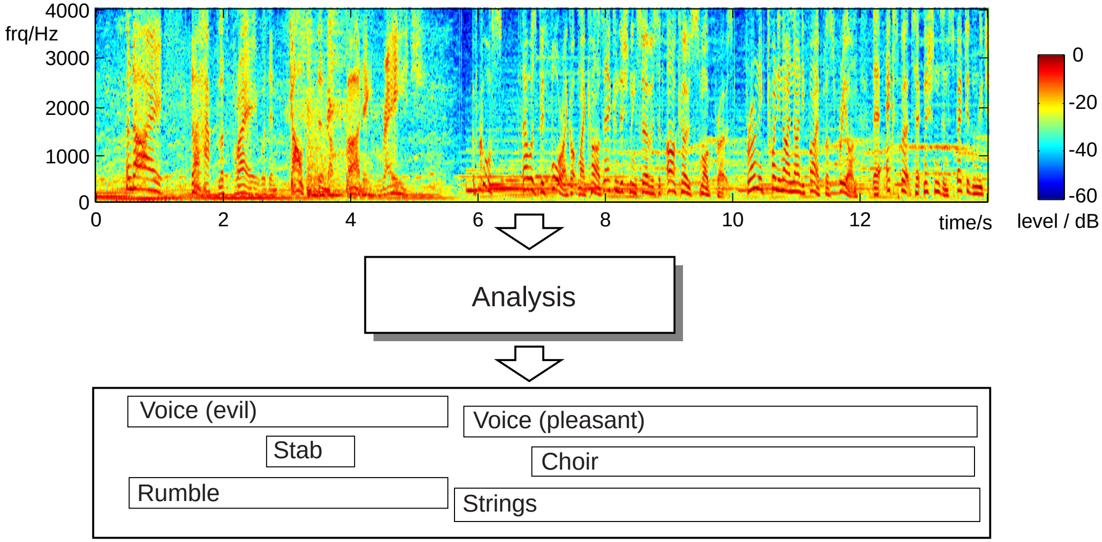
Sound, Mixtures, and Learning: LabROSA overview

1. **Sound Content Analysis**
2. **Recognizing sounds**
3. **Organizing mixtures**
4. **Accessing large datasets**

Dan Ellis <dpwe@ee.columbia.edu>

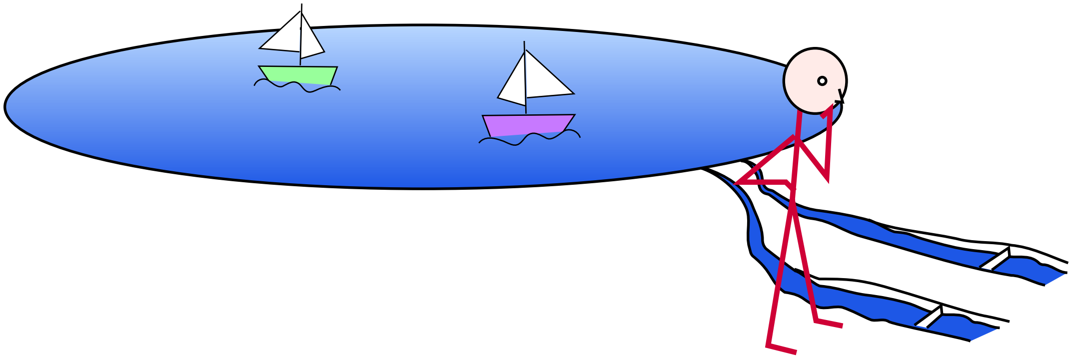
Laboratory for Recognition and Organization of Speech and Audio
(LabROSA)
Columbia University, New York
http://labrosa.ee.columbia.edu/

1 Sound Content Analysis



- **Sound understanding: the key challenge**
  - what listeners do
  - understanding = abstraction
- **Applications**
  - indexing/retrieval
  - robots
  - prostheses

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)

- **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 = constraints
  - subjective, not absolute

Approaches to handling sound mixtures

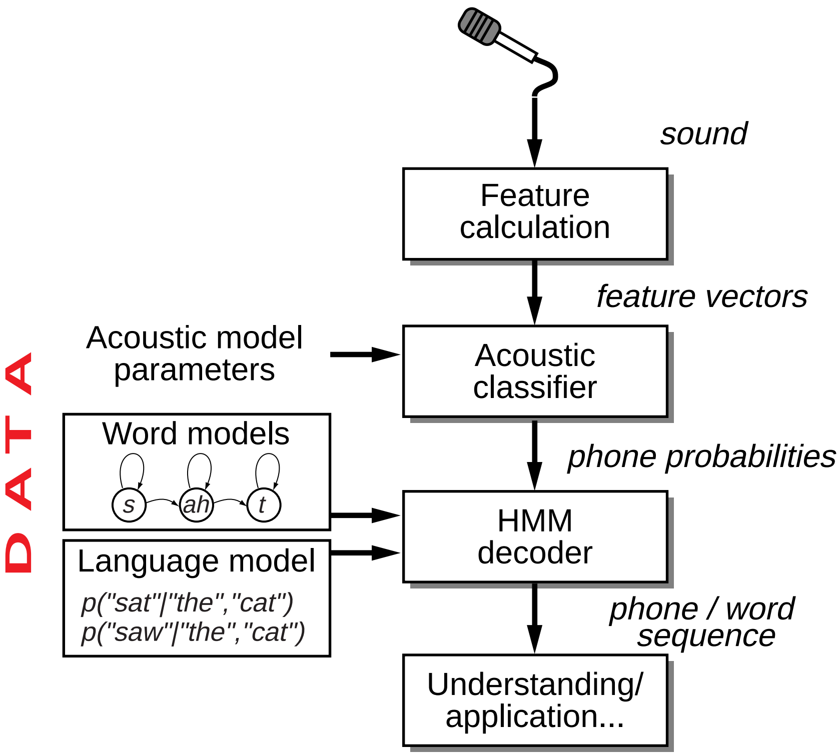
- **Separate signals, then recognize**
  - e.g. CASA, ICA
  - nice, if you can do it
- **Recognize combined signal**
  - 'multicondition training'
  - combinatorics..
- **Recognize with parallel models**
  - full joint-state space?
  - or: divide signal into fragments, then use missing-data recognition

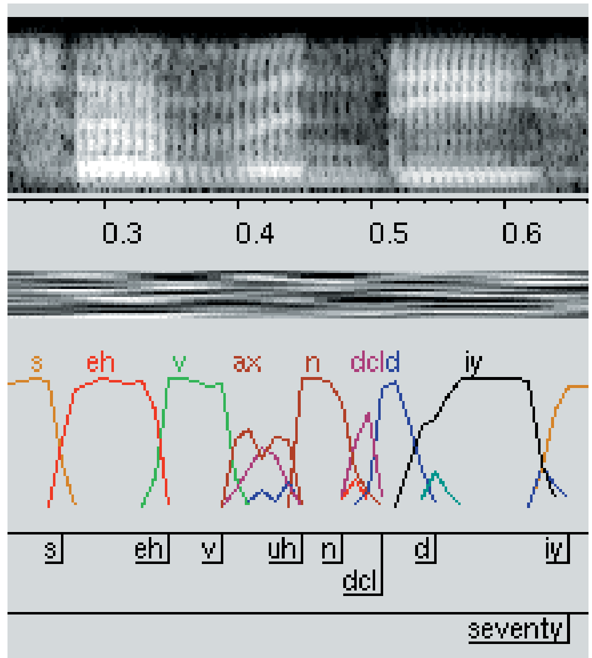
Outline

1. Sound Content Analysis
2. **Recognizing sounds**
   - Speech recognition
   - Nonspeech
3. Organizing mixtures
4. Accessing large datasets

2 Recognizing Sounds: Speech

- **Standard speech recognition structure:**



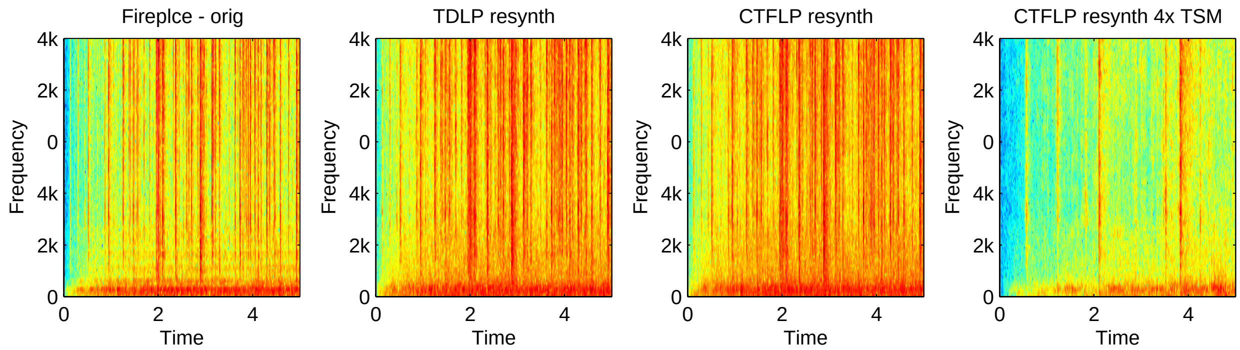


- **How to handle additive noise?**
    - just train on noisy data: 'multicondition training'

