignment locates MIDI notes in real sound:**



Music Similarity Browsing
(Adam Berenzweig)

- **'Anchor models' : music on subjective axes**



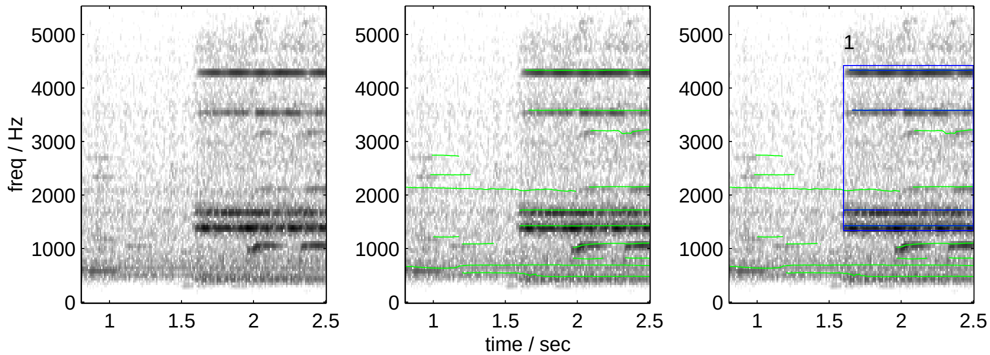
Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. Music Analysis & Similarity
4. **General Sound Organization**
   - alarm detection
   - sound texture modeling
   - recognition of multiple sources
5. Future Work

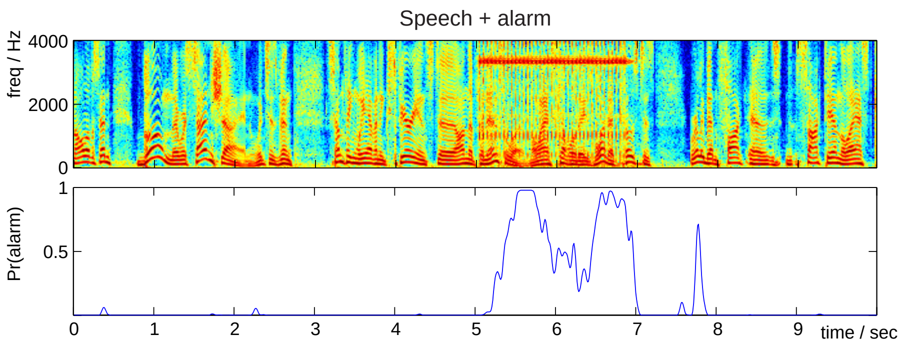
4

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

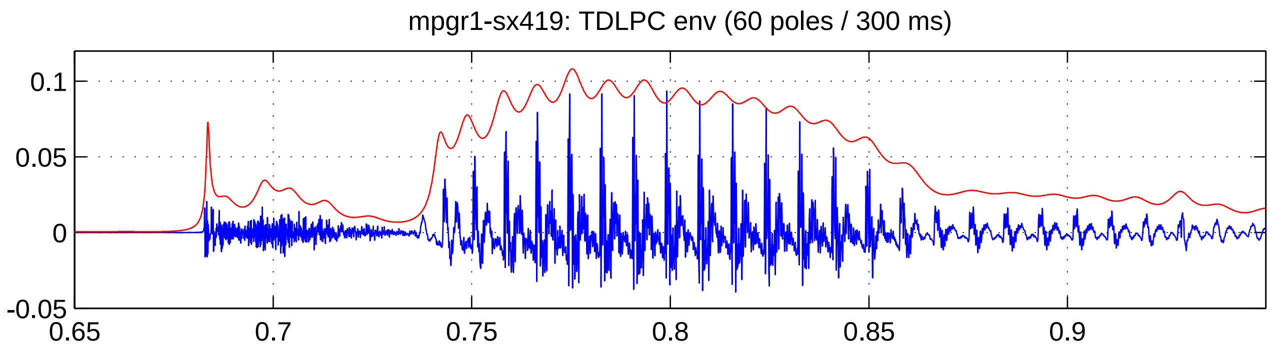


  - cepstral coefficients are easy to model

Sound Texture Modeling

(Marios Athineos)

- **Best sound models are based on sinusoids**
  - noise residual modeled quite simply
- **Noise 'textures' have extra temporal structure**
  - need a more detailed model
- **Linear prediction of spectrum defines a parametric temporal envelope:**



