Audio Information Extraction

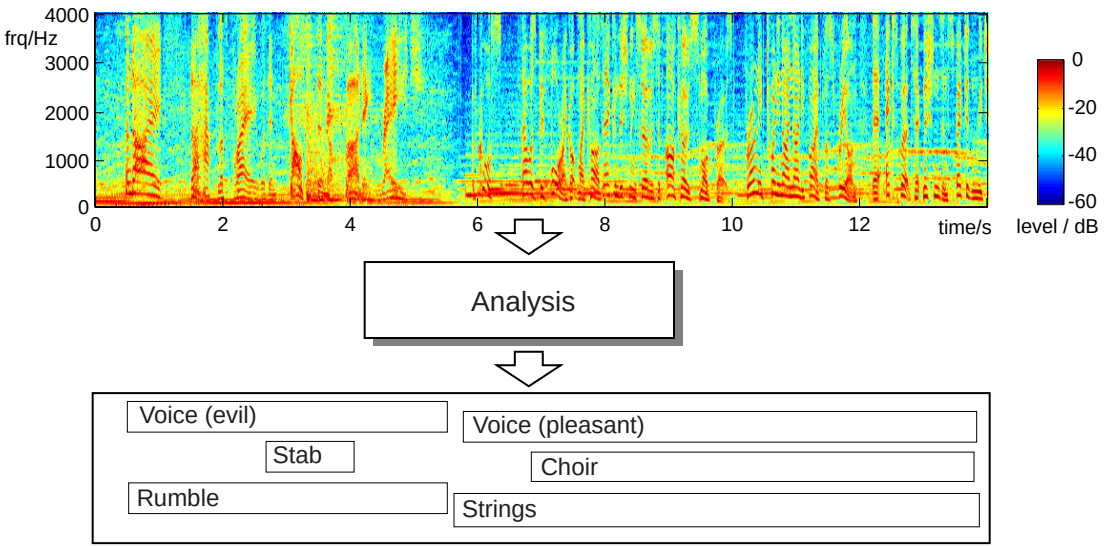
Dan Ellis
<dpwe@ee.columbia.edu>

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

Outline

1. **Audio information extraction**
2. **Speech, music, and other**
3. **General sound organization**
4. **Future work & summary**

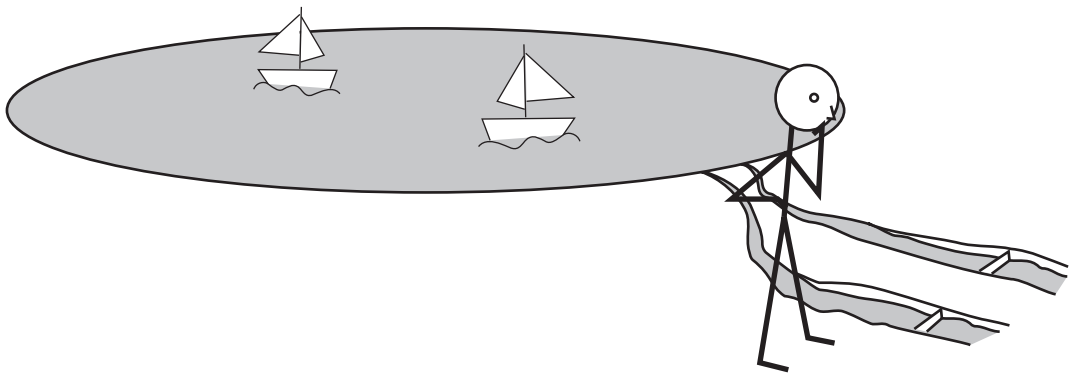
1 Audio Information Extraction



- **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

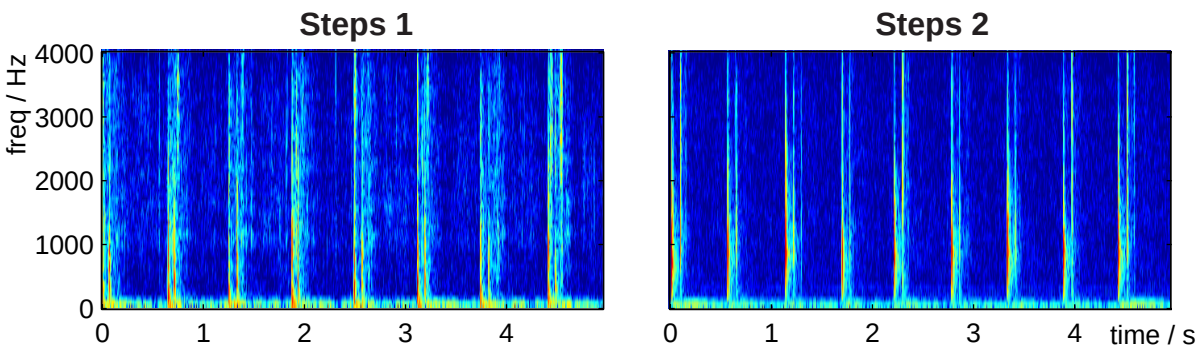
Bregman's lake



"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 ...
- **Disentangling mixtures as primary goal**
  - perfect solution is not possible
  - need knowledge-based *constraints*

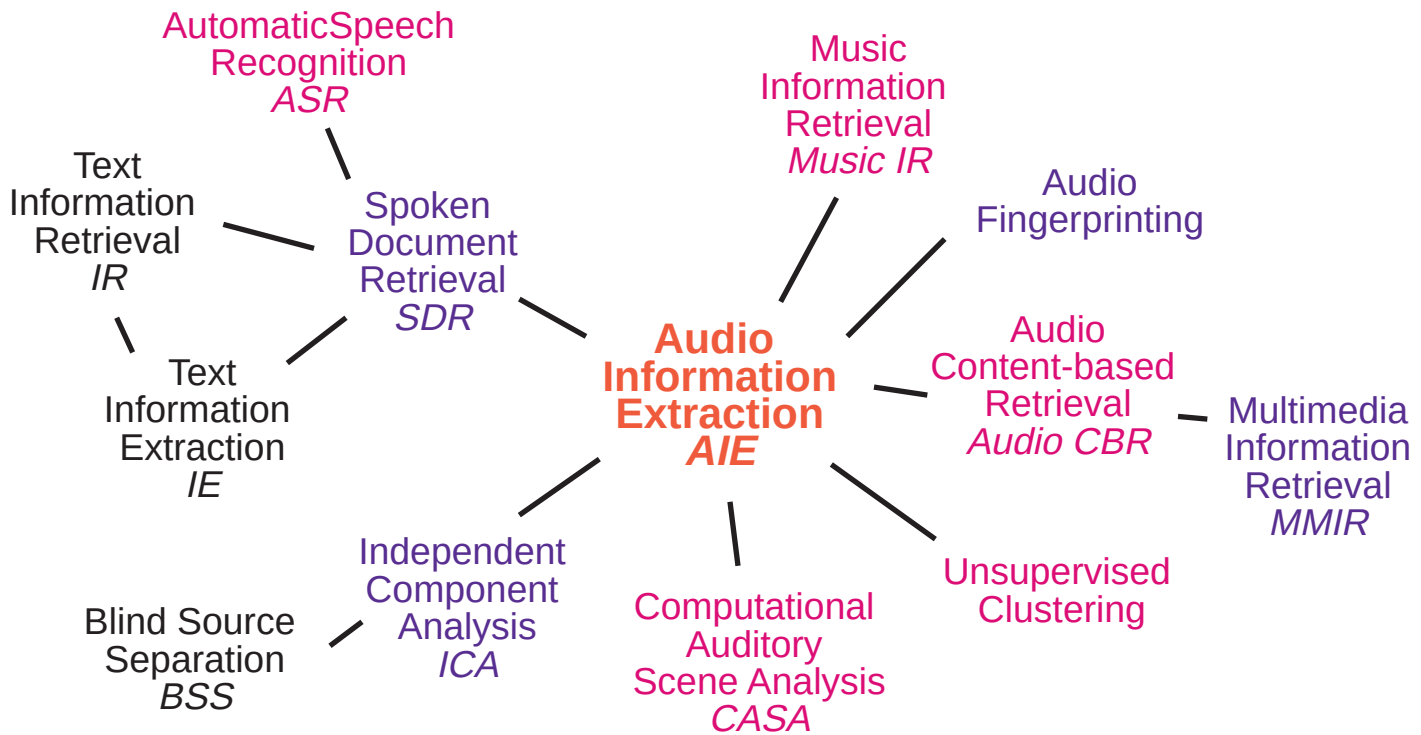
The information in sound



- **Hearing confers evolutionary advantage**
  - optimized to get 'useful' information from sound
- **Auditory perception is *ecologically* grounded**
  - scene analysis is preconscious ($\rightarrow$ illusions)
  - special-purpose processing reflects 'natural scene' properties
  - subjective *not* canonical (ambiguity)