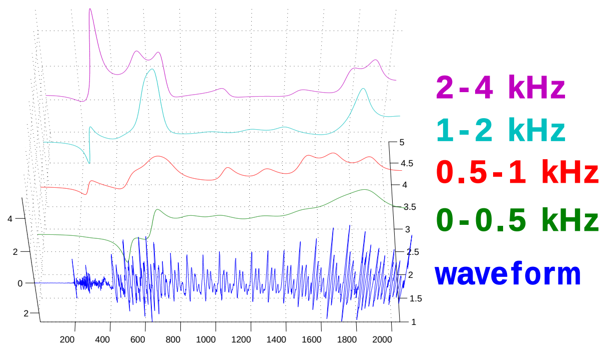
Novel speech signal representations

(with Marios Athineos)

- **Common sound models use 10ms frames**
  - but: sub-10ms envelope is perceptible



- **Use a parametric (LPC) model on spectrum**

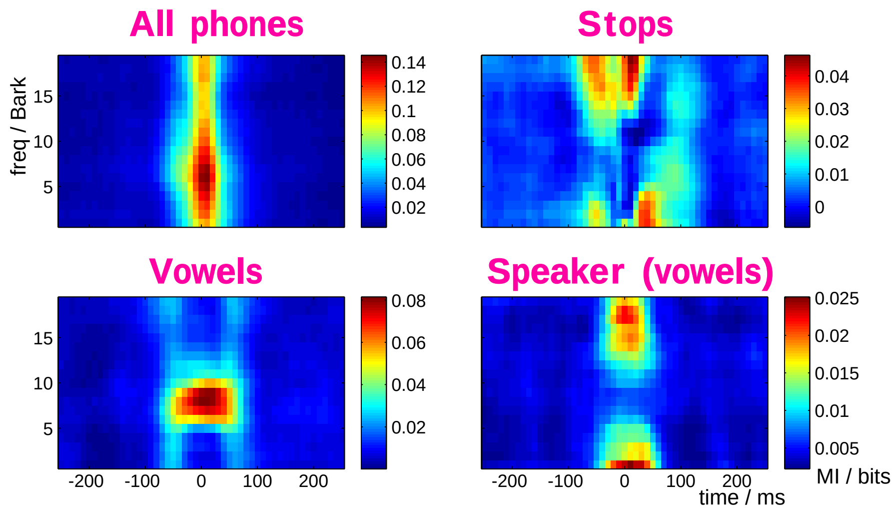


- **Convert to features for ASR**
  - improvements esp. for stops

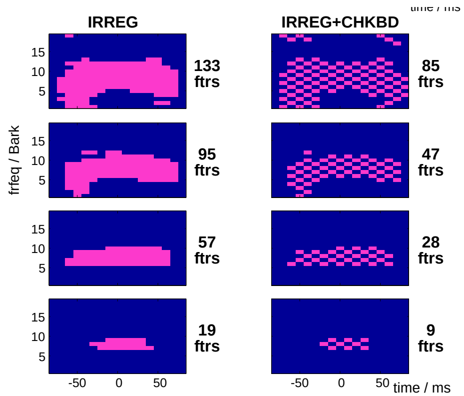
Finding the Information in Speech

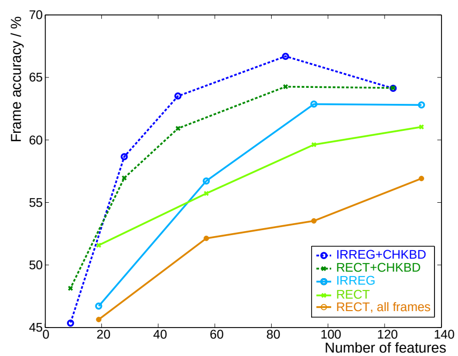
(with Patricia Scanlon)

- **Mutual Information in time-frequency:**



- **Use to select classifier input features**

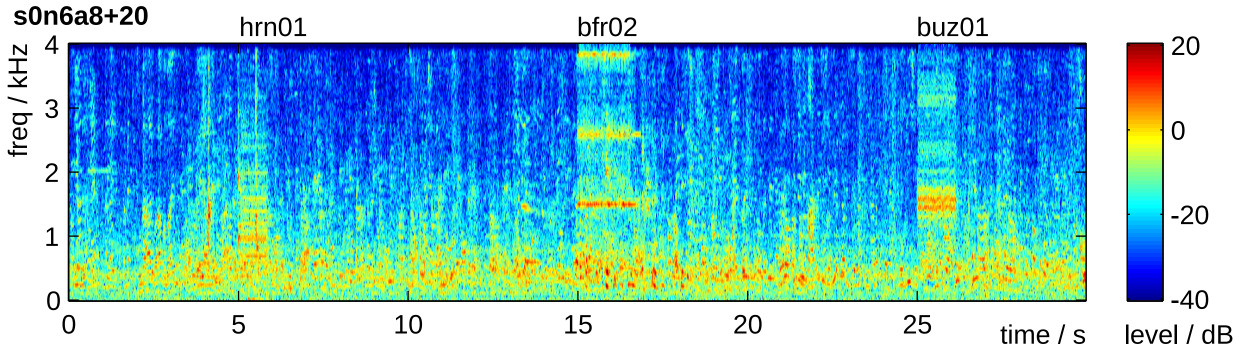




Alarm sound detection

(Ellis 2001)

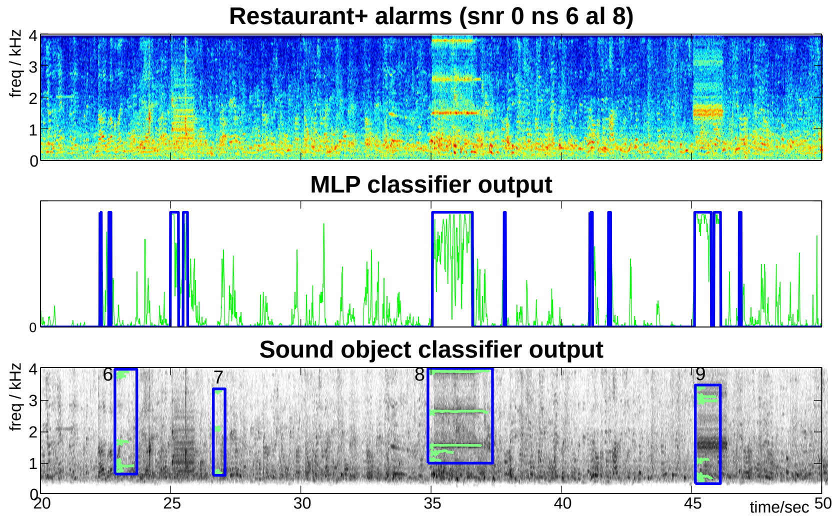
- **Alarm sounds have particular structure**
  - people ‘know them when they hear them’
  - clear even at low SNRs



- **Why investigate alarm sounds?**
  - they’re supposed to be easy
  - potential applications...

- **Contrast two systems:**
  - standard, global features, $P(X|M)$
  - sinusoidal model, fragments, $P(M,S|Y)$

Alarms: Results



- **Both systems commit many insertions at 0dB SNR, but in different circumstances:**

Noise	*Neural net system* *Del*	*Ins*	*Tot*	*Sinusoid model system* *Del*	*Ins*	*Tot*
1 (amb)	7 / 25	2	36%	14 / 25	1	60%
2 (bab)	5 / 25	63	272%	15 / 25	2	68%
3 (spe)	2 / 25	68	280%	12 / 25	9	84%
4 (mus)	8 / 25	37	180%	9 / 25	135	576%
Overall	**22 / 100**	170	**192%**	**50 / 100**	147	**197%**

