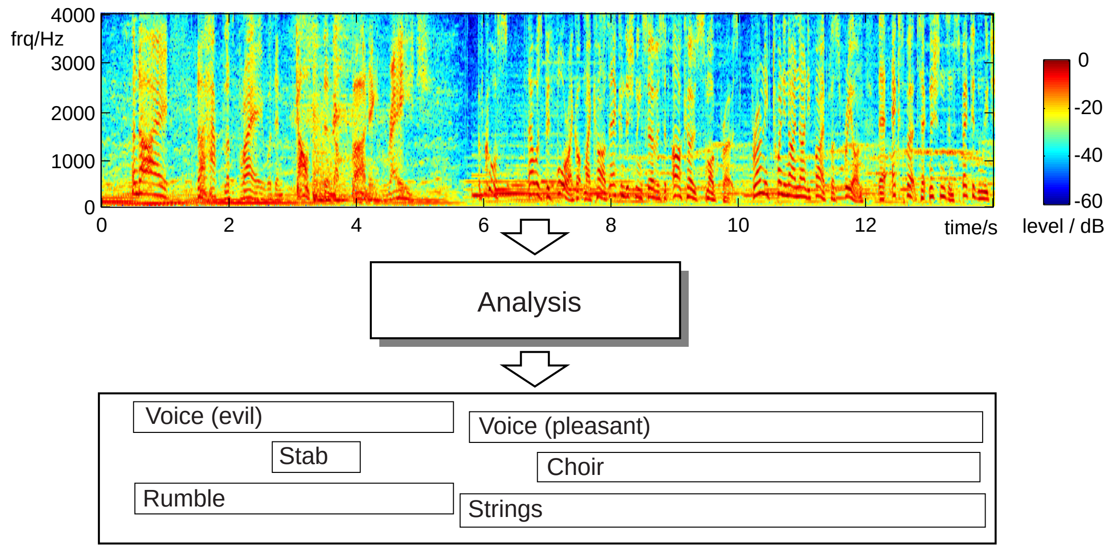
# Sound, Mixtures, and Learning: LabROSA overview

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

Dan Ellis <dpwe@ee.columbia.edu>

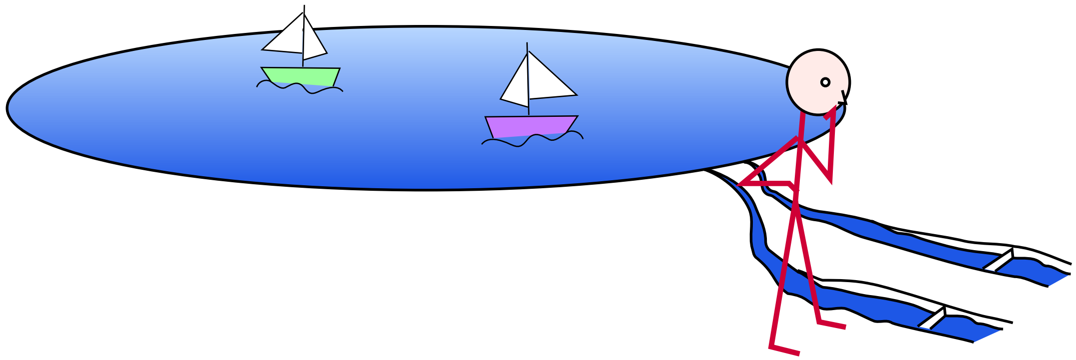
Laboratory for Recognition and Organization of Speech and Audio
(Lab**ROSA**)
Columbia University, New York
http://labrosa.ee.columbia.edu/

# 1 Sound Content Analysis




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

# The problem with recognizing mixtures



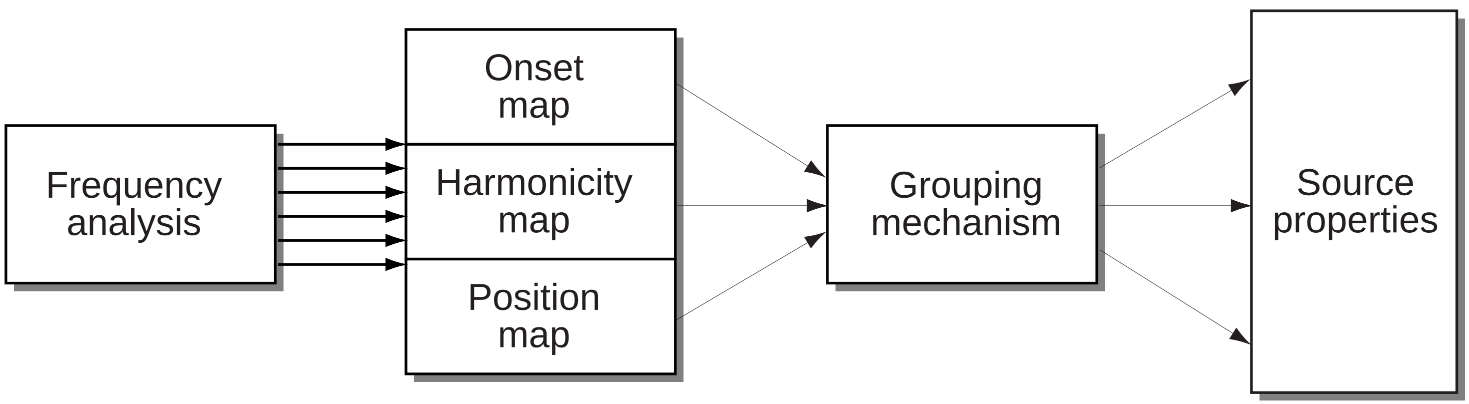
*“Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?”* (after Bregman’90)

- **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

# Auditory Scene Analysis

(Bregman 1990)

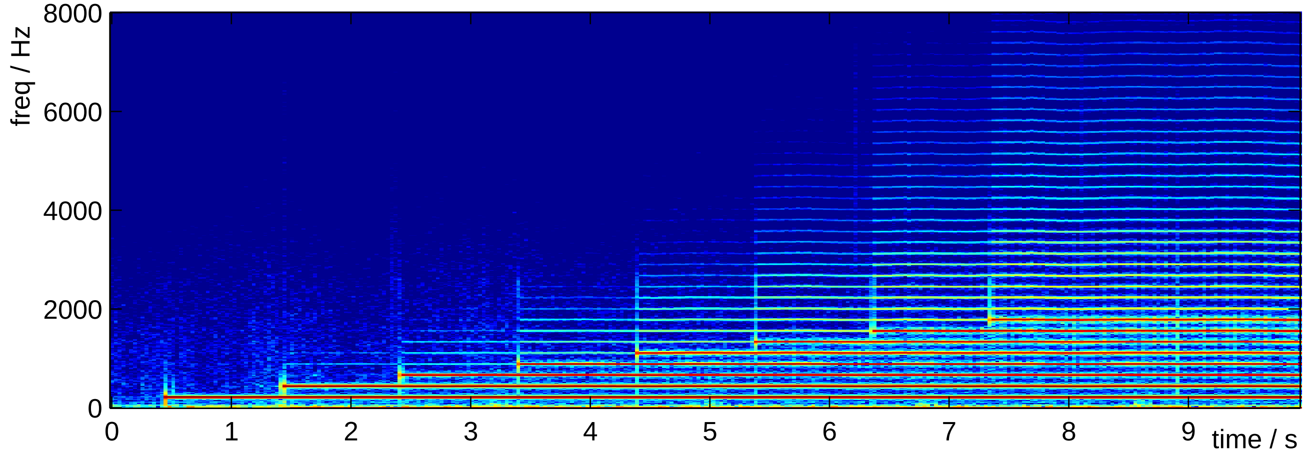
- **How do people analyze sound mixtures?**
  - break mixture into small *elements* (in time-freq)
  - elements are *grouped* in to sources using *cues*
  - sources have aggregate *attributes*
- **Grouping 'rules' (Darwin, Carlyon, ...):**
  - cues: common onset/offset/modulation, harmonicity, spatial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