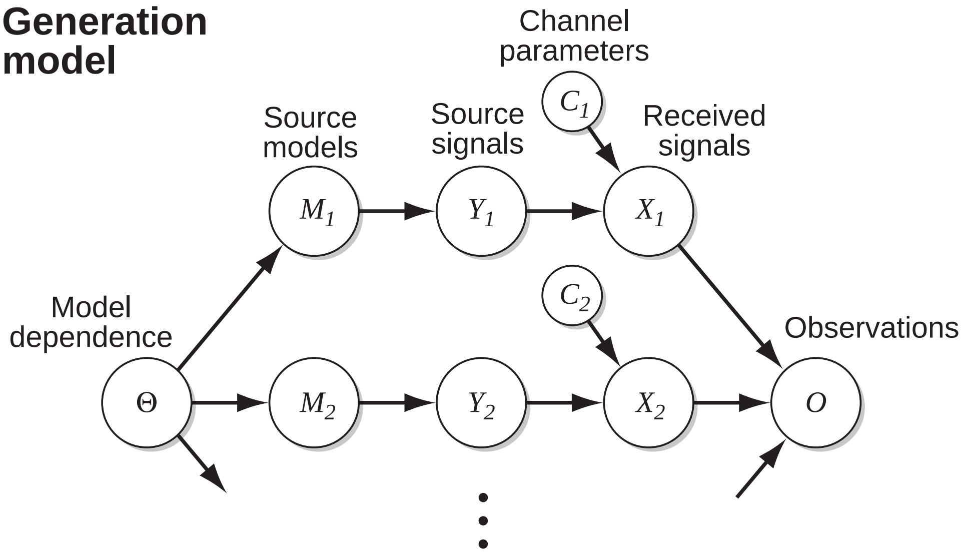
- **High-quality noise-excited resynthesis:**
  - original - resynth - x2 TSM - c/w PVOC

Sound mixture decomposition

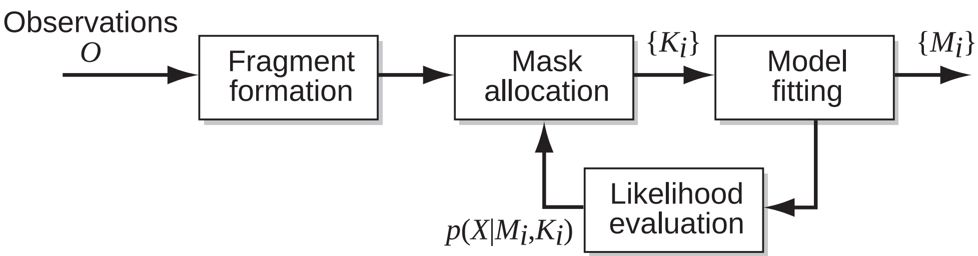
(Manuel Reyes)

- **Full or approximate Bayesian inference to model multiple, independent sound sources:**

Generation model



Analysis structure



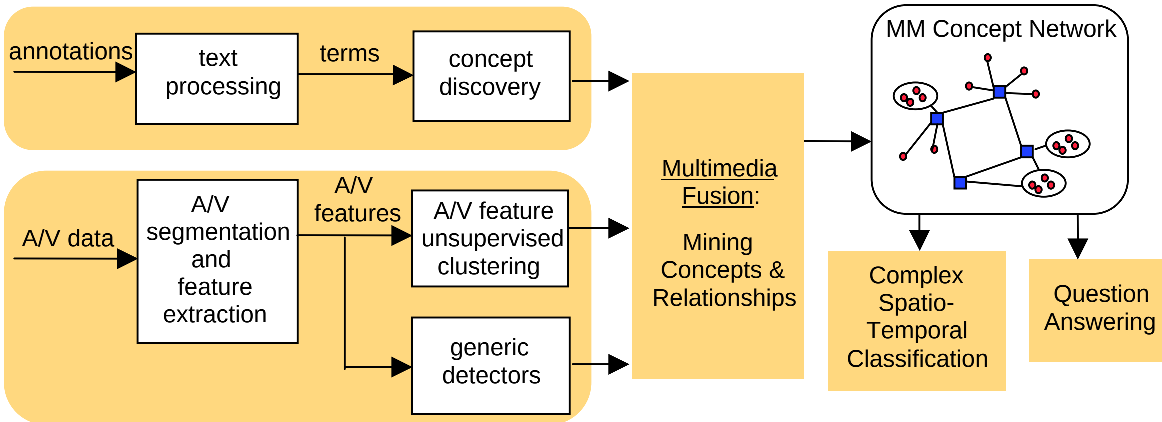
Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. Music Analysis & Similarity
4. General Sound Organization
5. **Future Work**
   - audio-visual information
   - real-world sound indexing

Future work: Automatic audio-video analysis

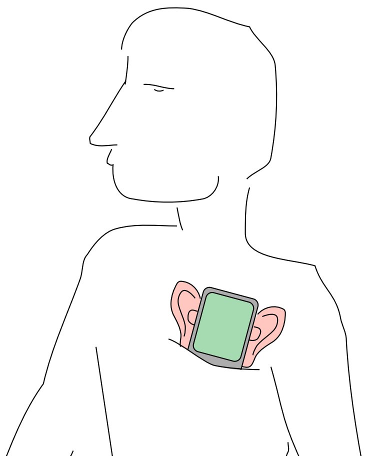
(Shih-Fu Chang, Kathy McKeown)

- **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 ‘Listening Machine’

- **Smart PDA records everything**
- **Only useful if we have index, summaries**
  - monitor for particular sounds
  - real-time description
- **Scenarios**



  - personal listener → summary of your day
  - future prosthetic hearing device
  - autonomous robots
- **Meeting data, ambulatory audio**

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