Lab ROSA

Positioning AIE



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

Lab ROSA

AIE Applications

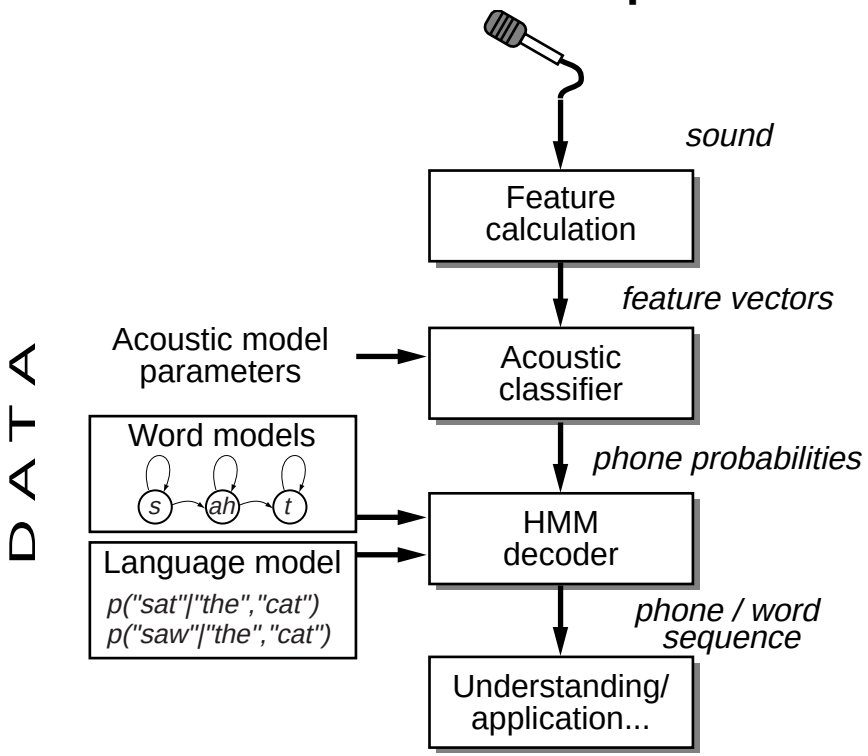
- **Multimedia access**
  - sound as complementary dimension
  - need all modalities for complete information
- **Personal audio**
  - continuous sound capture quite practical
  - different kind of indexing problem
- **Machine perception**
  - intelligence requires awareness
  - necessary for communication
- **Music retrieval**
  - area of hot activity
  - specific economic factors

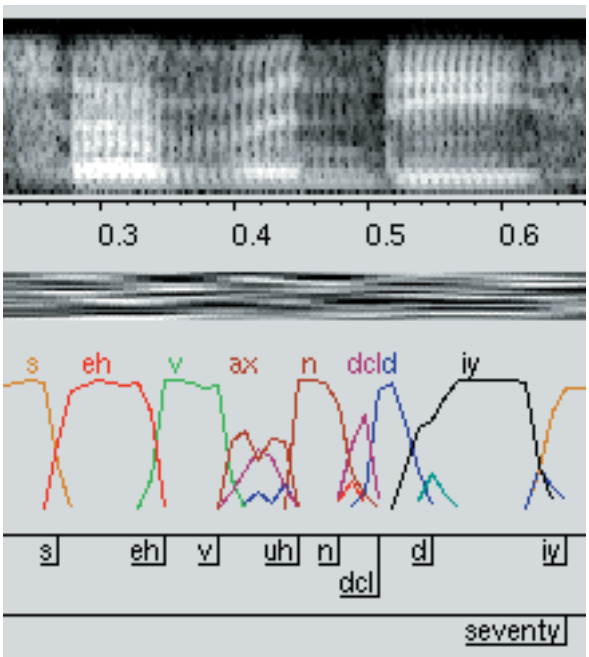
Outline

1. Audio information extraction
2. **Speech, music, and other**
   - Speech recognition: Tandem modeling
   - Multi-speaker processing: Meeting recorder
   - Music classification
   - Other sounds
3. General sound organization
4. Future work & summary

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...**

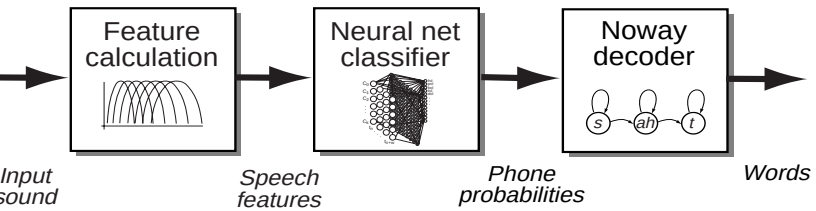
Lab ROSA

Tandem speech recognition

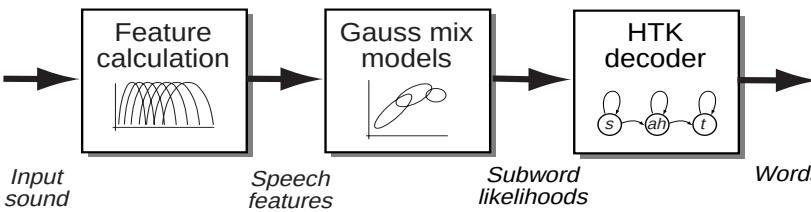
(with Hermansky, Sharma & Sivadas/OGI, Singh/CMU, ICSI)

- **Neural net estimates phone posteriors; but Gaussian mixtures model finer detail**
- **Combine them!**

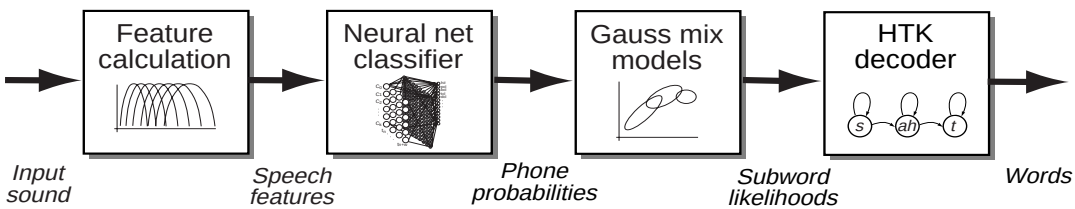
Hybrid Connectionist-HMM ASR



Conventional ASR (HTK)



Tandem modeling

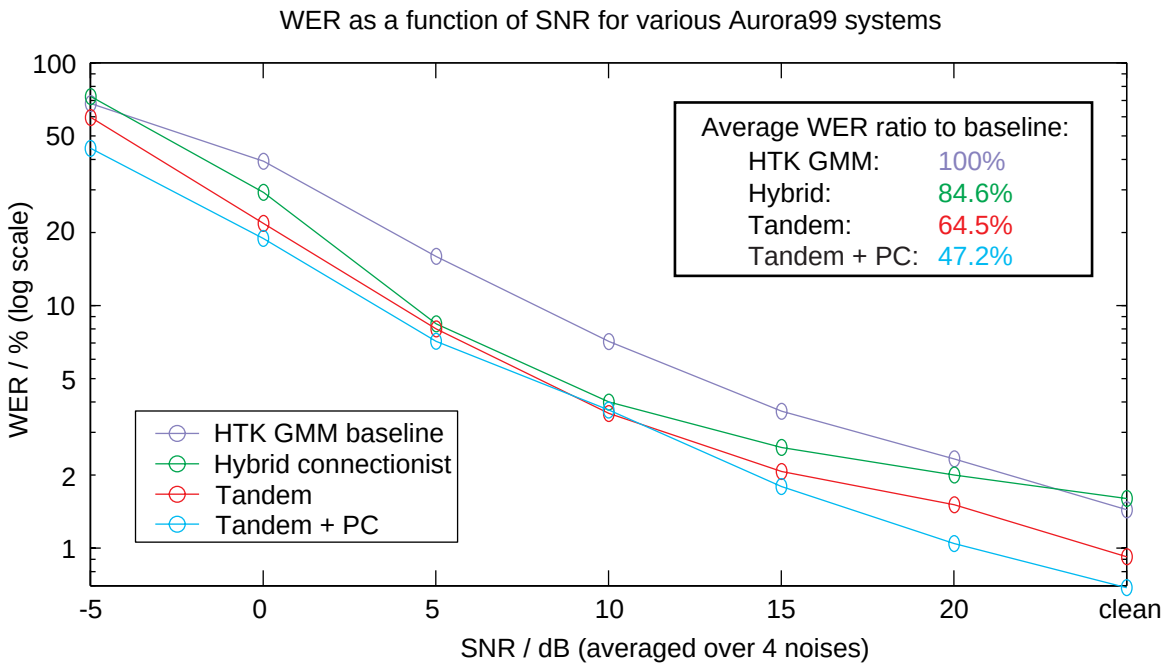


- **Train net, then train GMM on net output**
  - GMM is ignorant of net output 'meaning'