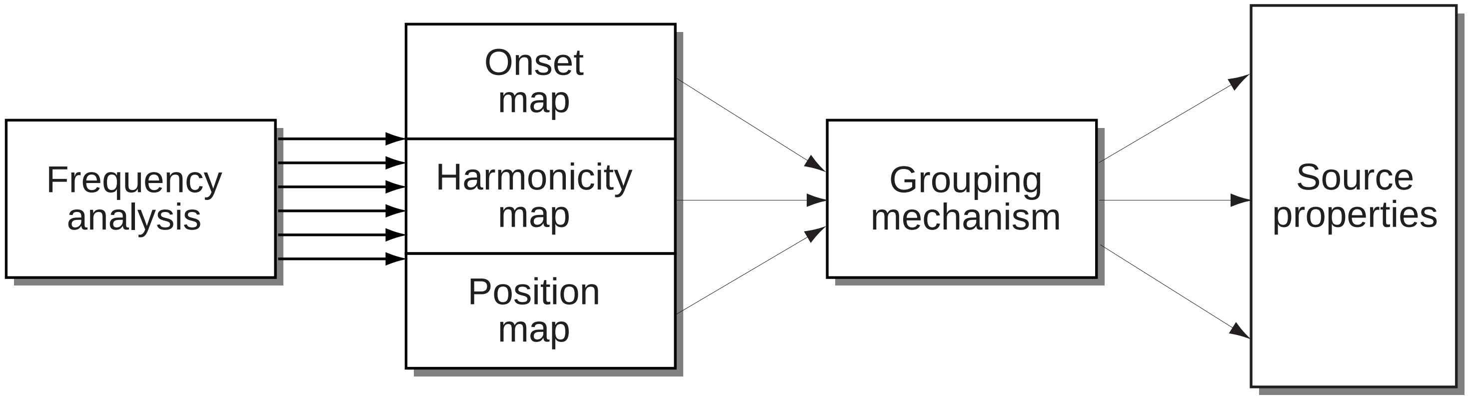
Outline

1. **Sound Content Analysis**
2. **Recognizing sounds**
3. **Organizing mixtures**
   - Auditory Scene Analysis
   - Missing data recognition
   - Parallel model inference
4. **Accessing large datasets**

3 Auditory Scene Analysis

(Bregman 1990)

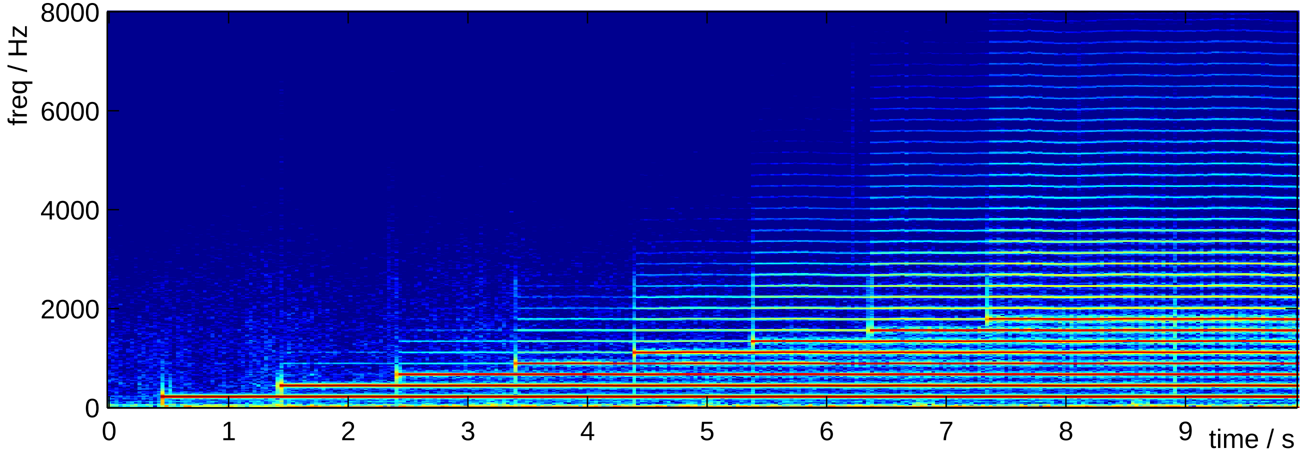
- **How do people analyze sound mixtures?**
  - break mixture into small *elements* (in time-freq)
  - elements are *grouped* in to sources using *cues*
  - sources have aggregate *attributes*
- **Grouping 'rules' (Darwin, Carlyon, ...):**
  - cues: common onset/offset/modulation, harmonicity, spatial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



(after Darwin, 1996)

Cues to simultaneous grouping

- **Elements + attributes**

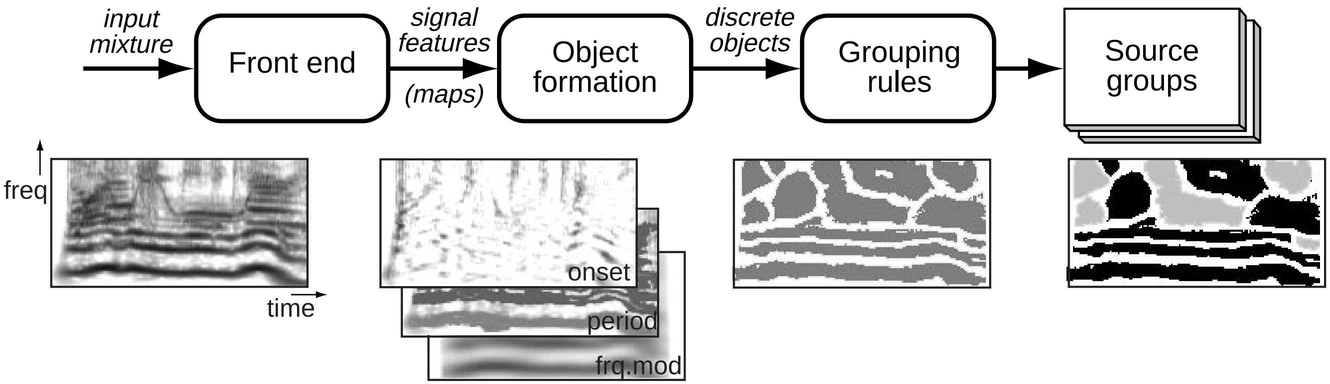


- **Common onset**
  - simultaneous energy has common source
- **Periodicity**
  - energy in different bands with same cycle
- **Other cues**
  - spatial (ITD/IID), familiarity, ...
- **But: Context ...**

Computational Auditory Scene Analysis: The Representational Approach

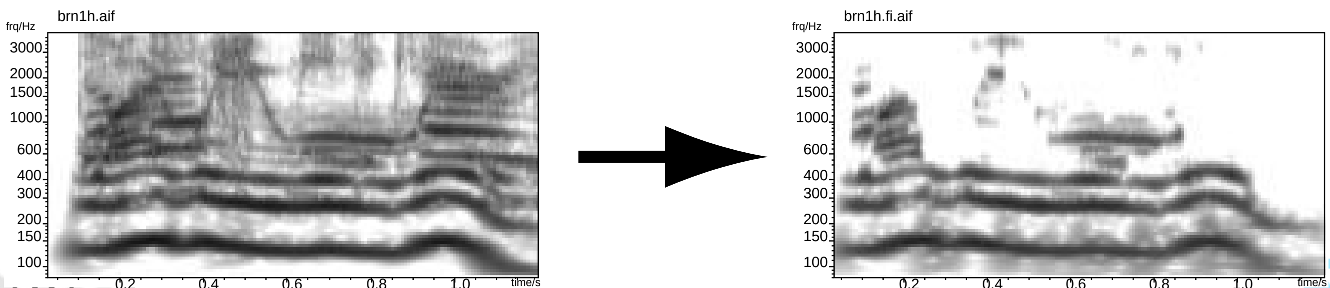
(Cooke & Brown 1993)

- **Direct implementation of psych. theory**



  - 'bottom-up' processing
  - uses common onset & periodicity cues

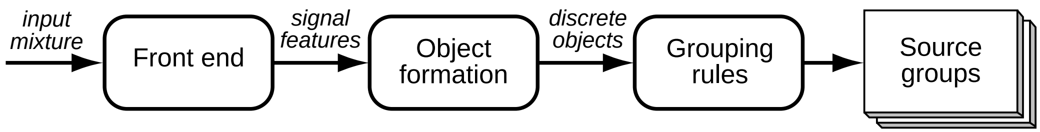
- **Able to extract voiced speech:**



Adding top-down constraints

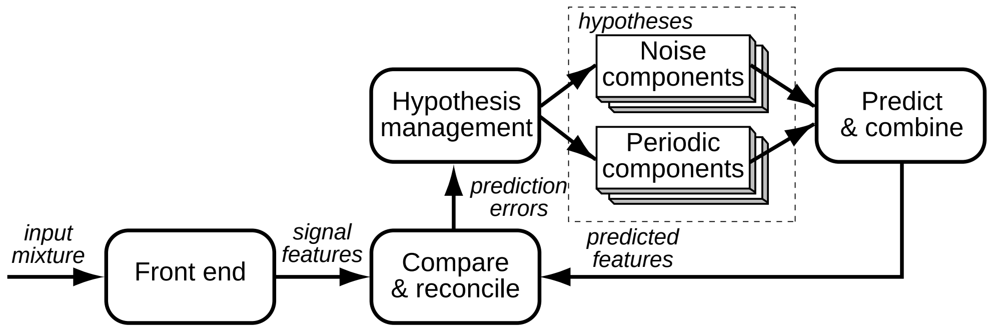
Perception is not direct
but a search for plausible hypotheses