*(after Darwin, 1996)*

# Cues to simultaneous grouping

- **Elements + attributes**



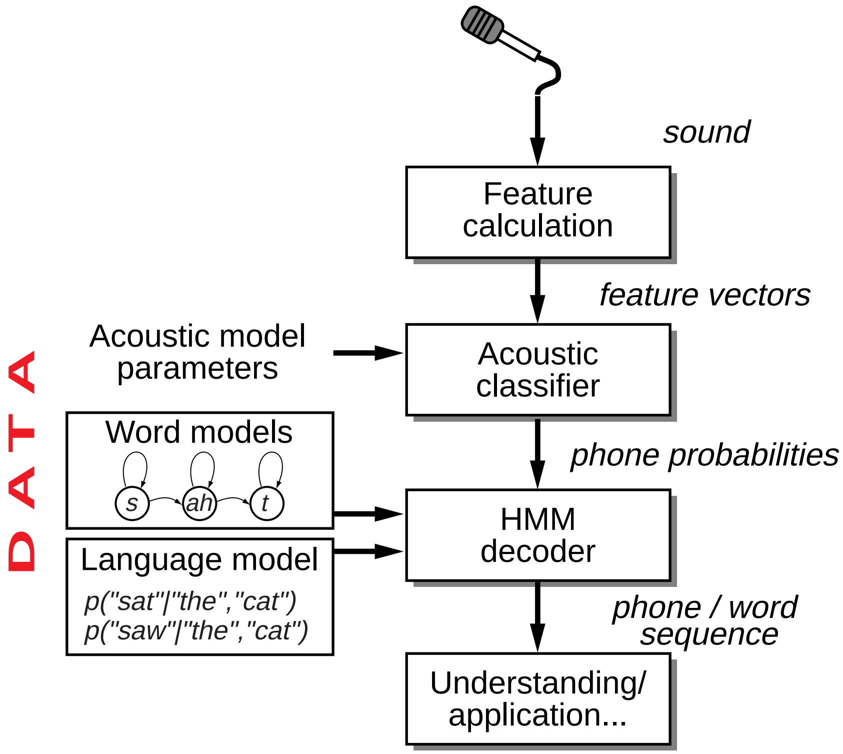

- **Common onset**
  - simultaneous energy has common source
- **Periodicity**
  - energy in different bands with same cycle
- **Other cues**
  - spatial (ITD/IID), familiarity, ...
- **But: Context ...**

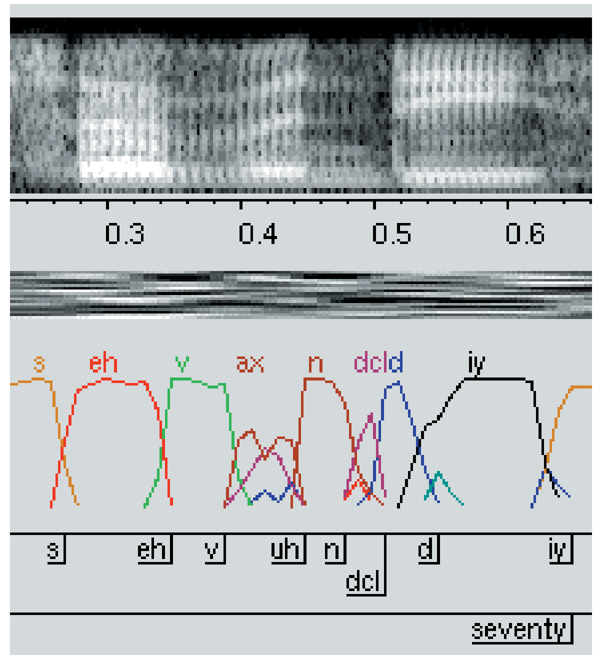
# Outline

1. **Sound Content Analysis**
2. **Recognizing sounds**
   - Clean speech
   - Speech-in-noise
   - Nonspeech
3. **Organizing mixtures**
4. **Accessing large datasets**
5. **Music Information Retrieval**

# 2 Recognizing Sounds: Speech

- **Standard speech recognition structure:**

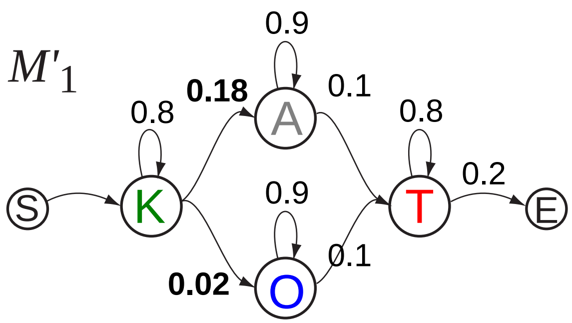


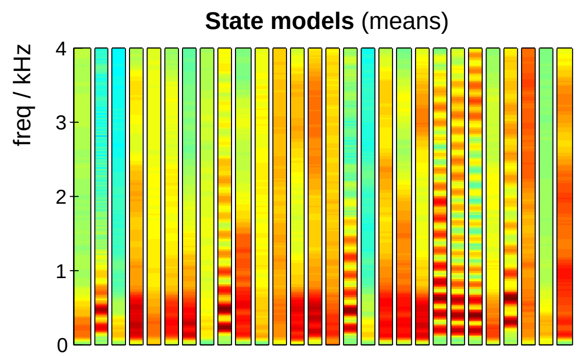



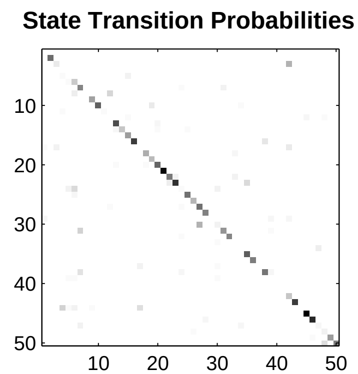

- **How to handle additive noise?**
  - just train on noisy data: 'multicondition training'

# How ASR Represents Speech

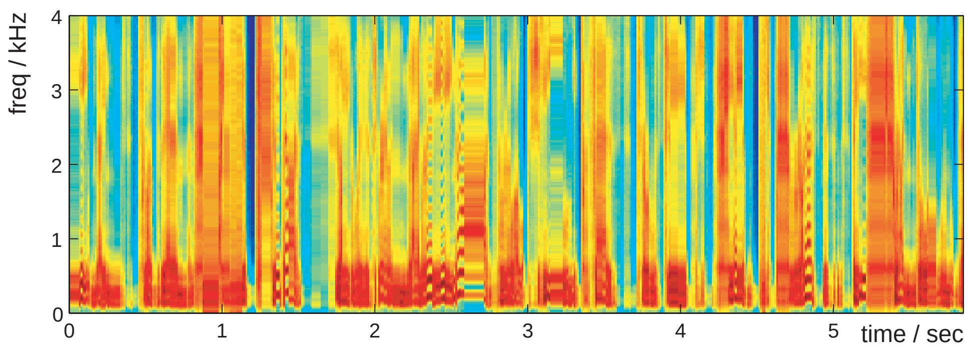
- **Markov model structure: states + transitions**




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

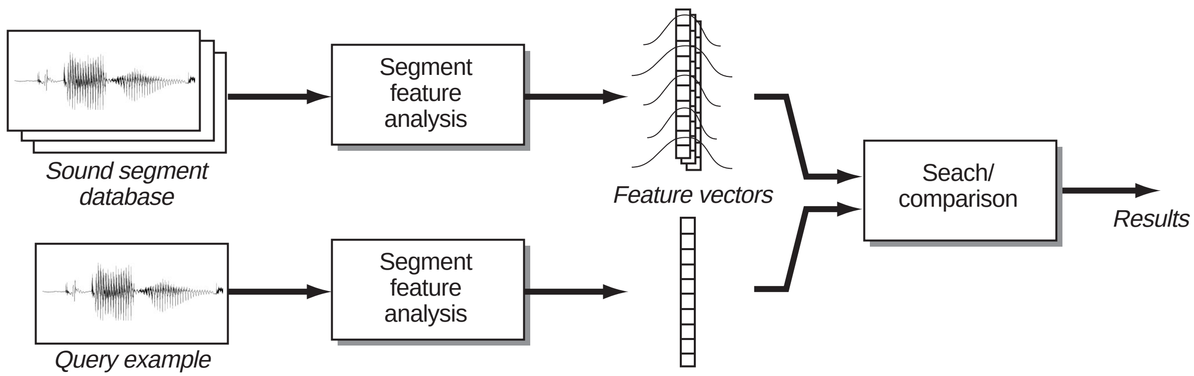



  - only meant for inference of $p(X|M)$

# General Audio Recognition

(with Manuel Reyes)

- **Searching audio databases**
  - 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

# Audio Recognition: Results

- **Musclefish corpus**
  - most commonly reported set
- **Features**
  - MFCC, brightness, bandwidth, pitch ...
  - no temporal structure
- **Results:**
  - 208 examples, 16 classes