Lab ROSA

Tandem system results

- **It works very well ('Aurora' noisy digits):**

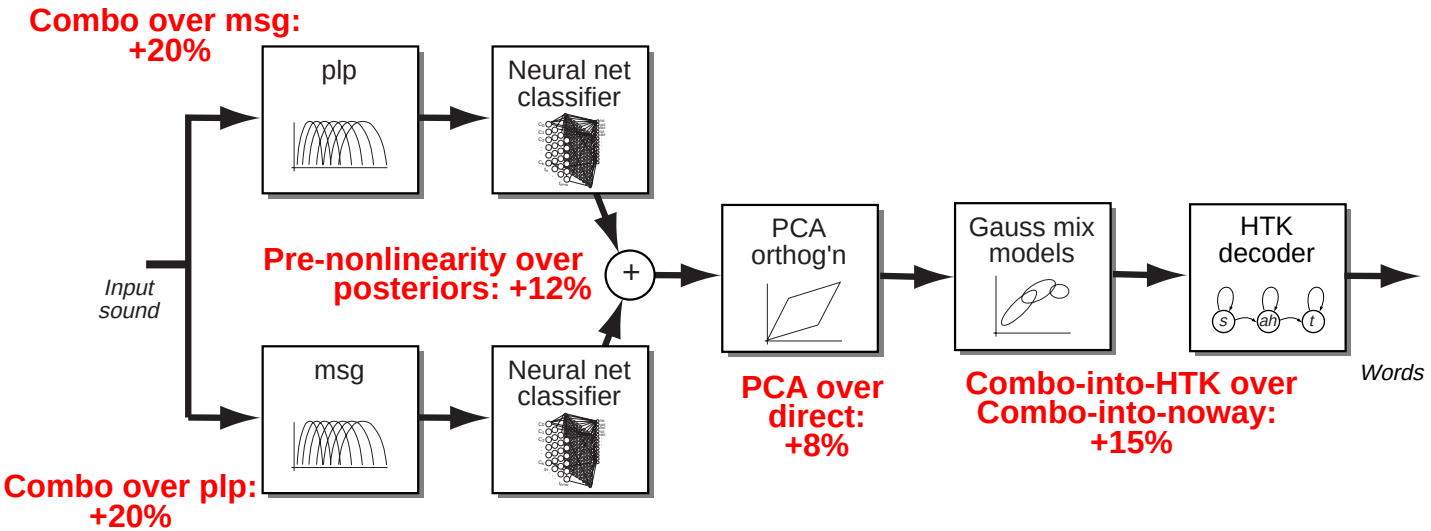


System-features	*Avg. WER 20-0 dB*	*Baseline WER ratio*
HTK-mfcc	13.7%	100%
Neural net-mfcc	9.3%	84.5%
Tandem-mfcc	7.4%	64.5%
Tandem-msg+plp	**6.4%**	**47.2%**

Lab ROSA

Relative contributions

- **Approx relative impact on baseline WER ratio for different components:**

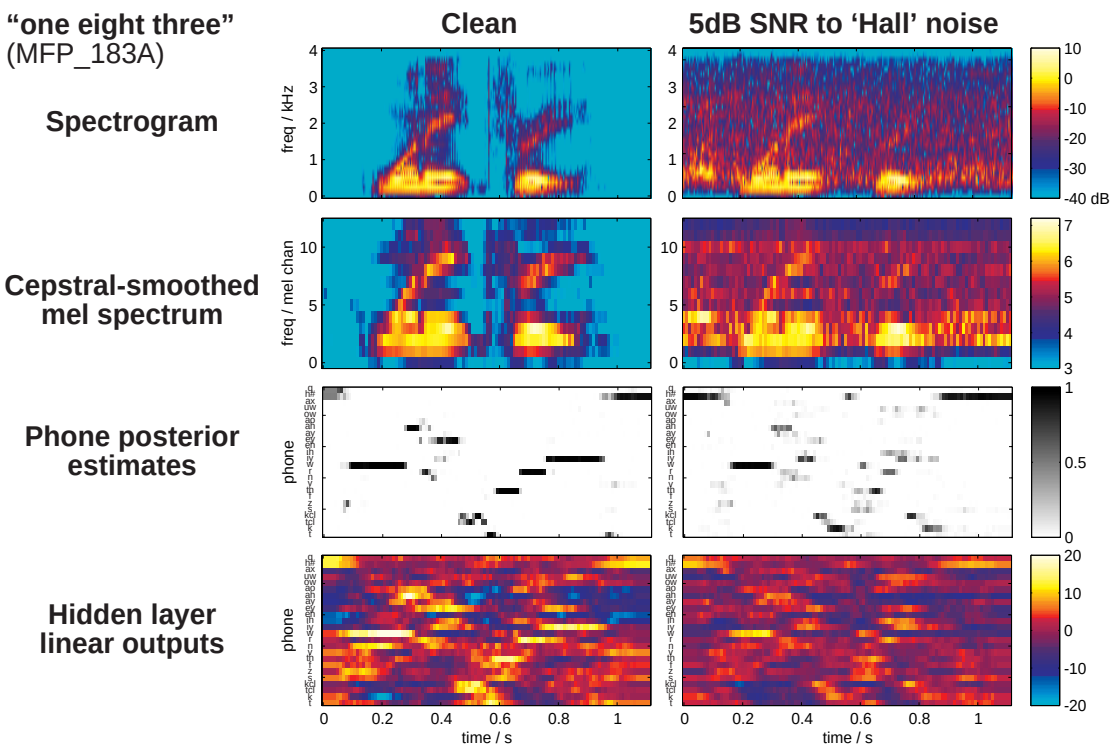


Tandem combo over HTK mfcc baseline: +53%

Lab ROSA

Inside Tandem systems: What's going on?

- **Visualizations of the net outputs**

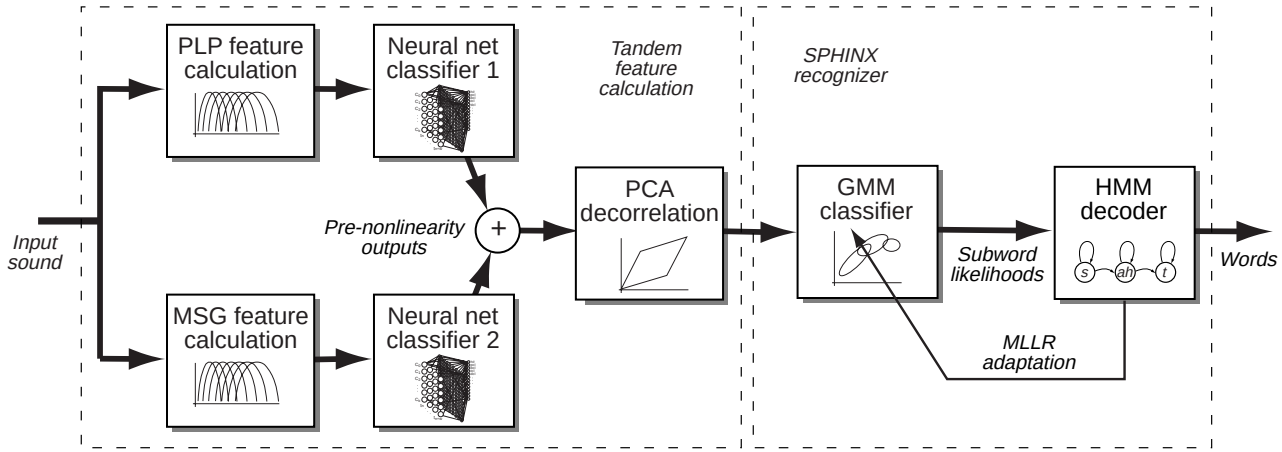


- **Neural net normalizes away noise**

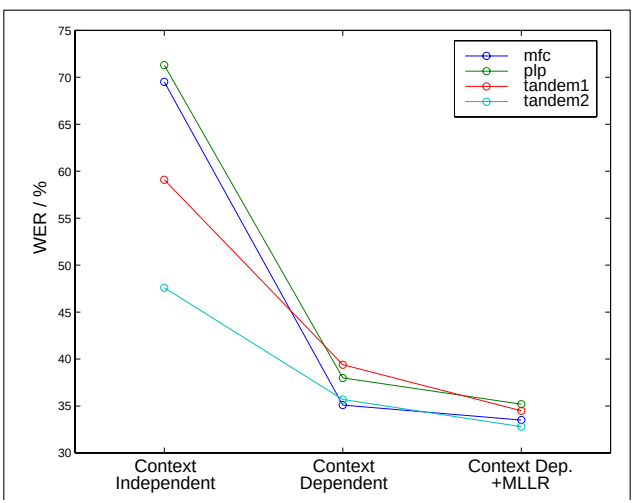
Lab ROSA

Tandem for large vocabulary recognition

- **CI Tandem front end + CD LVCSR back end**



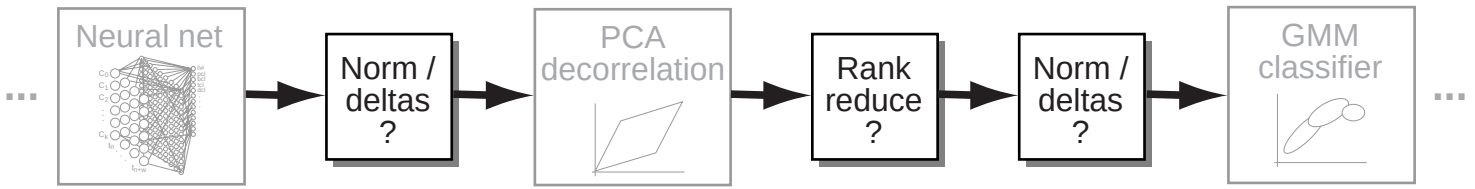
- **Tandem benefits reduced:**



‘Tandem-domain’ processing

(with Manuel Reyes)