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



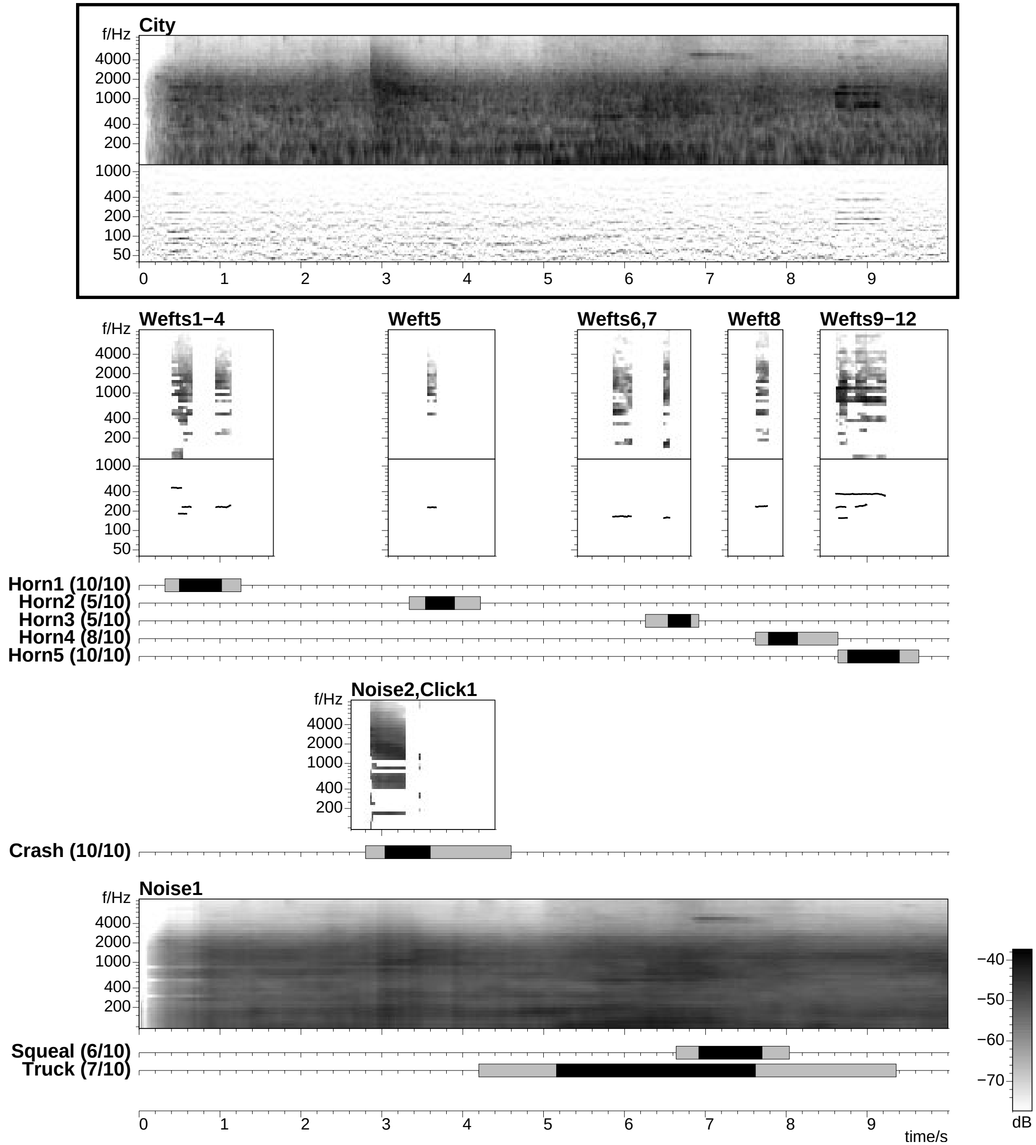
  - objects irresistibly appear

vs. Prediction-driven (top-down)



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

Prediction-Driven CASA



Segregation vs. Inference

- **Source separation requires attribute separation**
  - sources are characterized by attributes (pitch, loudness, timbre + finer details)
  - need to identify & gather different attributes for different sources ...
- **Need representation that segregates attributes**
  - spectral decomposition
  - periodicity decomposition
- **Sometimes values can't be separated**
  - e.g. unvoiced speech
  - maybe infer factors from probabilistic model?

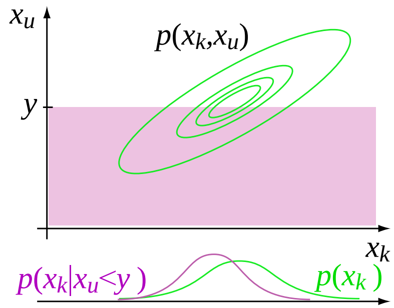
$$p(O, x, y) \rightarrow p(x, y | O)$$

  - or: just skip those values, infer from higher-level context
  - do both: missing-data recognition

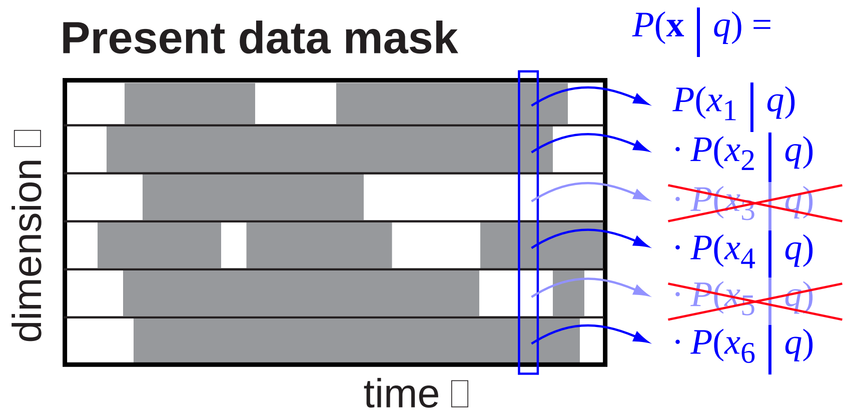
Missing Data Recognition

- **Speech models $p(\mathbf{x}|m)$ are multidimensional...**
  - i.e. means, variances for every freq. channel
  - need values for all dimensions to get $p(\bullet)$
- **But: can evaluate over a subset of dimensions x_k**

$$p(\mathbf{x}_k|m) = \int p(\mathbf{x}_k, \mathbf{x}_u|m) d\mathbf{x}_u$$



- **Hence, missing data recognition:**



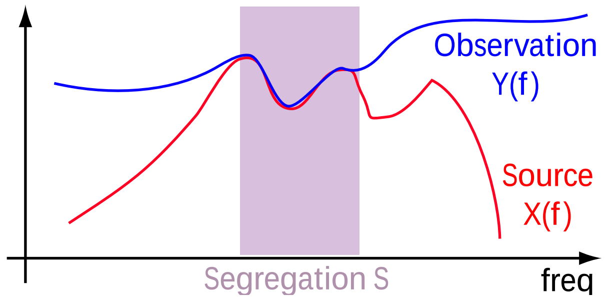
  - hard part is finding the mask (segregation)

Comparing different segregations

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



  - spectral features allow clean relationship

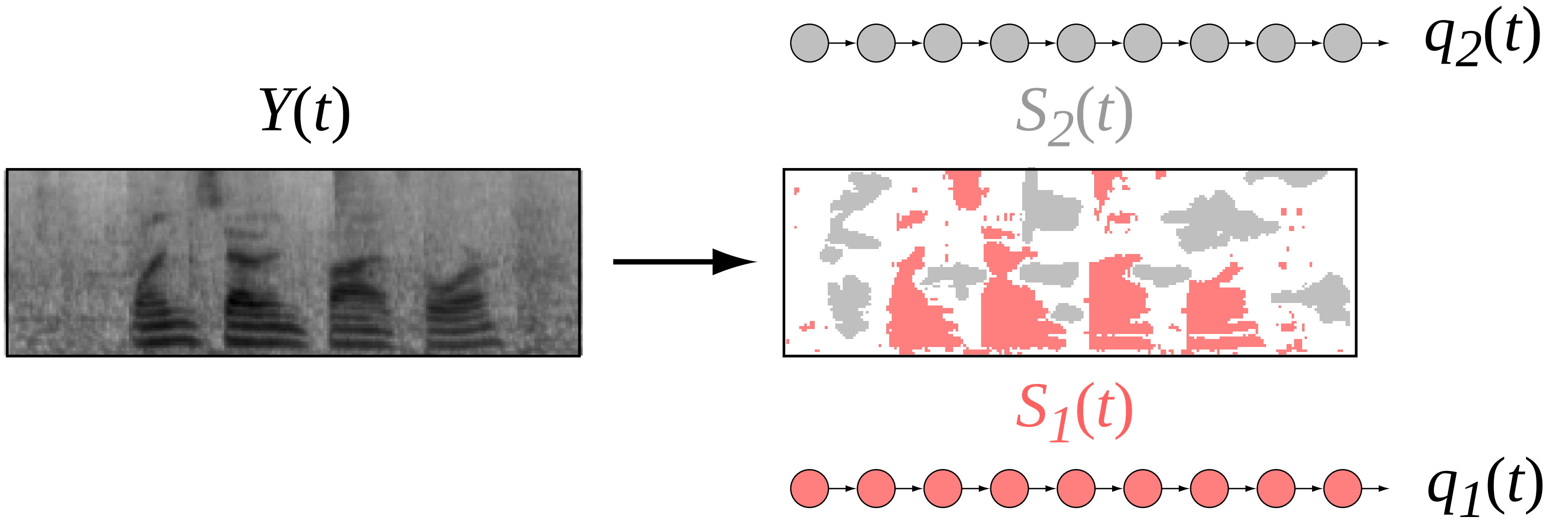
- **Joint classification of model and segregation:**