**Global features**: 41% corr

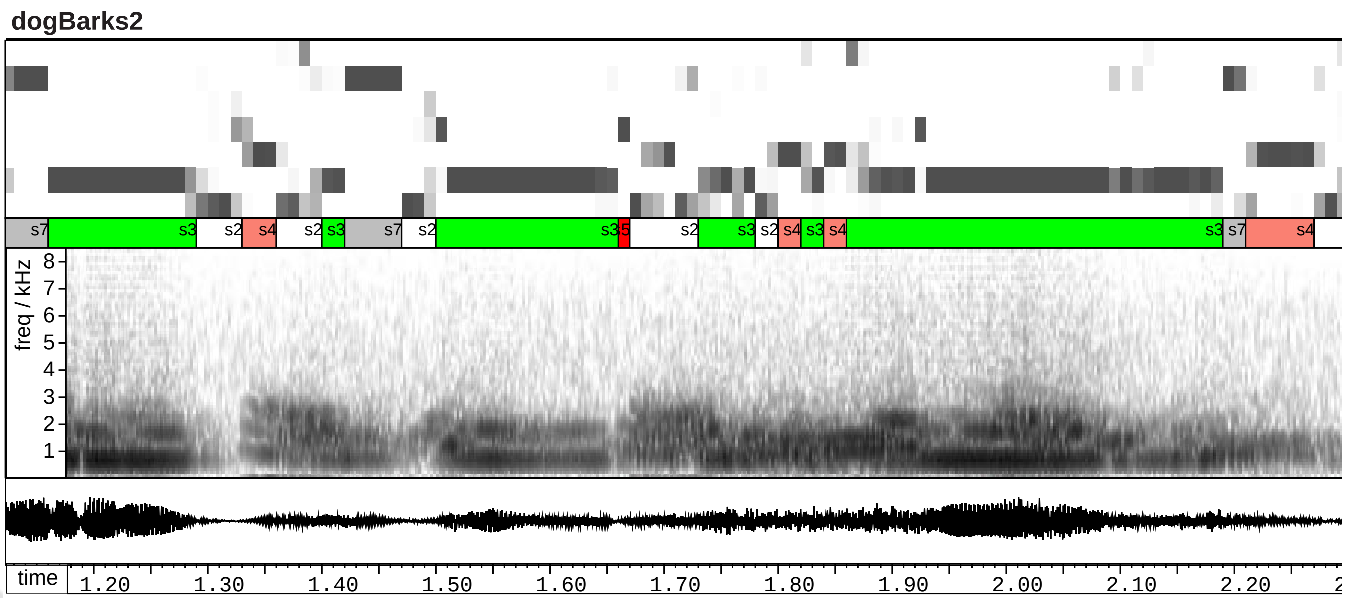
| | *Mu* | *Sp* | *Env* | *An* | *Mec* |
|---|---|---|---|---|---|
| *Musical* | 59/ **46** | | **24** | **2** | **19** |
| *Speech* | | 11/ **6** | **4** | **5** | |
| *Eviron.* | | | 7/ **2** | | |
| *Animals* | | | **2** | 1/ | **2** |
| *Mechan* | **1** | | **4** | **1** | 8/ **4** |

**HMM models**: 81% corr.

| | *Mu* | *Sp* | *Env* | *An* | *Mec* |
|---|---|---|---|---|---|
| *Musical* | 136/ **6** | | **2** | **1** | **5** |
| *Speech* | **1** | 14/ **2** | **5** | **3** | **1** |
| *Eviron.* | **1** | | 7/ | **1** | |
| *Animals* | | | | 4/ | **1** |
| *Mechan* | **3** | | **3** | | 12/ |

# What are the HMM states?

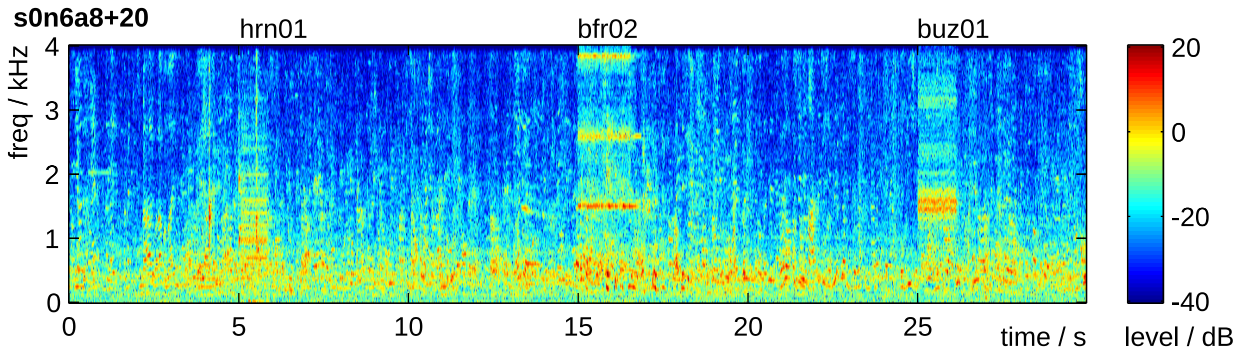
- **No sub-units defined for nonspeech sounds**
- **Final states depend structure, initialization**
  - number of states
  - initial clusters / labels / transition matrix
  - EM update objective
- **Have ideas of what we'd like to get**
  - investigate features/initialization to get there



# Alarm sound detection

(Ellis 2001)

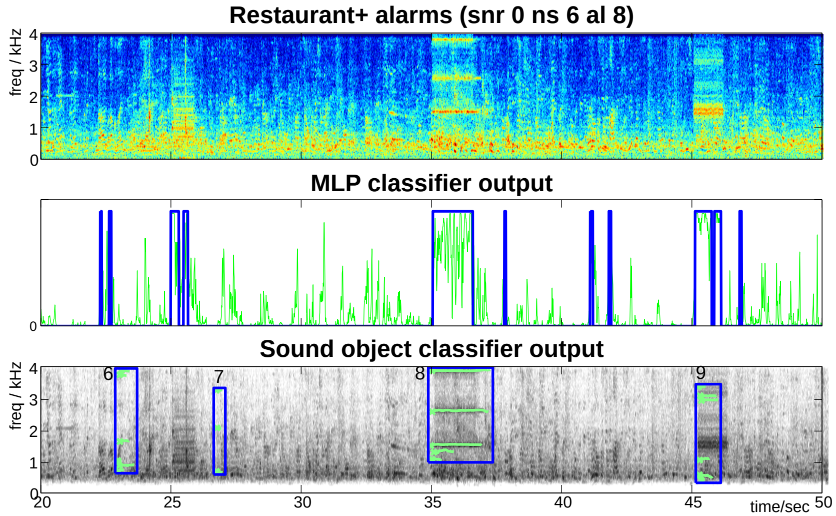
- **Alarm sounds have particular structure**
  - people ‘know them when they hear them’
  - clear even at low SNRs




- **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
   - Parallel model inference
4. Accessing large datasets
5. Music Information Retrieval

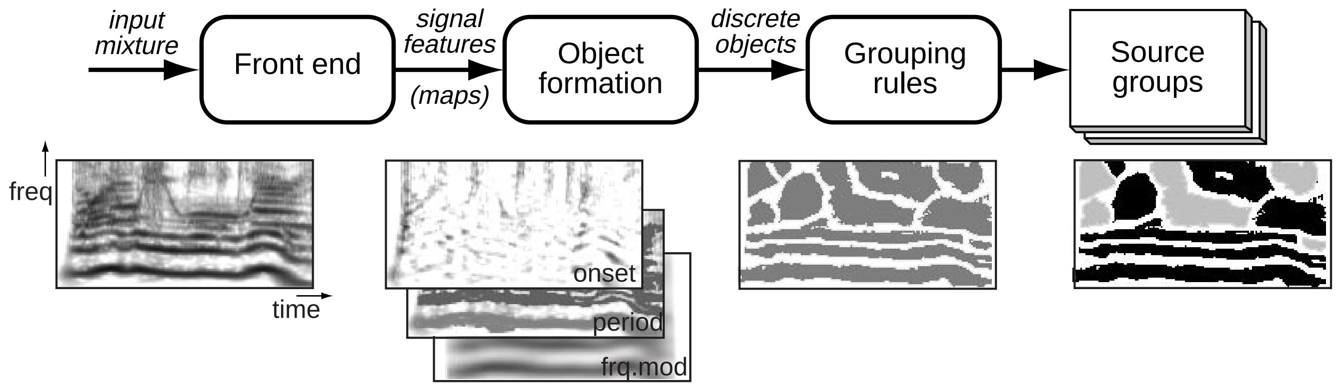
## 3 Organizing mixtures: Approaches to handling overlapped sound