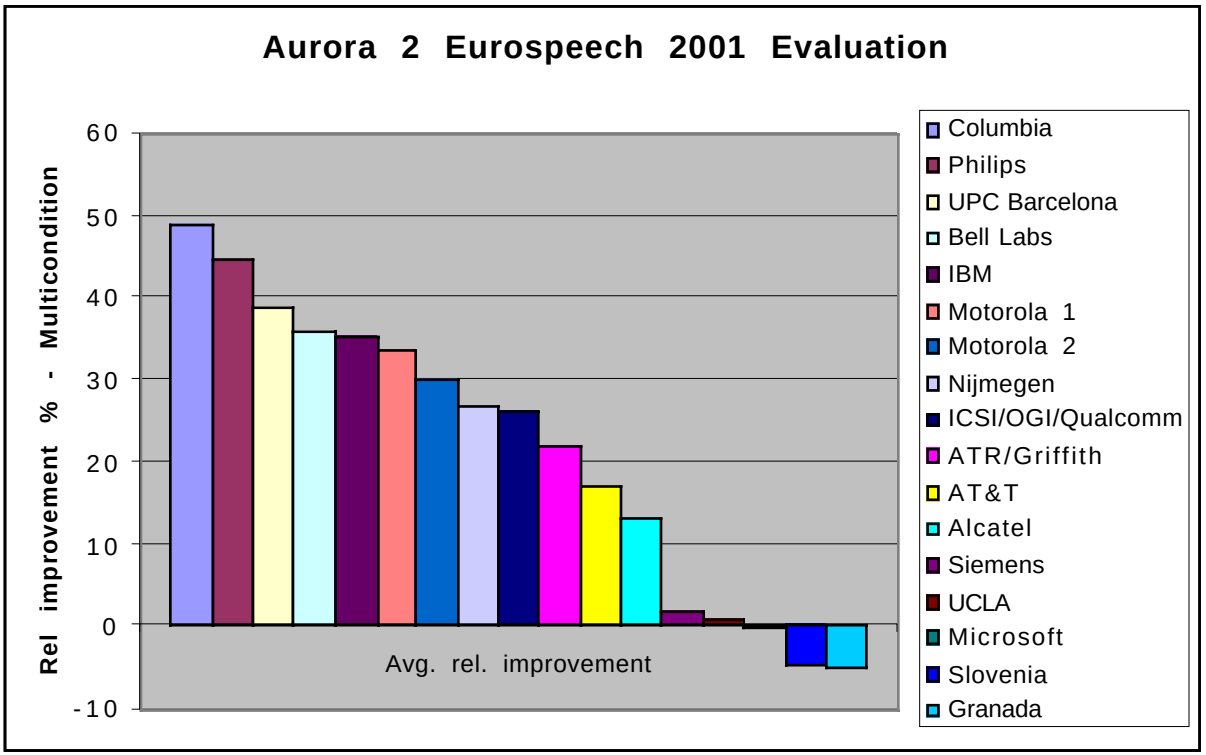
- **Can we improve the ‘tandem’ features with conventional processing (deltas, normalization)?**



- **Somewhat..**

Processing	*Avg. WER 20-0 dB*	*Baseline WER ratio*
Tandem PLP mis-match baseline (24 els)	11.1%	70.3%
Rank reduce @ 18 els	11.8%	77.1%
Delta → PCA	9.7%	60.8%
PCA → Norm	9.0%	58.8%
Delta → PCA → Norm	**8.3%**	**53.6%**

Tandem vs. other approaches



- **50% of word errors corrected over baseline**
- **Beat 'bells and whistles' system using large-vocabulary techniques**

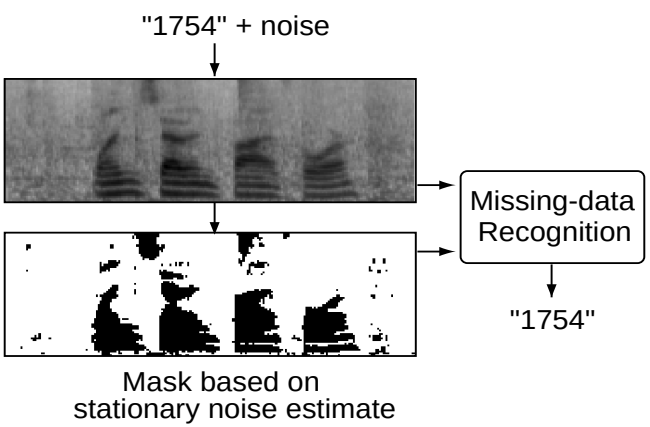
Missing data recognition

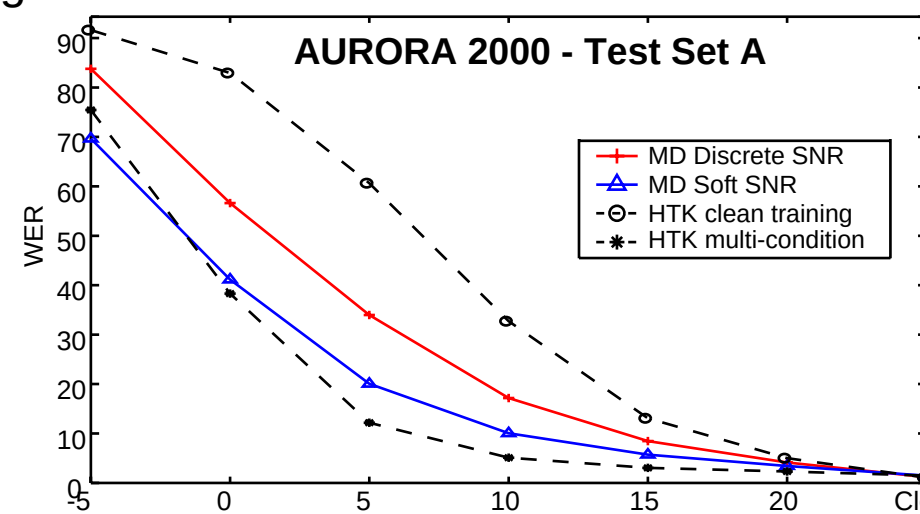
(Cooke, Green, Barker @ Sheffield)

- **Energy overlaps in time-freq. hide features**
  - some observations are effectively missing
- **Use missing feature theory...**
  - integrate over missing data x_m under model M

$$p(x|M) = \int p(x_p|x_m, M)\, p(x_m|M)\, dx_m$$

- **Effective in speech recognition**
  - trick is finding good/bad data mask

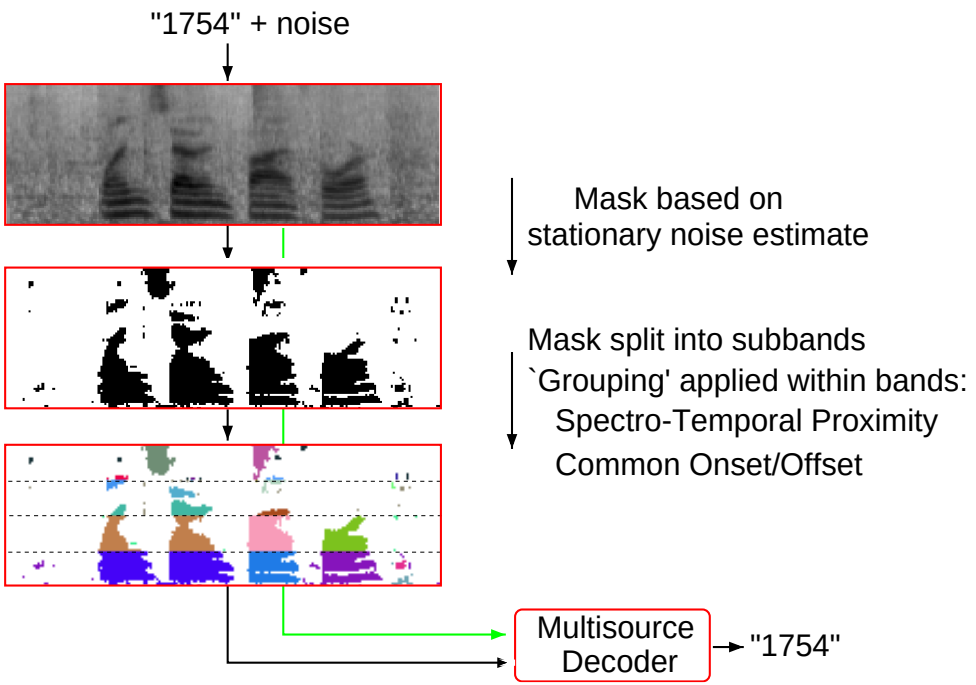




Maximum-likelihood data mask

(Jon Barker @ Sheffield)