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

  - probabilistic relation of models & segregation

Multi-source decoding

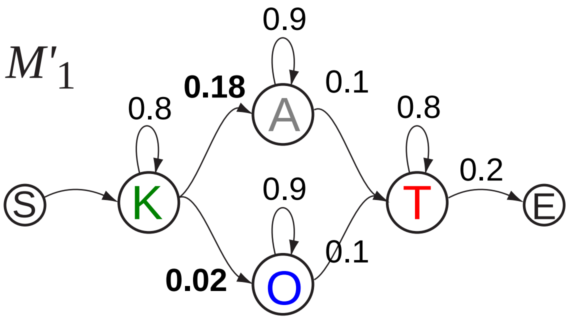
- **Search for more than one source**

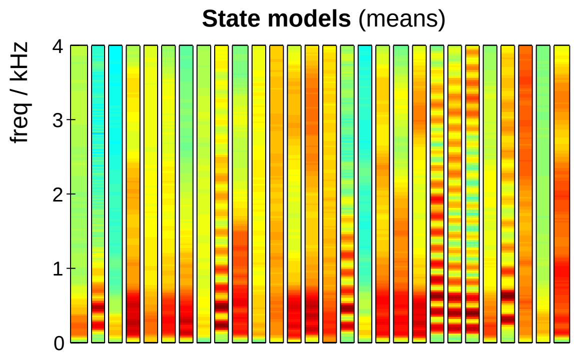


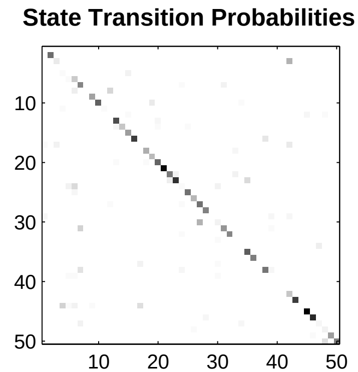
- **Mutually-dependent data masks**
- **Use e.g. CASA features to propose masks**
  - locally coherent regions
- **Lots of issues in models, representations, matching, inference...**

What a speech HMM contains

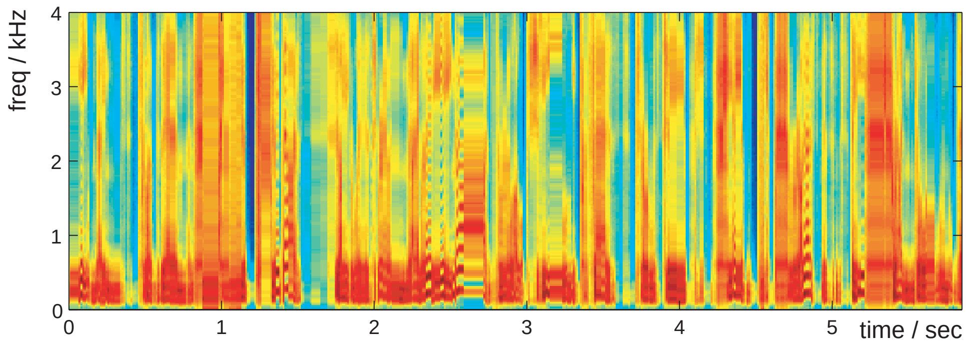
- **Markov model structure: states + transitions**







- **A** generative **model**
  - but not a good speech generator!

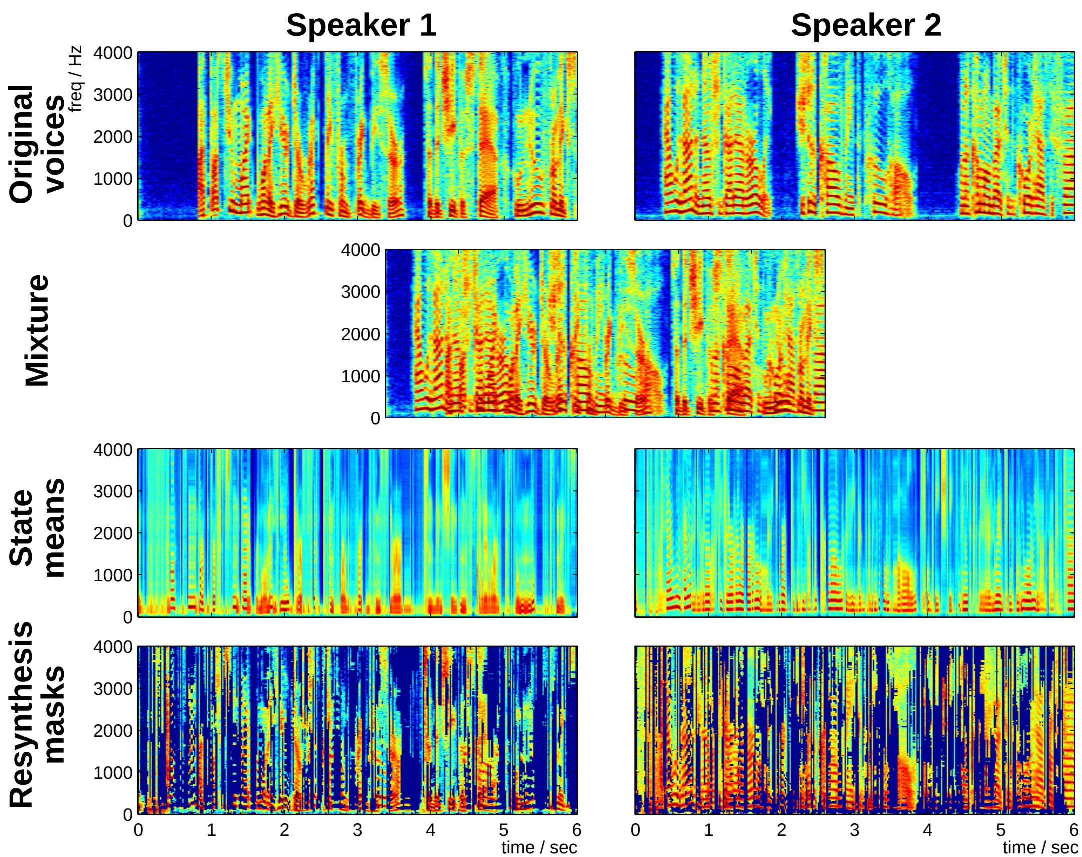


  - only meant for inference of $p(X|M)$

"One microphone source separation"

(Roweis 2000, Manuel Reyes)

- State sequences → t-f estimates → mask

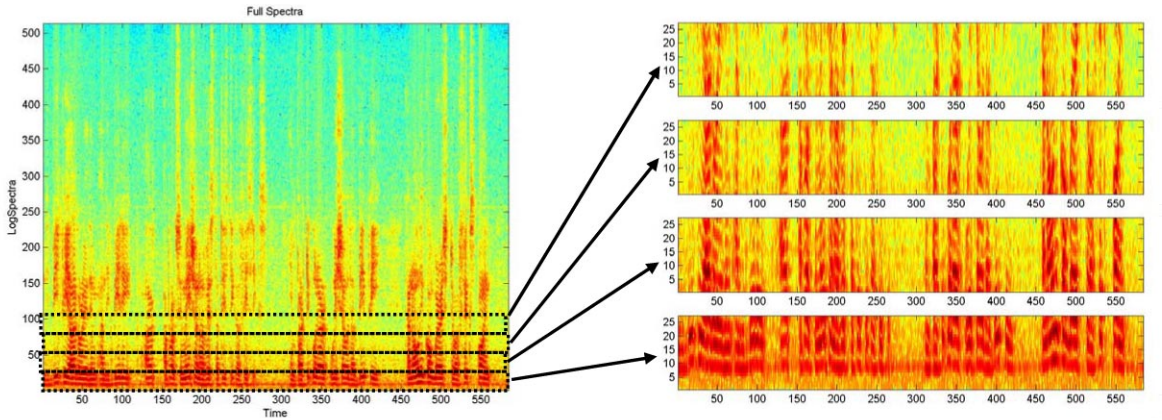


  - 1000 states/model (→ 10^6 transition probs.)
  - simplify by modeling subbands (coupled HMM)?