- **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

# Computational Auditory Scene Analysis: The Representational Approach

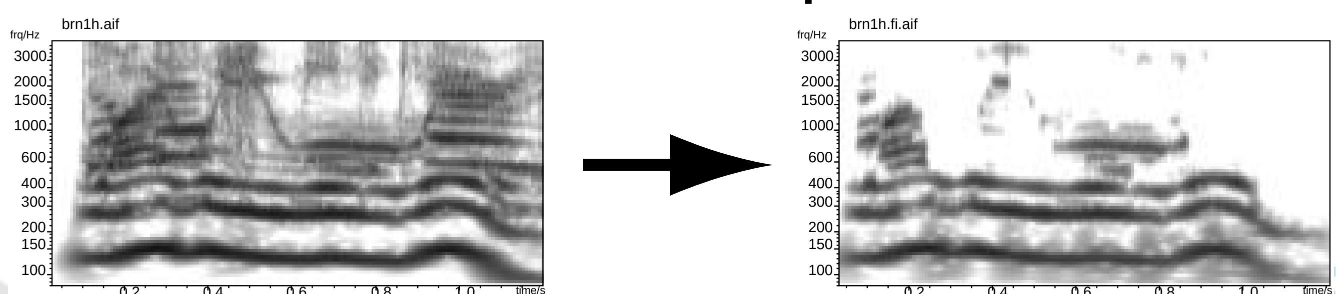
(Cooke & Brown 1993)

- **Direct implementation of psych. theory**




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

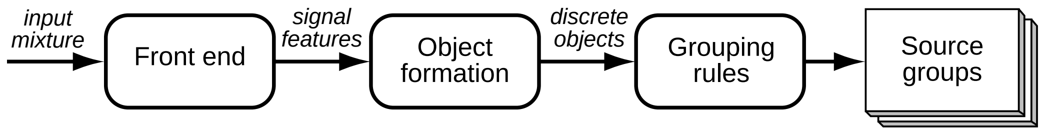
- **Able to extract voiced speech:**



# Adding top-down constraints

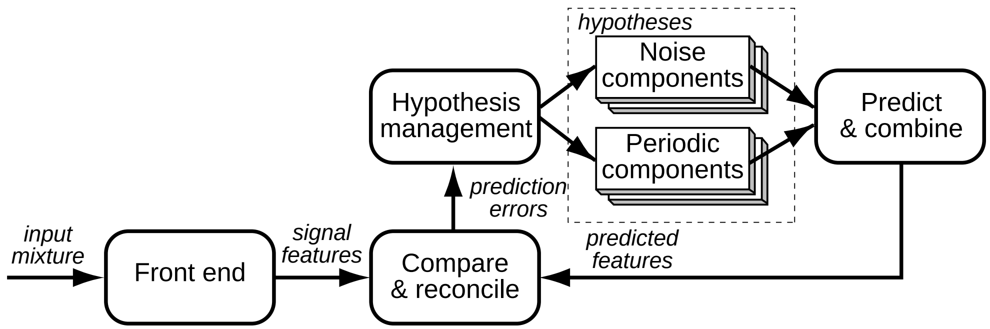
**Perception is not direct**
**but a search for plausible hypotheses**

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



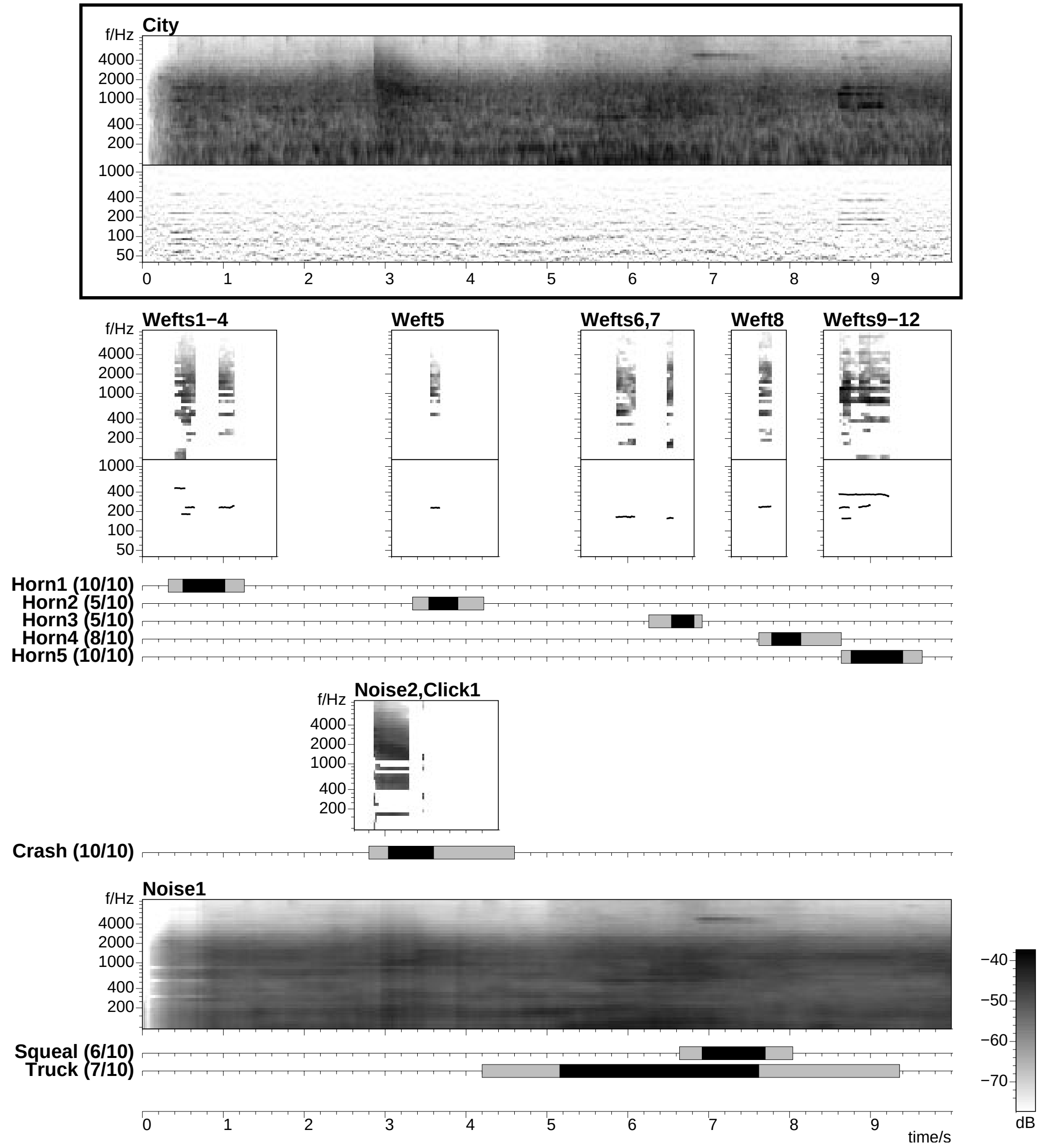

  - objects irresistibly appear

**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?

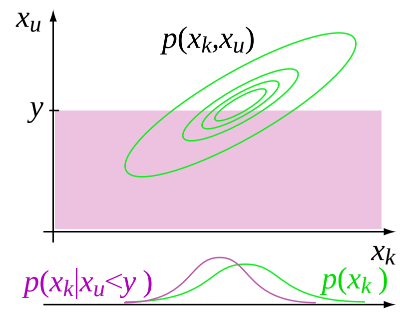
$$p(O, x, y) \rightarrow p(x, y | O)$$

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

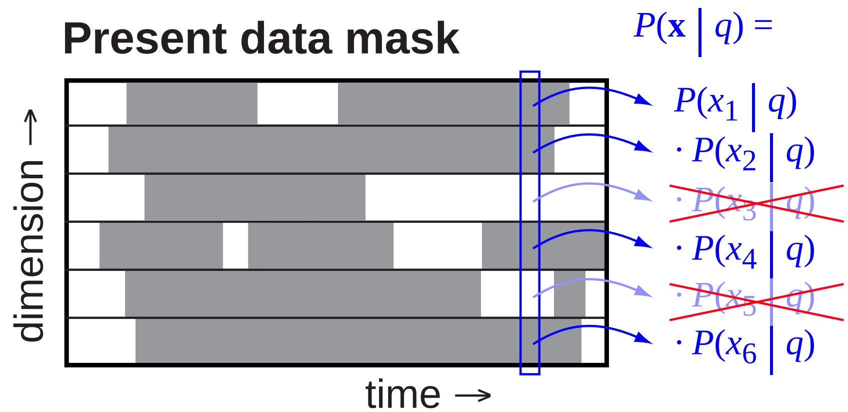
# Missing Data Recognition

- **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:**



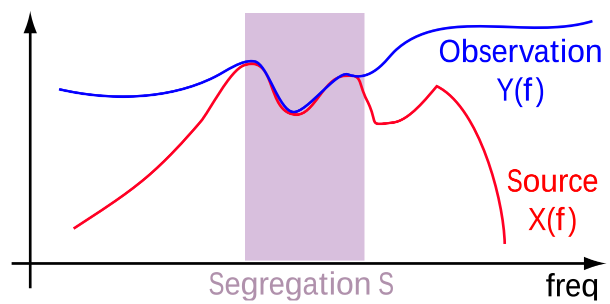

  - hard part is finding the mask (segregation)