- **Search of sound-fragment interpretations**



- also search over data mask K:

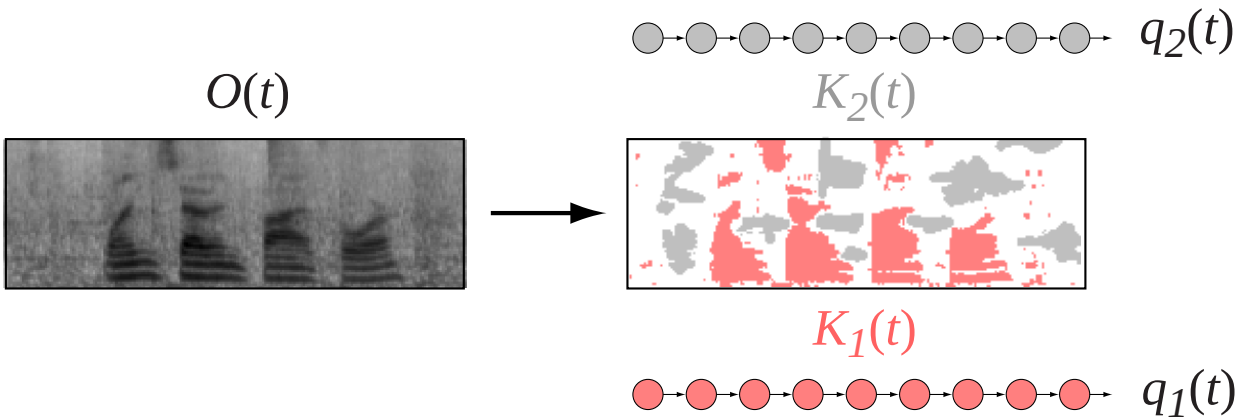
$$p(M, K|x) \propto p(x|K, M)p(K|M)p(M)$$

- **Modeling mask likelihoods $p(K)$ ...**

Lab ROSA

Multi-source decoding

- **Search for more than one source**



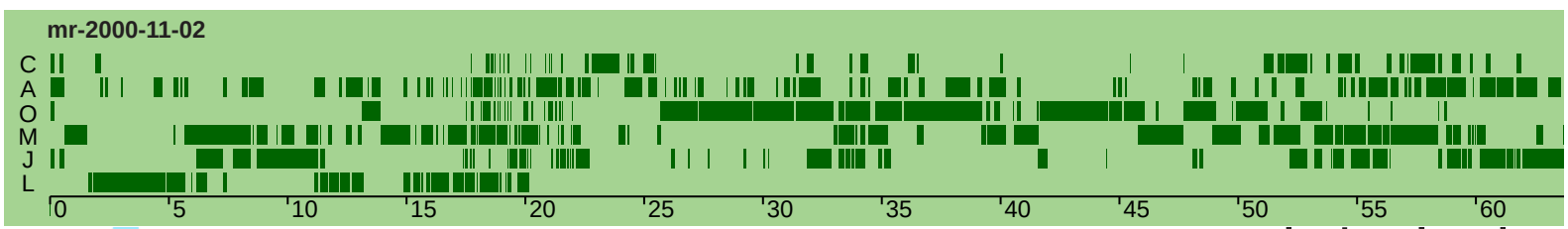
- **Mutually-dependent data masks**
- **Use CASA processing to propose masks**
  - locally coherent regions
  - $p(K|q)$
- **Theoretical vs. practical limits**

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

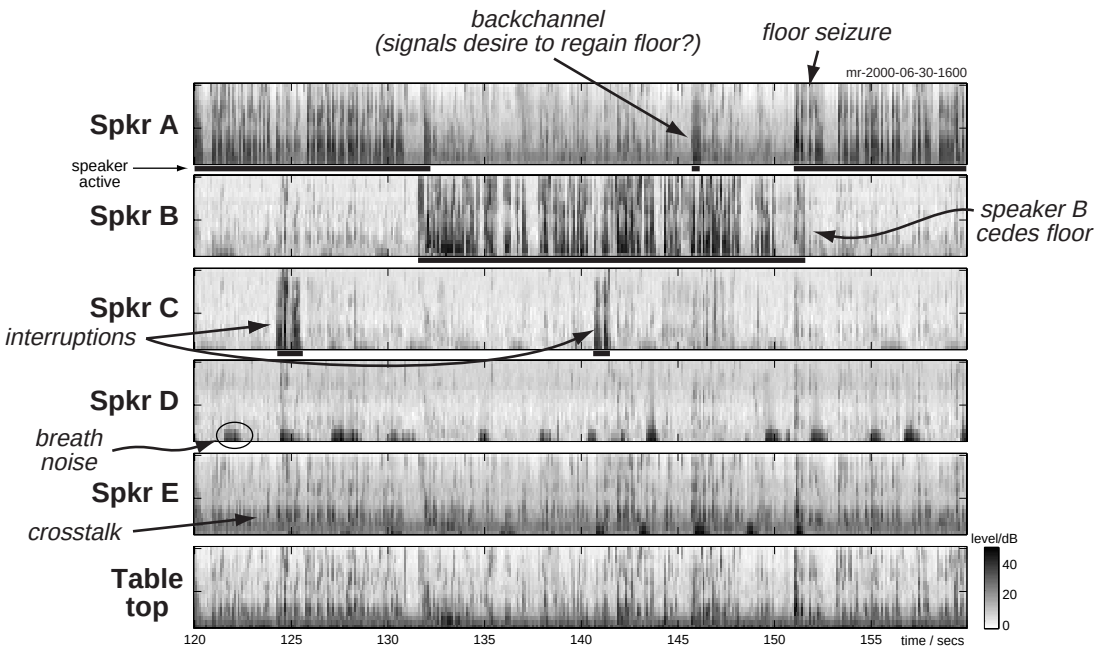


Lab ROSA

Crosstalk cancellation

(with Sam Keene)

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

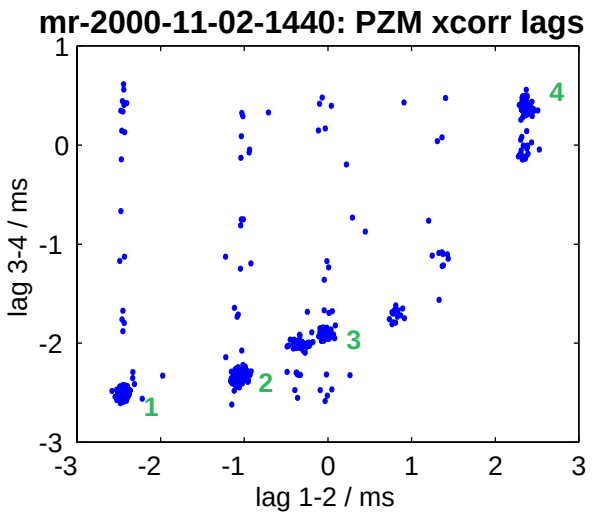


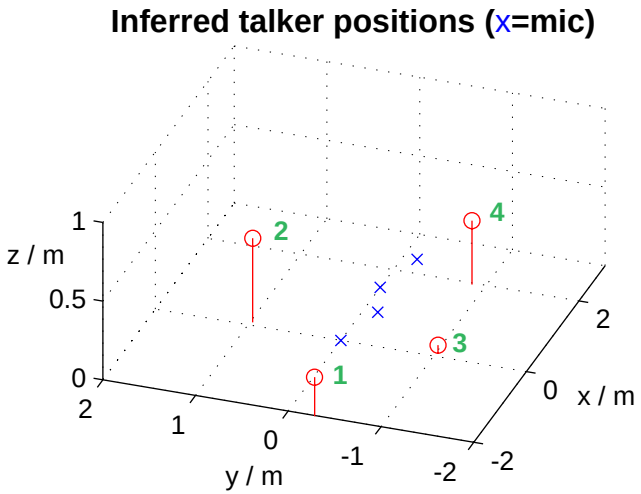
- **Noisy crosstalk model:** $\mathbf{m} = \mathbf{C} \cdot \mathbf{s} + \mathbf{n}$
- **Estimate subband C_{Aa} from A's peak energy**
  - ... then linear inversion

Speaker localization

(with Huan Wei Hee)