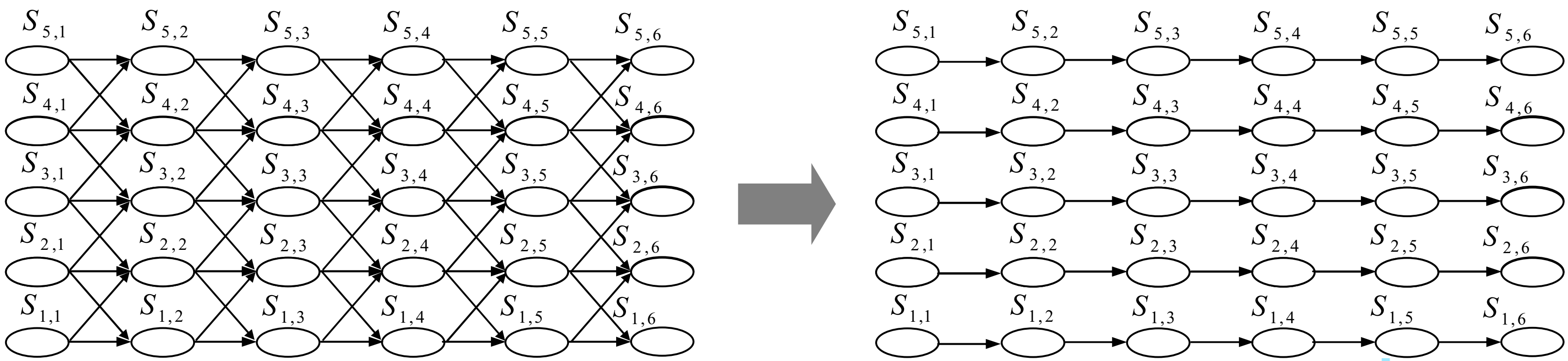
Subband models

(Reyes, Jojic)

- **Reduce the number of states required**
  - 4000 states ×1 band → 30 states×19 bands

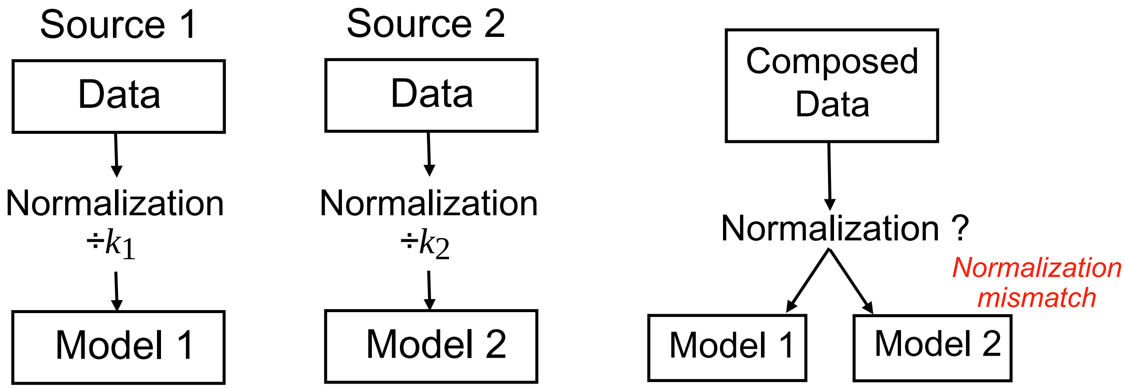


- **Train coupled HMMs via variational approx**

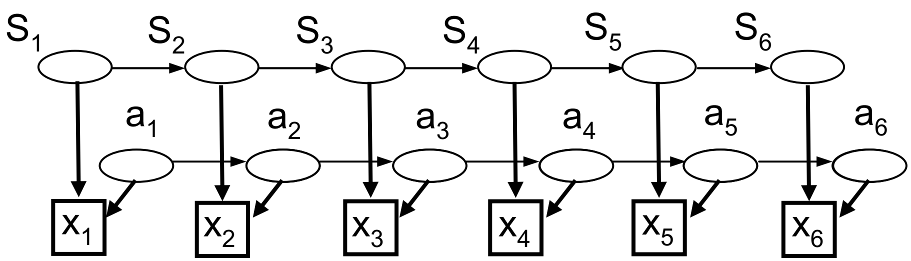


Tracking source normalization

- **Standard HMM sound models try to normalize away energy variation**
  - but: each source in mixture has different 'gain'



- **Instead, factor out scalar gain for each source**



  - solve with var. approx. to $P(\boldsymbol{S},\boldsymbol{a})$

Outline

1. Sound Content Analysis
2. Recognizing sounds
3. Organizing mixtures
4. **Accessing large datasets**
   - Meeting Recordings
   - The Listening Machine
   - Music Information Retrieval

4

Accessing large datasets: The Meeting Recorder Project

(with ICSI, UW, IDIAP, SRI, Sheffield)

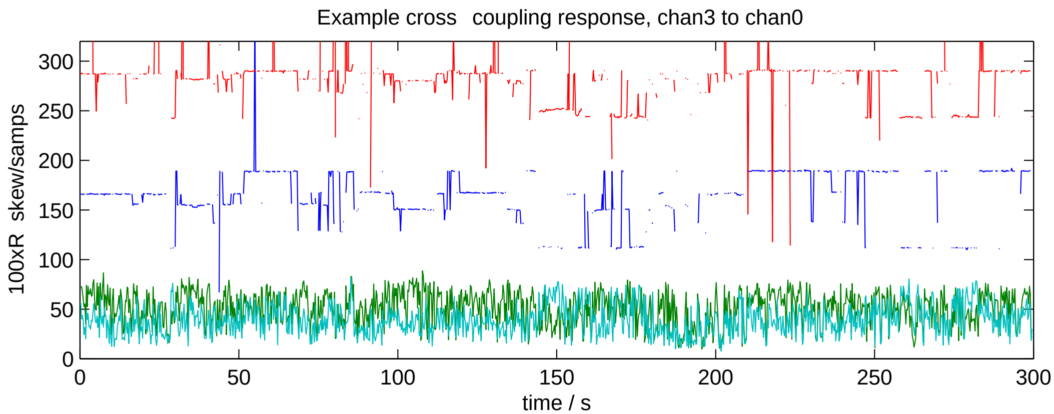
- **Microphones in conventional meetings**
  - for summarization / retrieval / behavior analysis
  - informal, overlapped speech
- **Data collection (ICSI, UW, IDIAP, NIST):**

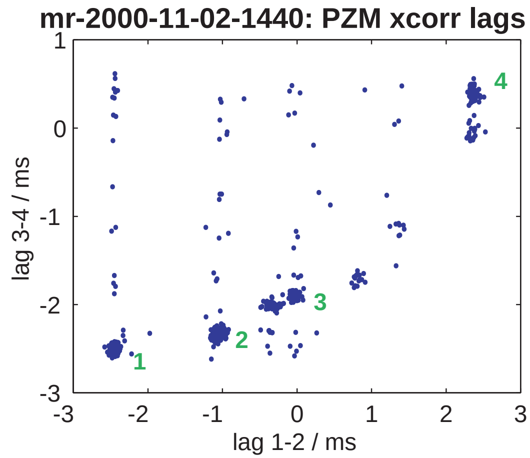
  - ~100 hours collected & transcribed
- **NSF 'Mapping Meetings' project**

Speaker Turn detection

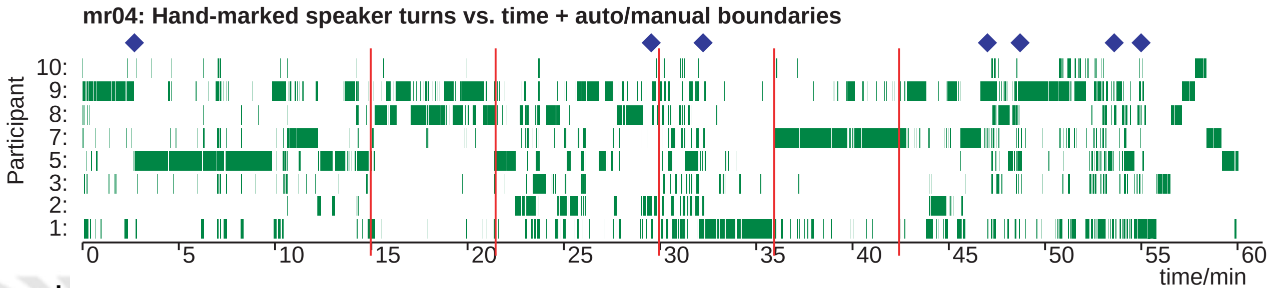
(Huan Wei Hee, Jerry Liu)