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




  - spectral features allow clean relationship

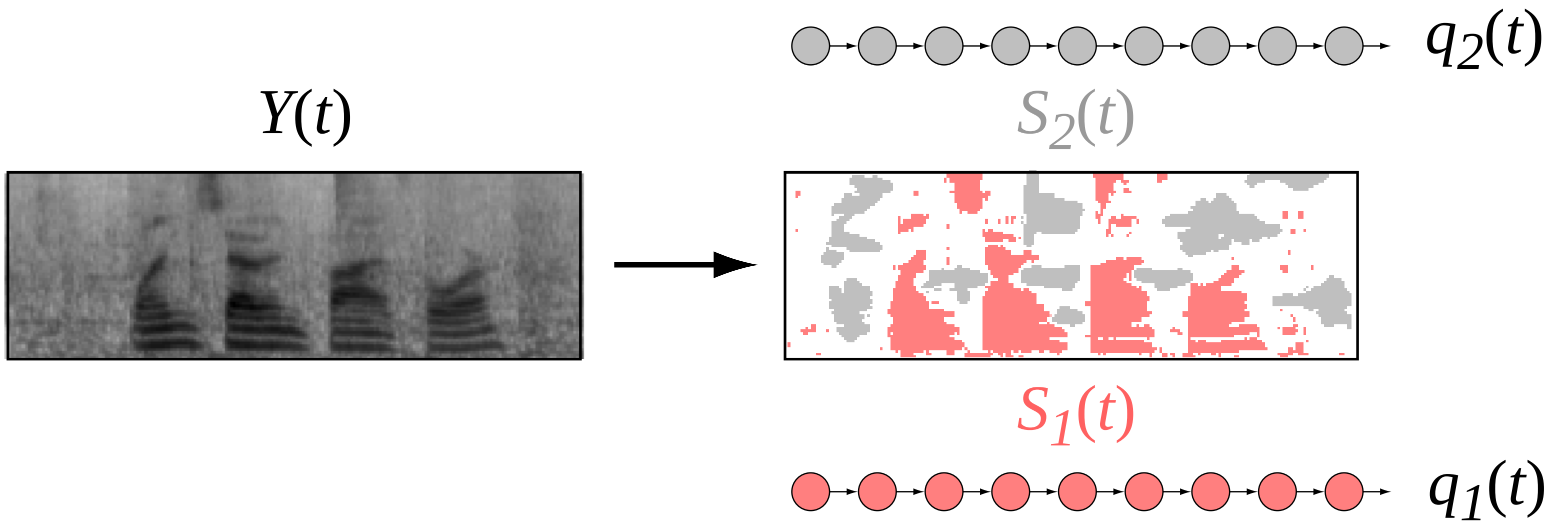
- **Joint classification of model and segregation:**

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

  - 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...**

# Outline

1. Sound Content Analysis
2. Recognizing sounds
3. Organizing mixtures
4. **Accessing large datasets**
    - Spoken documents
    - The Listening Machine
    - Music preference modeling
5. Music Information Retrieval

4

# Accessing large datasets: The Meeting Recorder Project

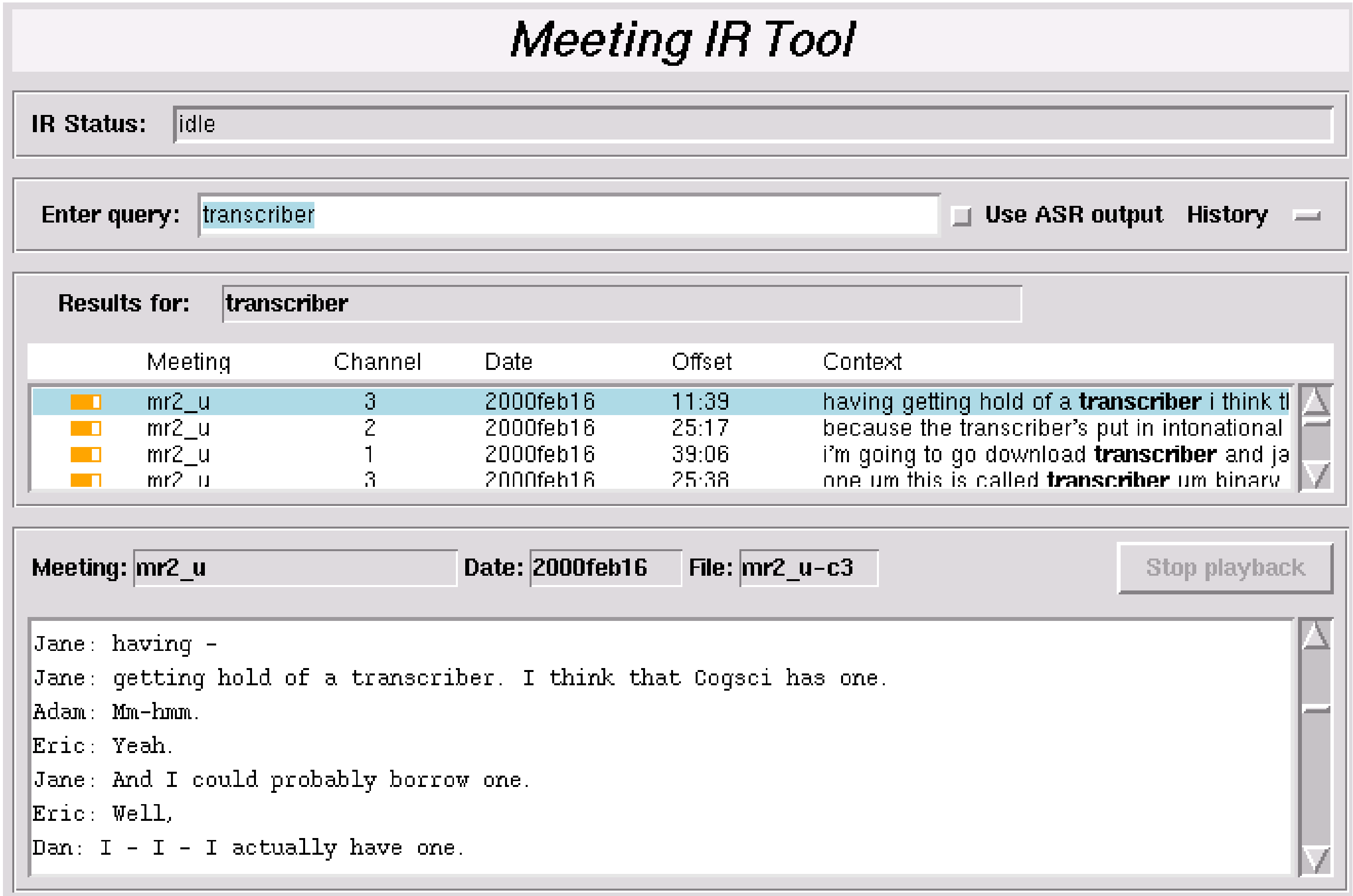
(with ICSI, UW, IDIAP, SRI, Sheffield)