- **Tabletop mics form an array; time differences locate speakers**





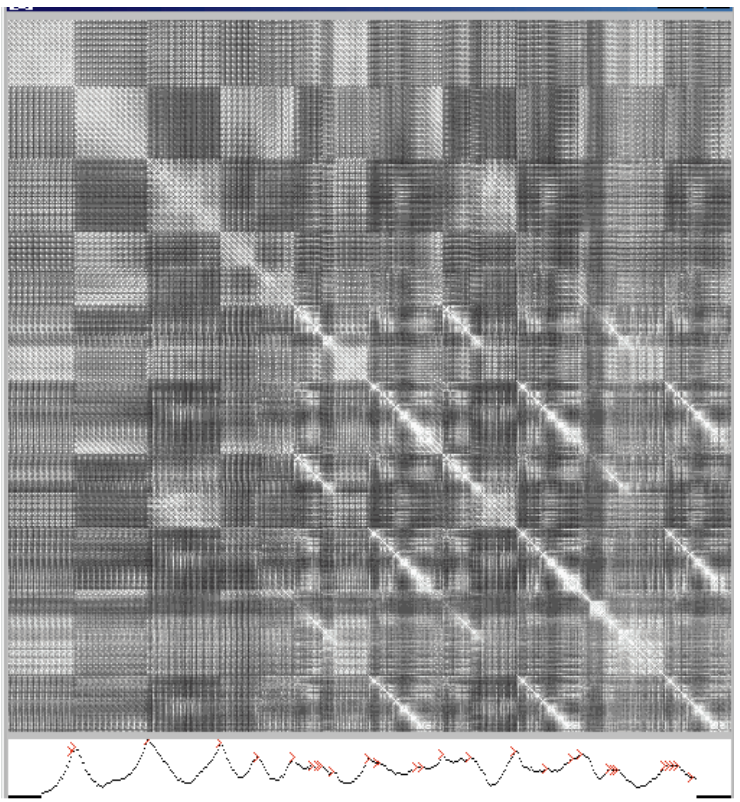
- **Ambiguity:**
  - mic positions not fixed
  - speaker motions
- **Detect speaker activity, overlap**

Lab ROSA

Music analysis: Structure recovery

(with Rob Turetsky)

- **Structure recovery by similarity matrices (after Foote)**

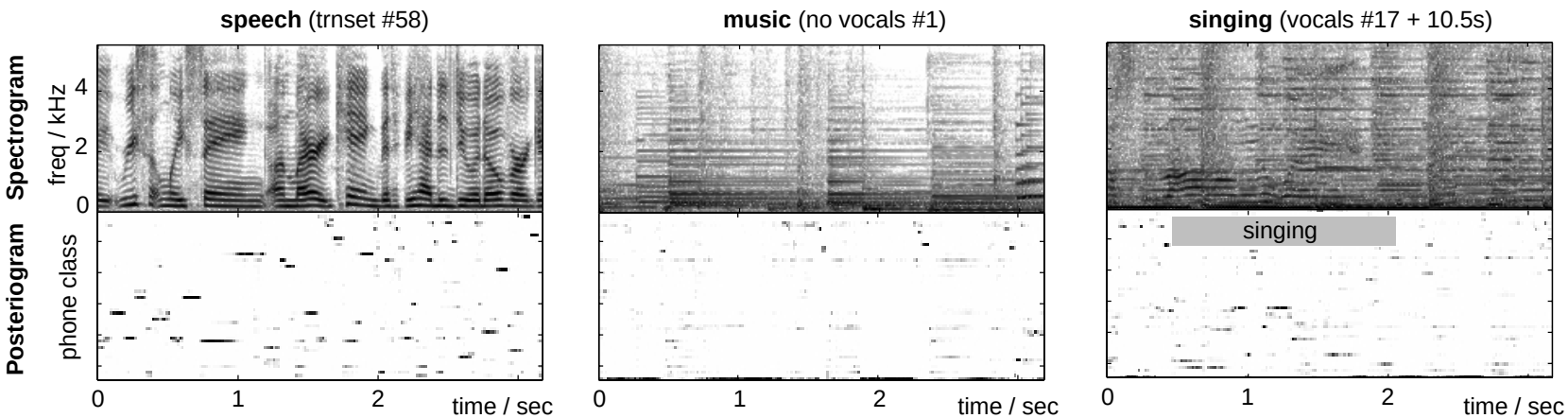


- similarity distance measure?
- segmentation & repetition structure
- interpretation at different scales: notes, phrases, movements
- incorporating musical knowledge: 'theme similarity'

Music analysis: Lyrics extraction

(with Adam Berenzweig)

- **Vocal content is highly salient, useful for retrieval**
- **Can we find the singing? Use an ASR classifier:**

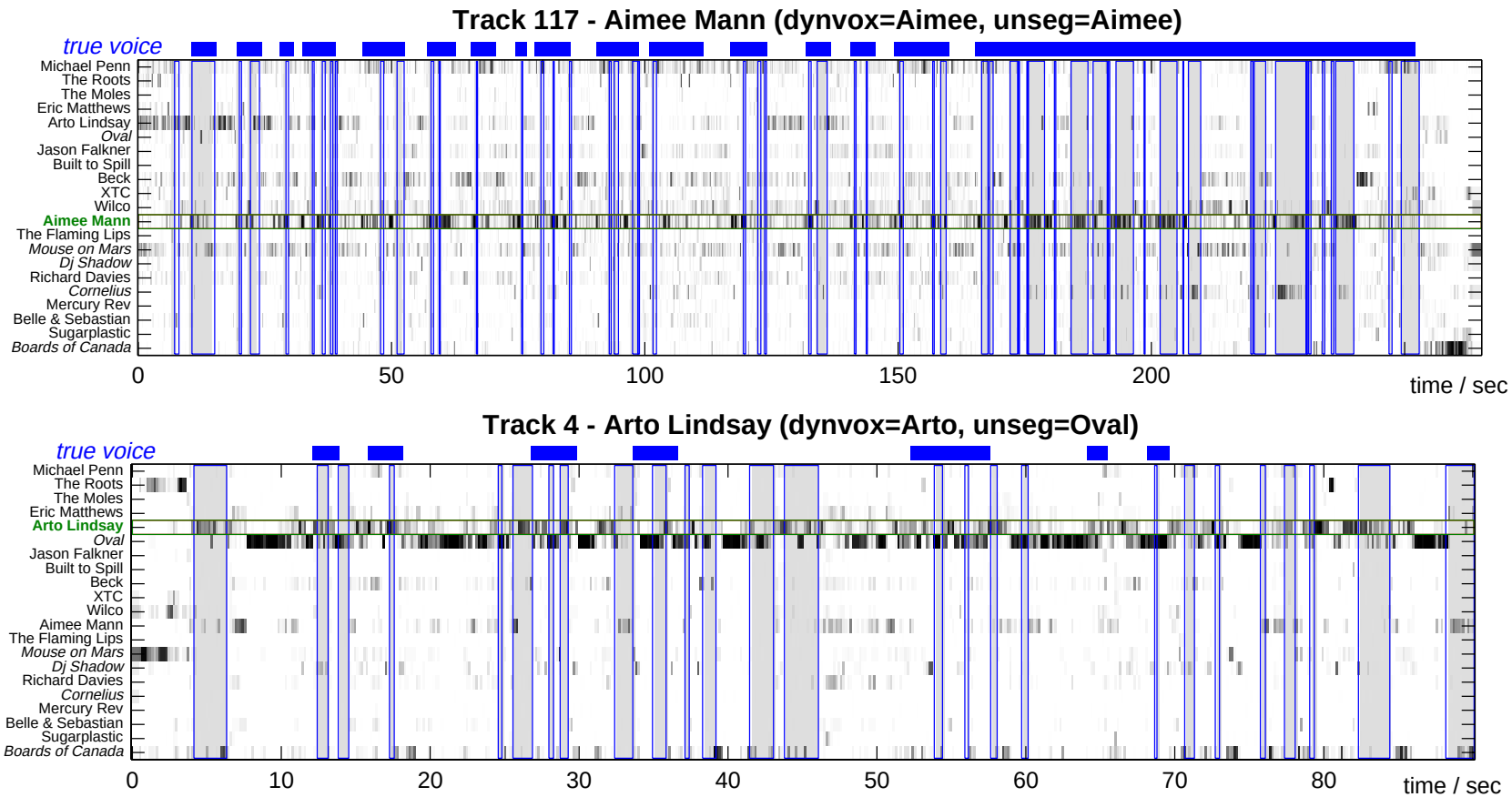


- **Frame error rate ~20% for segmentation based on posterior-feature statistics**
- **Lyric segmentation + transcribed lyrics → training data for lyrics ASR...**