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



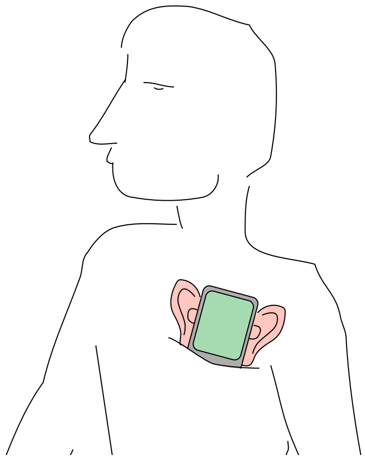


- **Behavioral: Look for patterns of speaker turns**



The Listening Machine

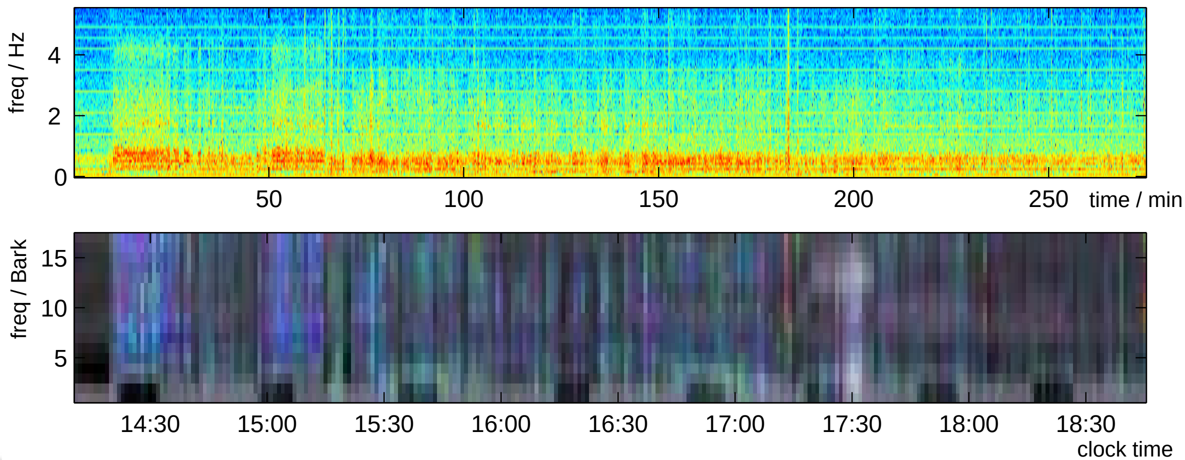
- Smart PDA records everything
- Only useful if we have index, summaries
  - monitor for particular sounds
  - real-time description
- Scenarios



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

Personal Audio

- **LifeLog / MyLifeBits / Remembrance Agent: Easy to record everything you hear**
- **Then what?**
  - prohibitively time consuming to search
  - but .. applications if access easier
- **Automatic content analysis / indexing...**

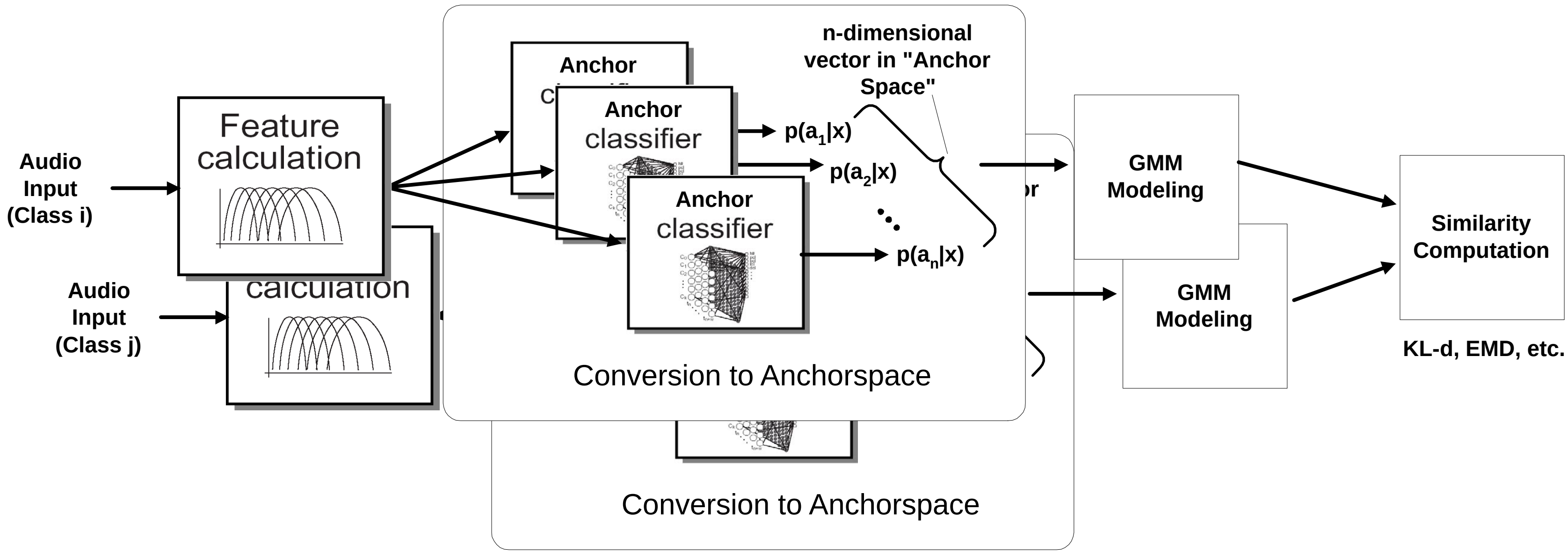


Music Information Retrieval

- **Transfer search concepts to music?**
  - "musical Google"
  - finding something specific / vague / browsing
  - is anything more useful than human annotation?
- **Most interesting area: finding new music**
  - is there anything on mp3.com that I would like?
  - audio is only information source for new bands
- **Basic idea:**
  Project music into a space where neighbors are "similar"
- **Also need models of personal preference**
  - where in the space is the stuff I like
  - relative sensitivity to different dimensions
- **Evaluation problems**
  - requires large, shareable music corpus!

Music similarity from Anchor space

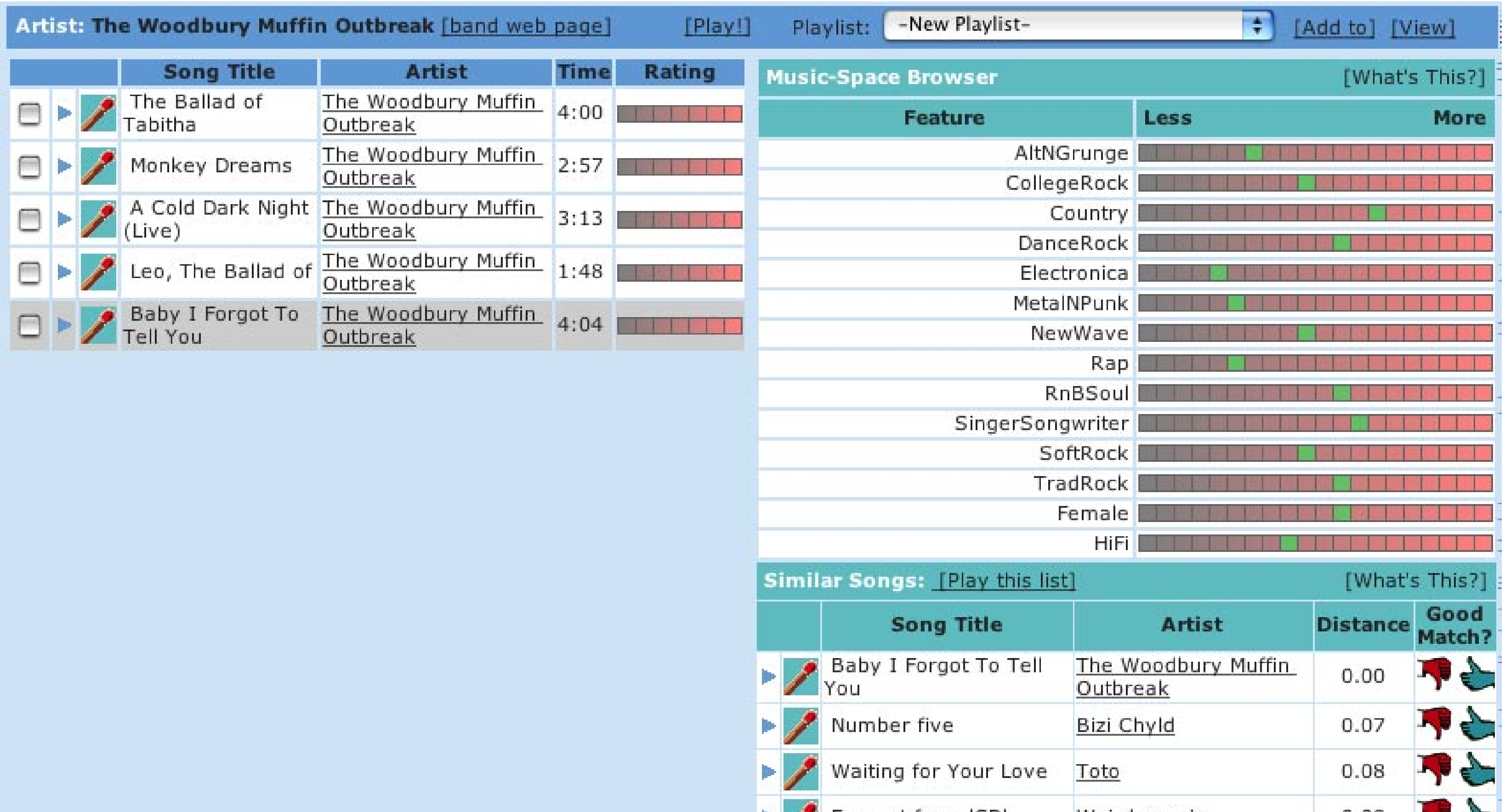
- A classifier trained for one artist (or genre) will respond partially to a similar artist
- Each artist evokes a particular pattern of responses over a set of classifiers
- We can treat these classifier outputs as a new feature space in which to estimate similarity



- “Anchor space” reflects subjective qualities?