- **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**

# Meeting IR tool

- **IR on (ASR) transcripts from meetings**

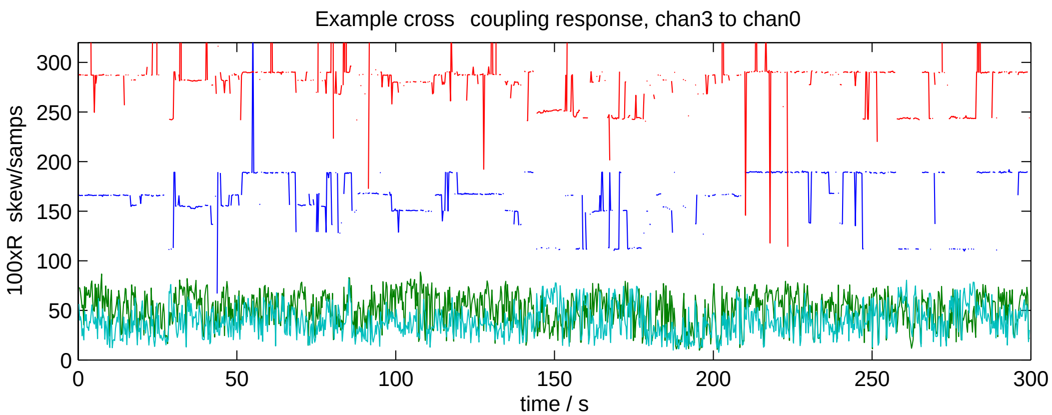


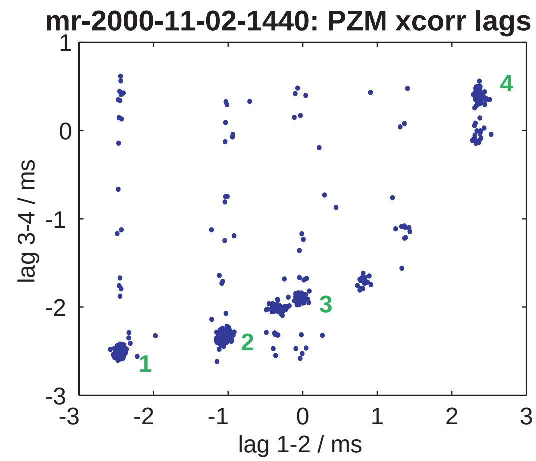

- ASR errors have limited impact on retrieval

# Speaker Turn detection

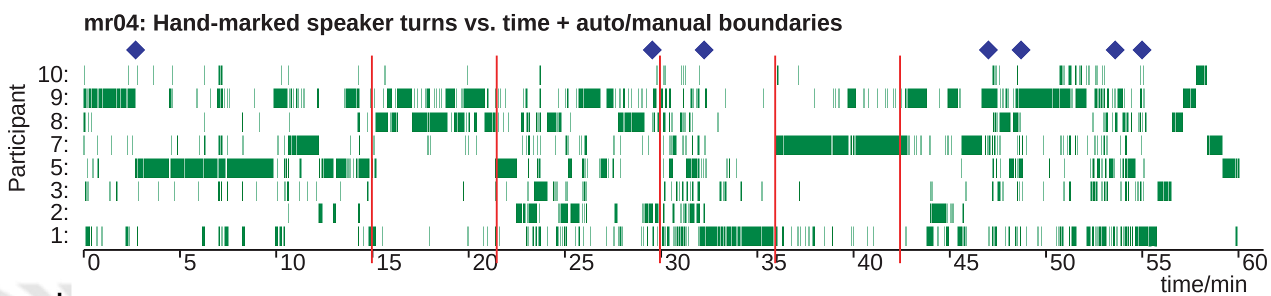
(Huan Wei Hee, Jerry Liu)

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



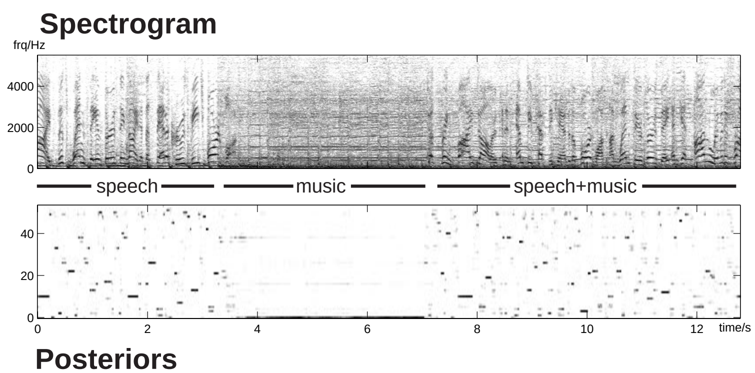

- **Behavioral**: **Look for patterns of speaker turns**

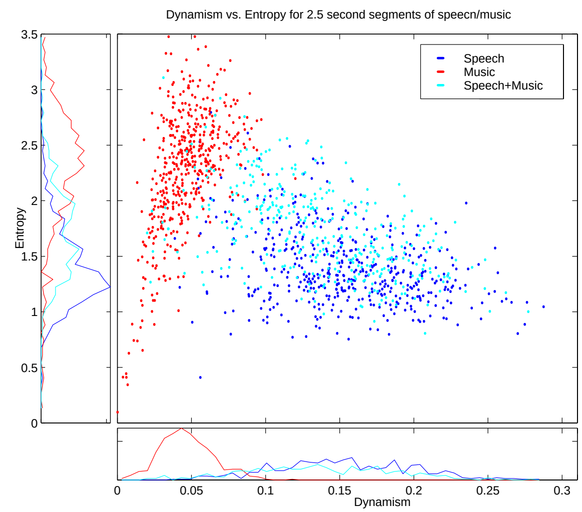


# Speech/nonspeech detection

(Williams & Ellis 1999)

- **ASR run over entire soundtracks?**
  - for nonspeech, result is nonsense
- **Watch behavior of speech acoustic model:**
  - average per-frame entropy
  - ‘dynamism’ - mean-squared 1st-order difference

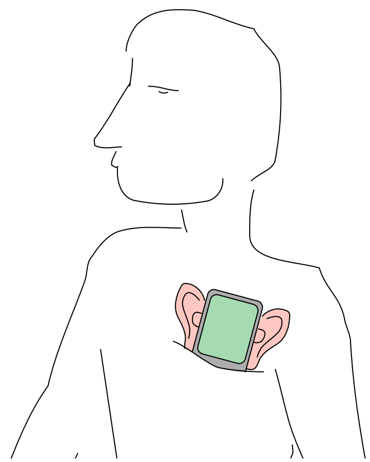



- **1.3% error on 2.5 second speech-music testset**

# The Listening Machine

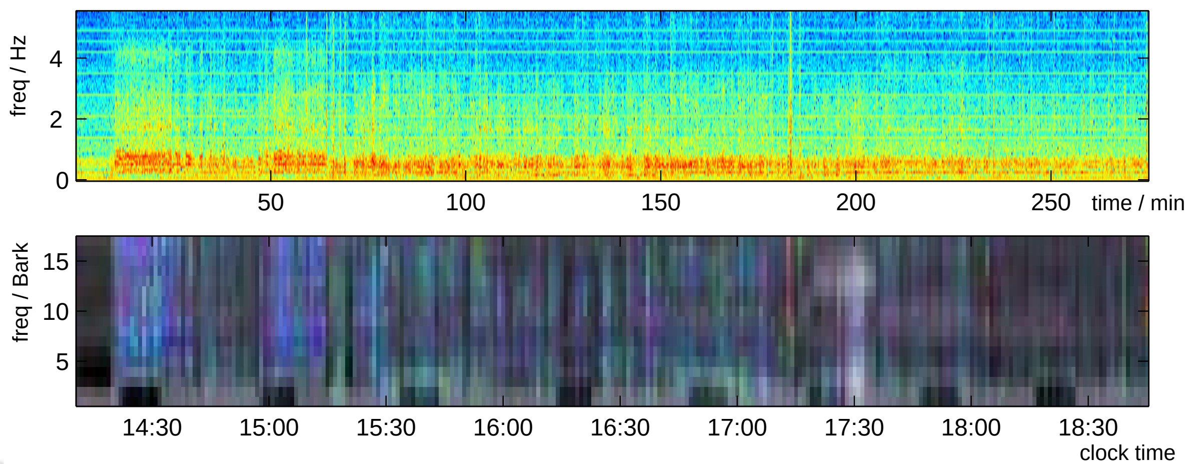
- **Smart PDA records everything**
- **Only useful if we have index, summaries**
  - monitor for particular sounds
  - real-time description
- **Scenarios**



  - 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...**



# Outline

1. Sound Content Analysis
2. Recognizing sounds
3. Organizing mixtures
4. Accessing large datasets
5. **Music Information Retrieval**
   - Anchor space
   - Playola browser