Lab ROSA

Artist similarity

- **Train network to discriminate specific artists:**



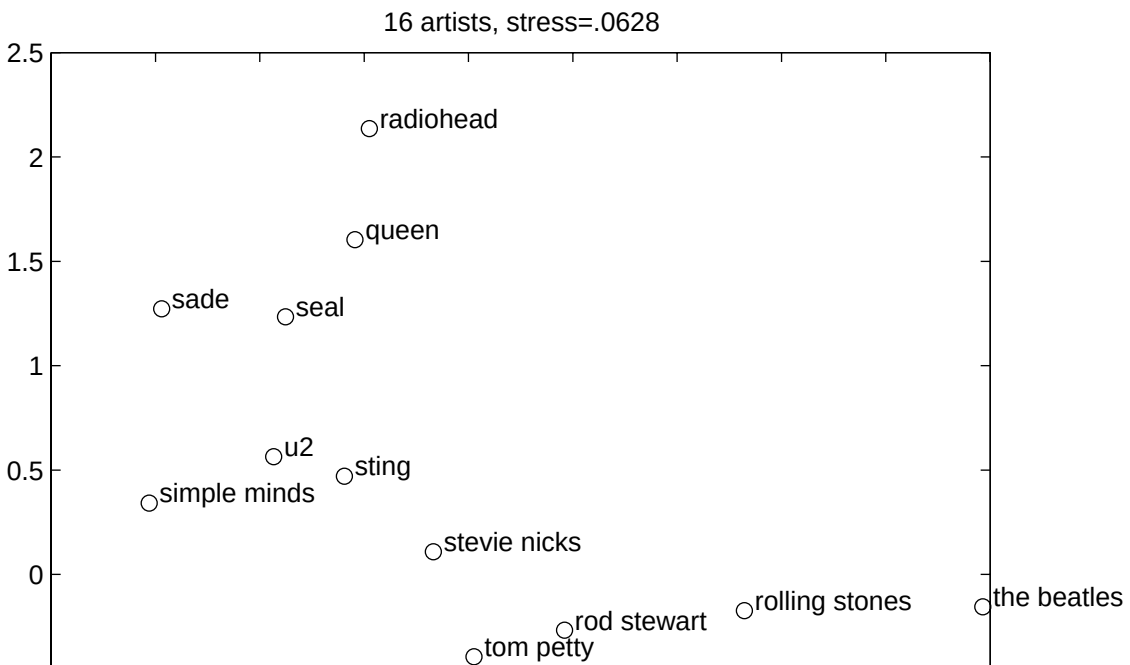
- **Focus on vocal segments for consistency**
  - accuracy (21 artists) 57% → 65%

Lab ROSA

Ground truth in artist similarity

(with Berenzweig, Whitman@MIT, Lawrence@NEC)

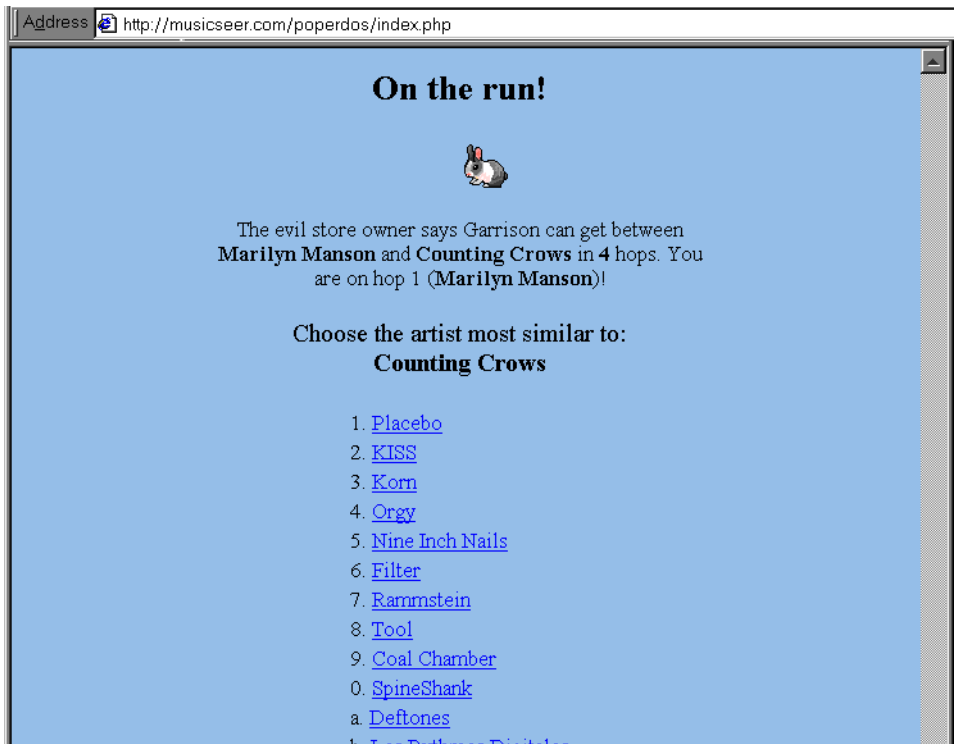
- **For training audio-based similarity measures**
- **Extend partial ratings to a complete set?**
- **e.g. Erdös distance**
  - music guide → first-order similarities
  - hop count → total distance



Lab ROSA

Subjective verification of metrics

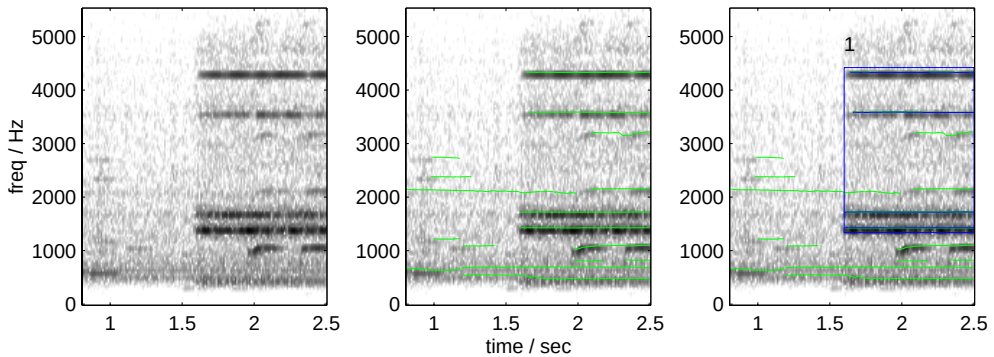
- How can we choose between different proposed ‘ground truth’ metrics?
- Collect subjective judgments via web ‘game’:



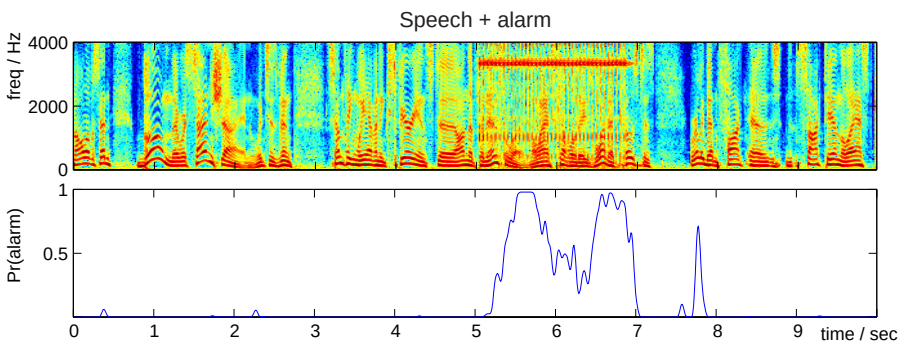
- Compare user responses to predictions from different models

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



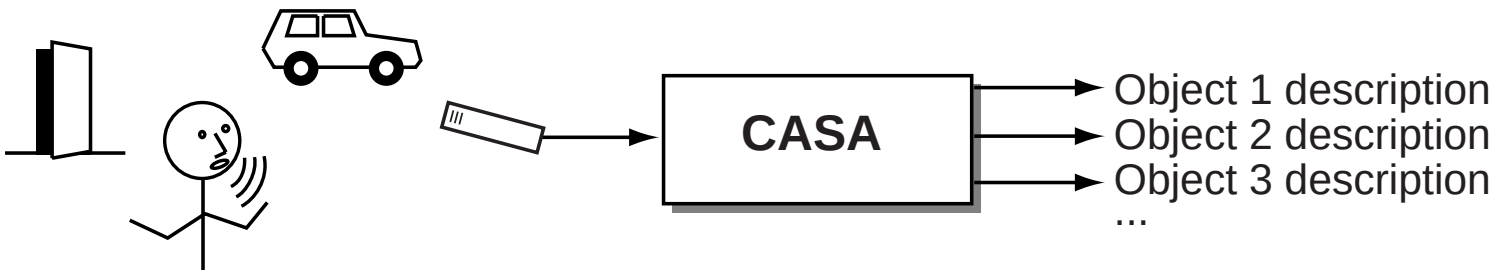
  - cepstral coefficients are easy to model

Lab ROSA

Outline

1. Sound organization
2. Speech, music, and other
3. **General sound organization**
   - Computational Auditory Scene Analysis
   - Audio Information Retrieval
4. Future work & summary