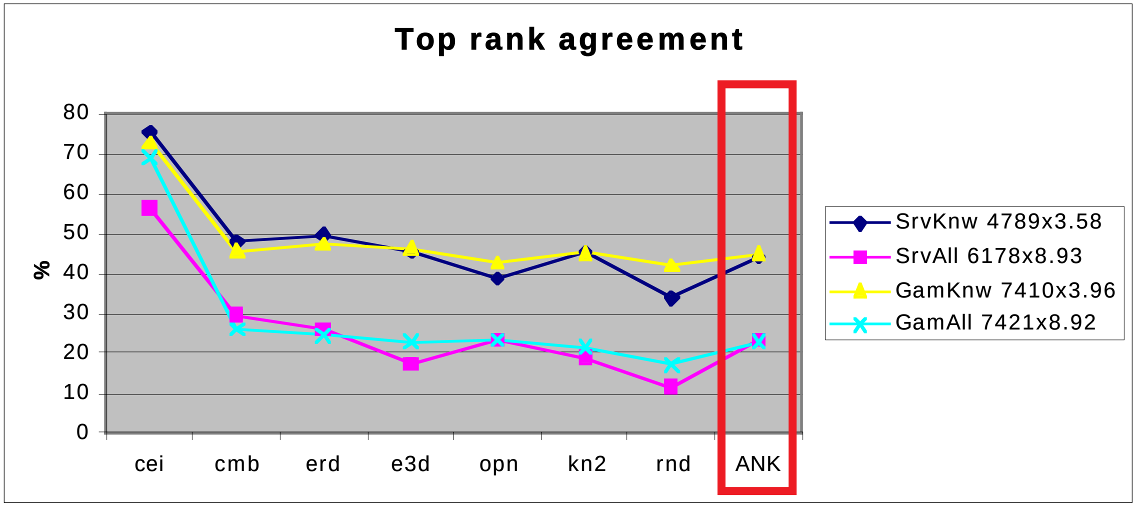
Playola interface (www.playola.org)

- Browser finds closest matches to single tracks or entire artists in anchor space
- Direct manipulation of anchor space axes



Evaluation

- **Are recommendations good or bad?**
- Subjective **evaluation is the ground truth**
  - .. but subjects aren't familiar with the bands being recommended
  - can take a long time to decide if a recommendation is good
- **Measure match to other similarity judgments**
  - e.g. musicseer data:



Summary

- **Sound**
  - .. contains much, valuable information at many levels
  - intelligent systems need to use this information
- **Mixtures**
  - .. are an unavoidable complication when using sound
  - looking in the right time-frequency place to find points of dominance
- **Learning**
  - need to acquire constraints from the environment
  - recognition/classification as the real task

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

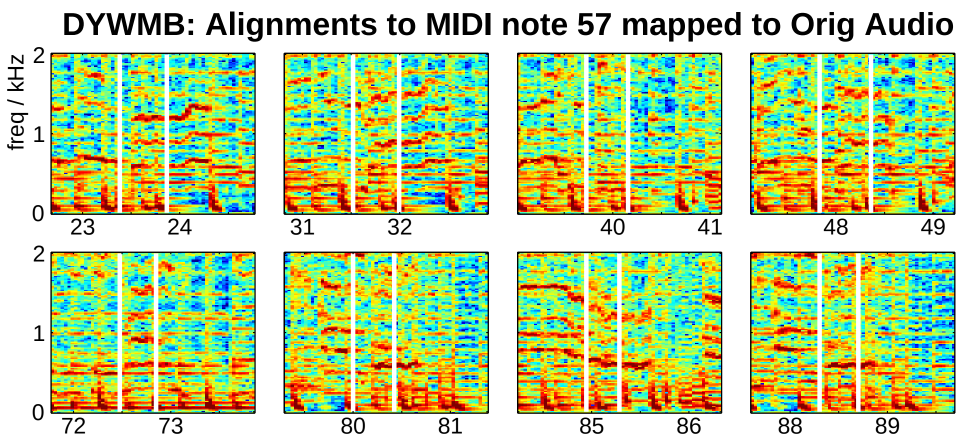
APPLICATIONS

- Structuring
- Search
- Summarization
- Awareness
- Understanding

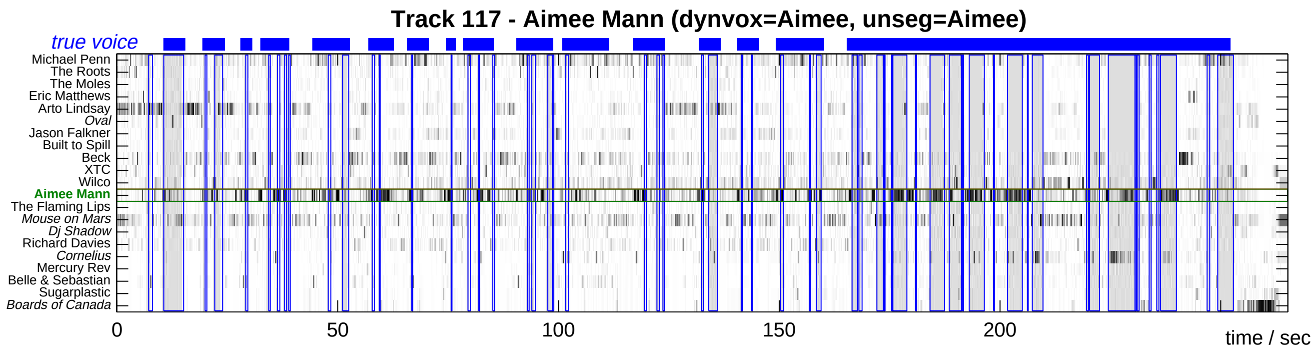
Extra Slides

Music Applications

- **Music as a complex, information-rich sound**
- **Applications of separation & recognition:**
  - note/chord detection & classification



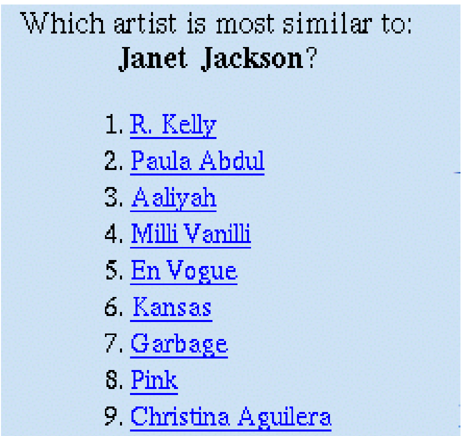
  - singing detection (→ genre identification ...)



Artist Similarity

- **Recognizing work from each artist is all very well...**
- **But: what is similarity between artists?**

- pattern recognition systems give a number...


Which artist is most similar to:
Janet Jackson?

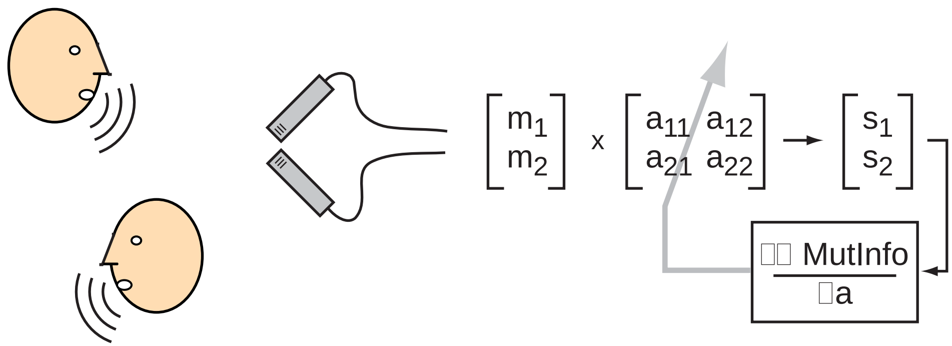
1. R. Kelly
2. Paula Abdul
3. Aaliyah
4. Milli Vanilli
5. En Vogue
6. Kansas
7. Garbage
8. Pink
9. Christina Aguilera

- **Need subjective ground truth: Collected via web site**
  www.musicseer.com
- **Results:**
  - 1,000 users, 22,300 judgments collected over 6 months

Independent Component Analysis (ICA)

(Bell & Sejnowski 1995 et seq.)

- **Drive a parameterized separation algorithm to maximize independence of outputs**



- **Advantages:**
  - mathematically rigorous, minimal assumptions
  - does not rely on prior information from models
- **Disadvantages:**
  - may converge to local optima...
  - separation, not recognition
  - does not exploit prior information from models