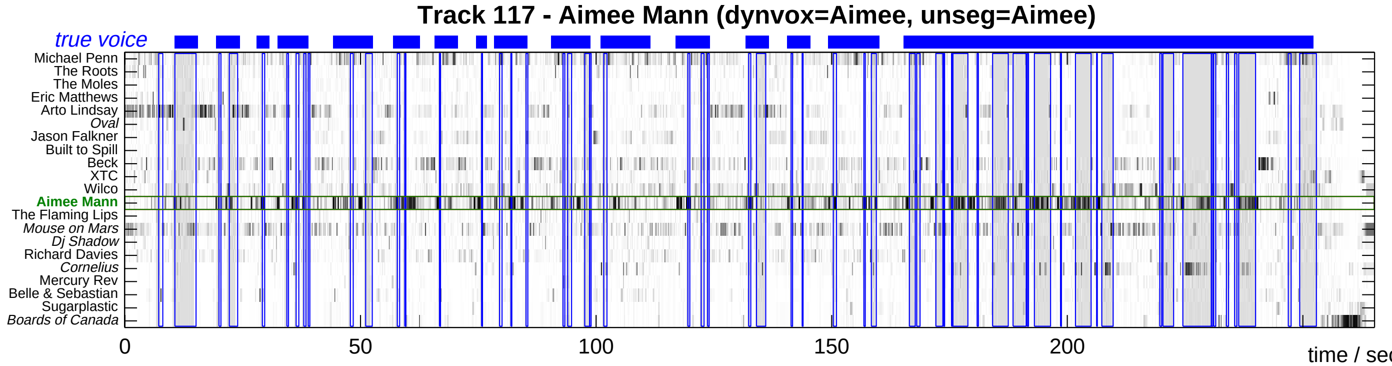
# 5 Music Information Retrieval

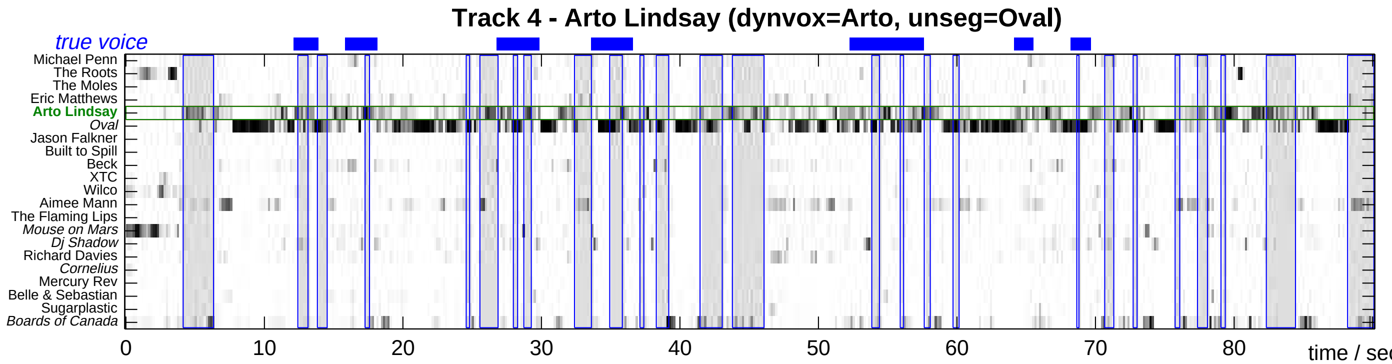
- **Transfer search concepts to music?**
  - "musical Google"
  - finding something specific / vague / browsing
  - is anything more useful than human annotation?
- **Most interesting area: finding new music**
  - is there anything on mp3.com that I would like?
  - audio is only information source for 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!

# Artist Classification

(Berenzweig et al. 2001)

- **Artists' oeuvres as similarity-sets**
- **Train MLP to classify frames among 21 artists**
- **Using (detected) voice segments:**
  **Song-level accuracy improves 56.7% → 64.9%**

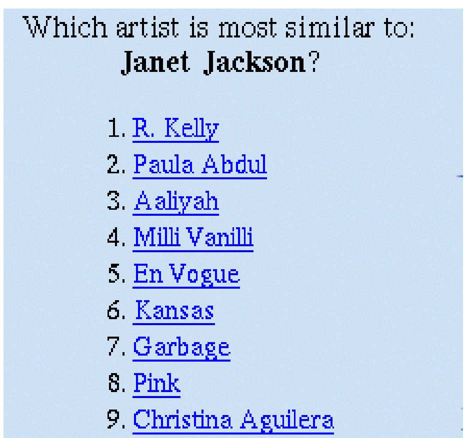


# Artist Similarity

- **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**?

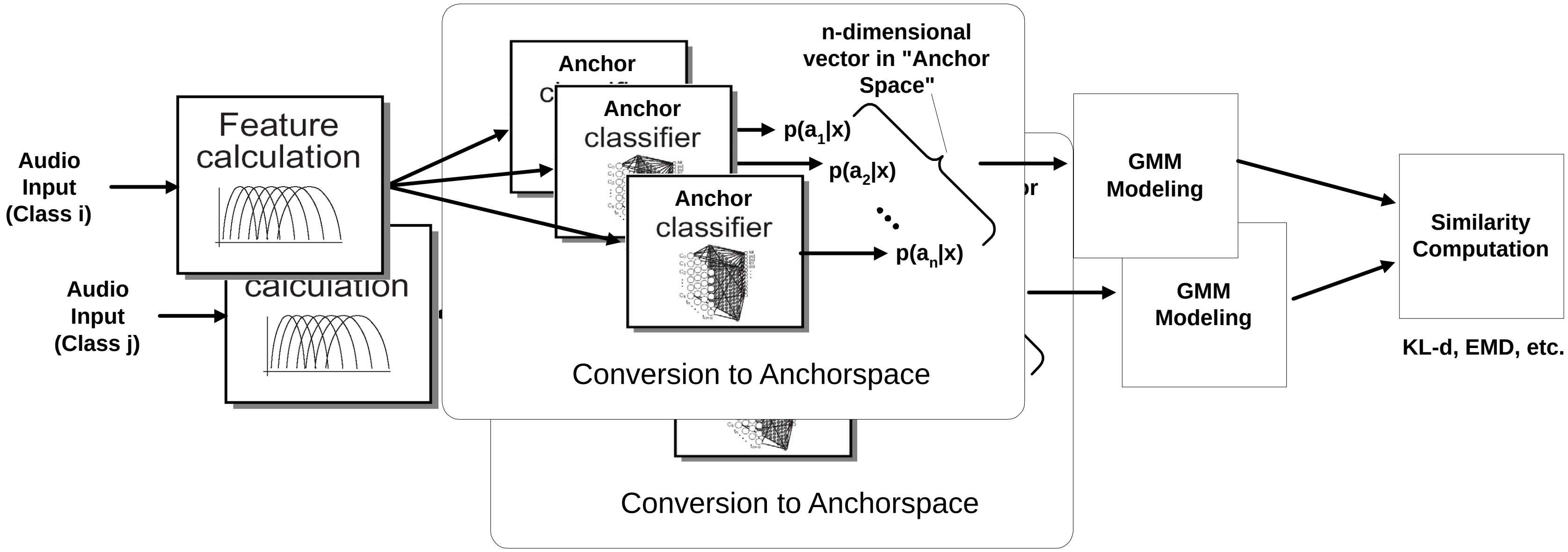
1. R. Kelly
2. Paula Abdul
3. Aaliyah
4. Milli Vanilli
5. En Vogue
6. Kansas
7. Garbage
8. Pink
9. Christina Aguilera

- **Need subjective ground truth: Collected via web site**
  www.musicseer.com

- **Results:**
  - 1,000 users, 22,300 judgments collected over 6 months

# Music similarity from Anchor space

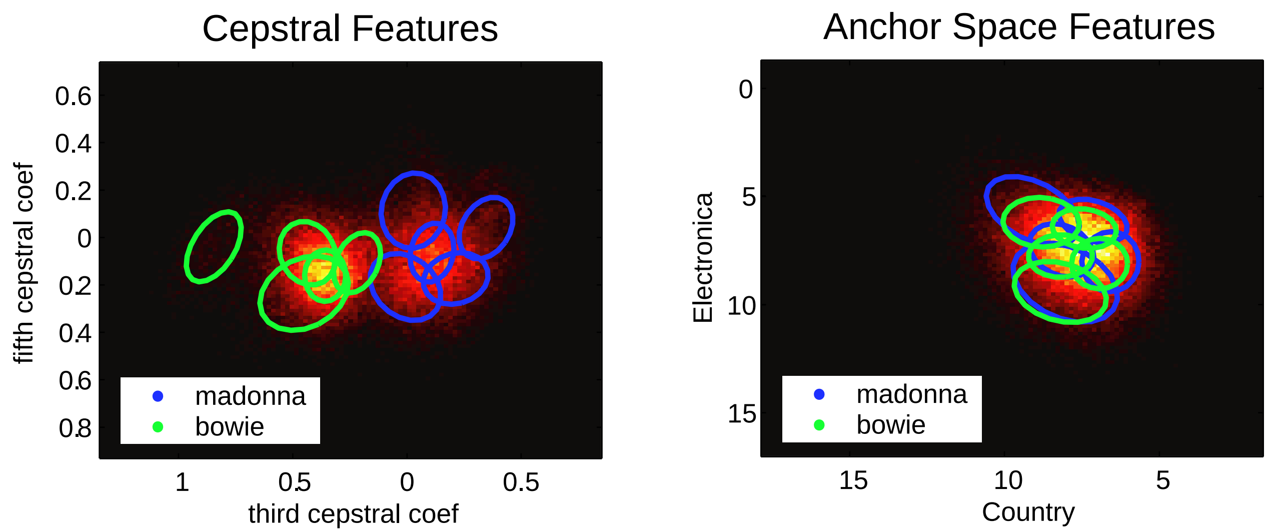
- **A classifier trained for one artist (or genre) will respond partially to a similar artist**
- **Each artist evokes a particular pattern of responses over a set of classifiers**
- **We can treat these classifier outputs as a new feature space in which to estimate similarity**




- **"Anchor space" reflects subjective qualities?**

# Anchor space visualization

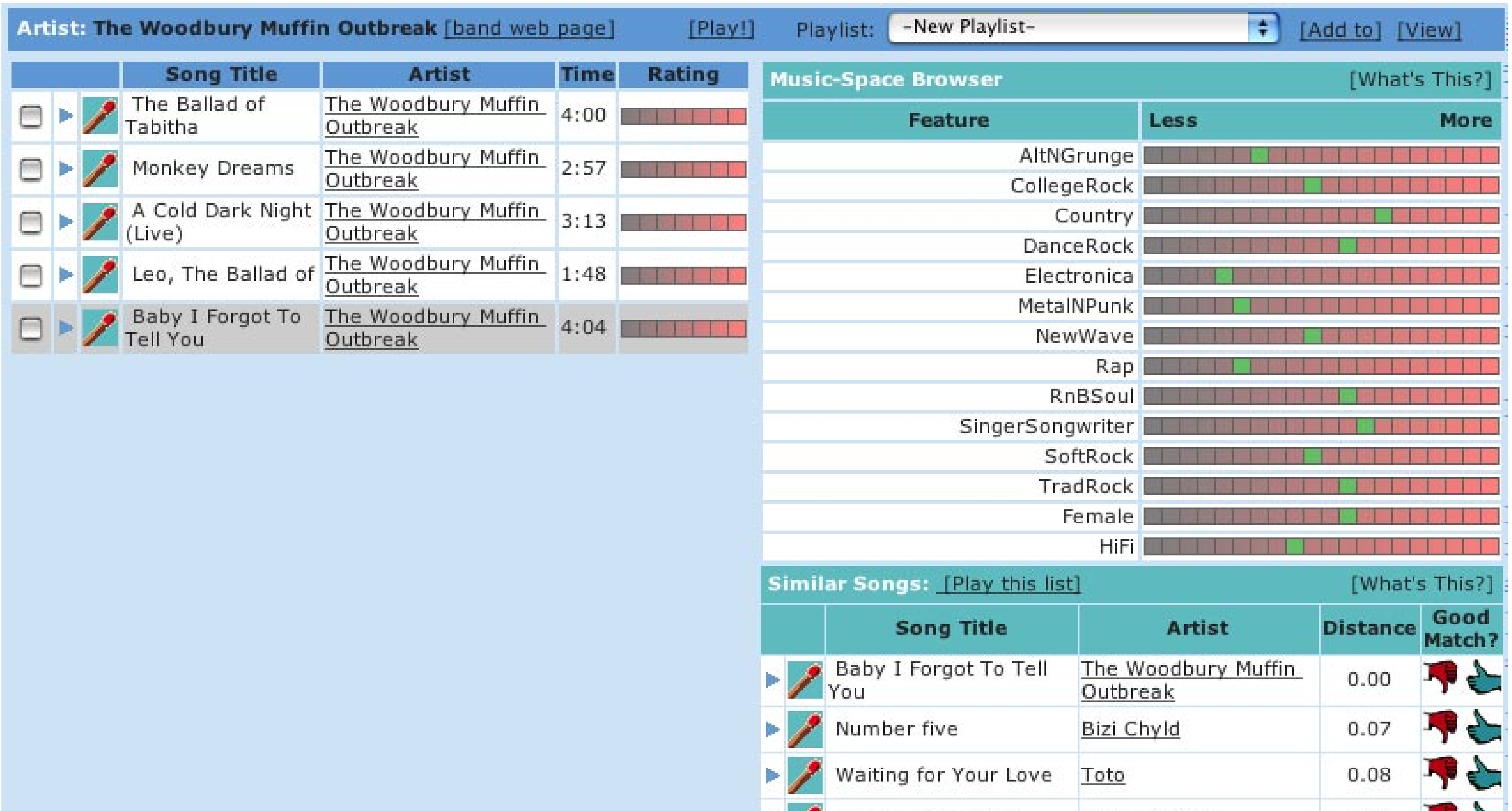
- **Comparing 2D projections of per-frame feature points in cepstral and anchor spaces:**




- each artist represented by 5GMM
- greater separation under MFCCs!
- but: relevant information?

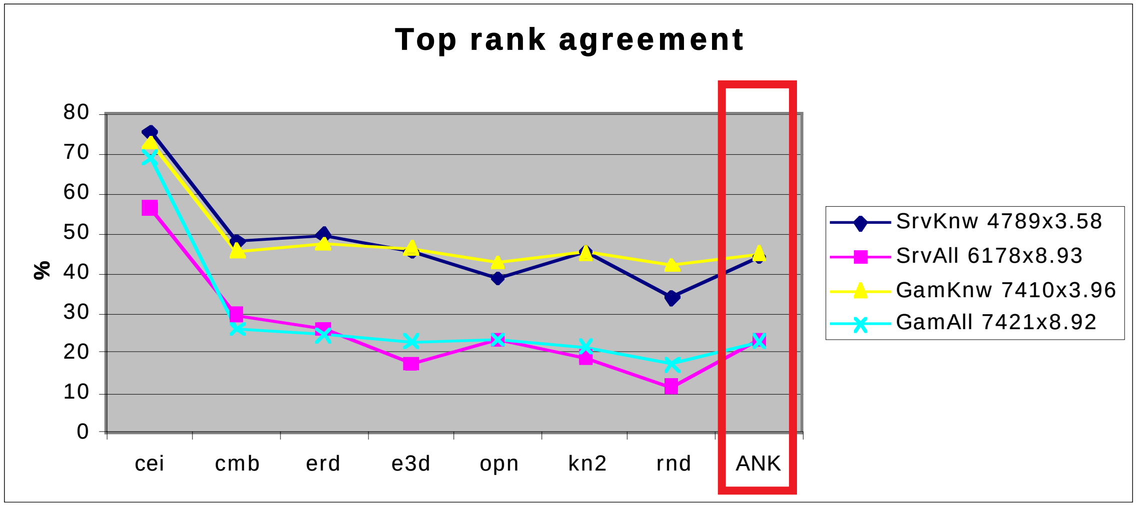
# Playola interface ( www.playola.org )

- **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

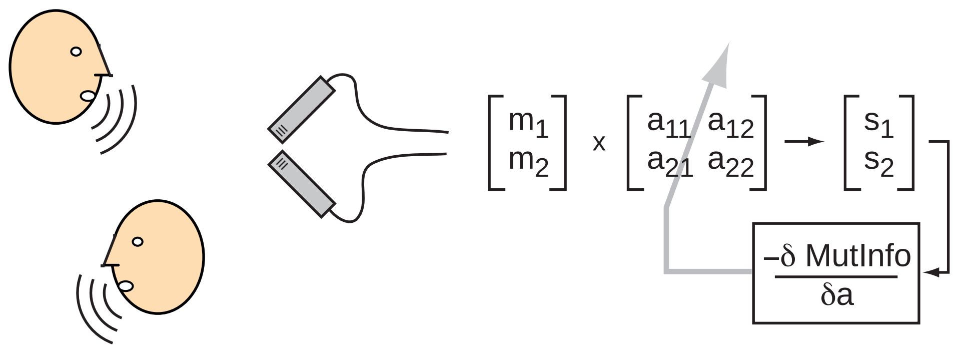
- Structuring
- Search
- Summarization
- Awareness
- Understanding

# *Extra Slides*

# 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