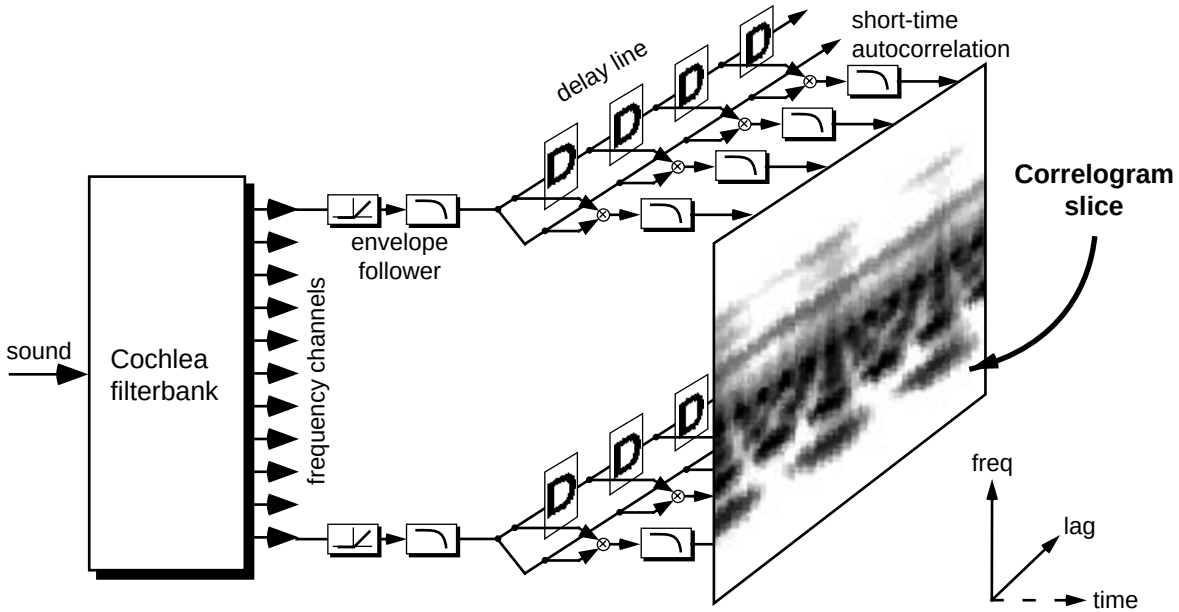
Computational Auditory Scene Analysis (CASA)



- **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'
  - ... can we implement them?

CASA front-end processing

- **Correlogram:**
  Loosely based on known/possible physiology



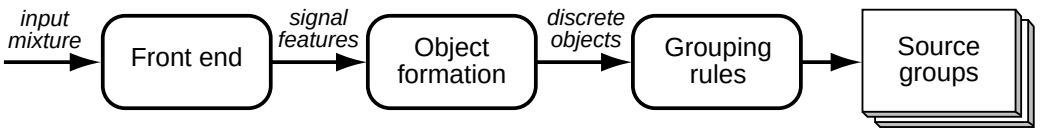
- linear filterbank cochlear approximation
- static nonlinearity
- zero-delay slice is like spectrogram
- periodicity from delay-and-multiply detectors

Lab ROSA

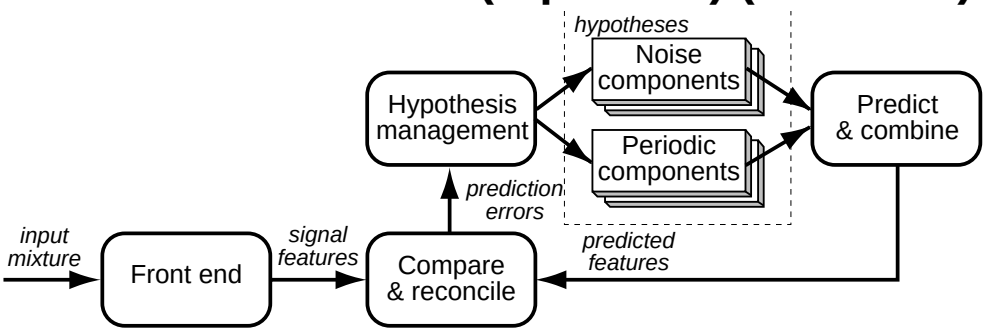
Adding top-down cues

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

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



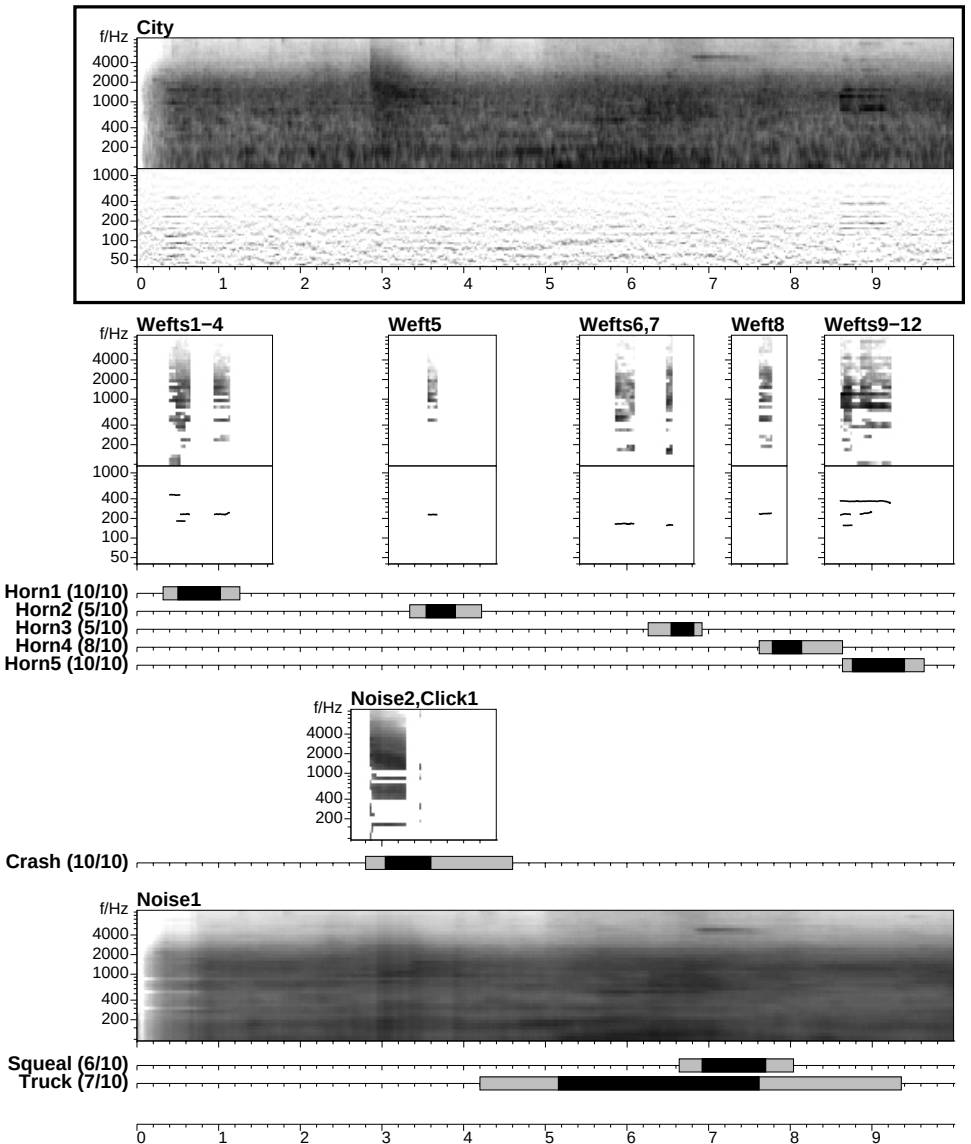
vs. Prediction-driven (top-down) (PDCASA)



- **Motivations**
  - detect non-tonal events (noise & click elements)
  - support ‘restoration illusions’...
- **Machine Learning for sound models**
  - corpus of isolated sounds?

Lab ROSA

PDCASA and complex scenes

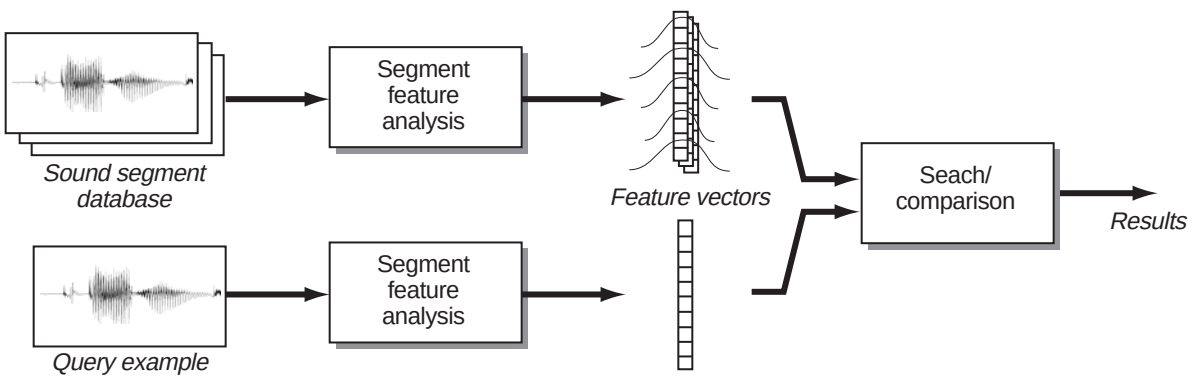


Lab ROSA

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

Audio Retrieval: Results

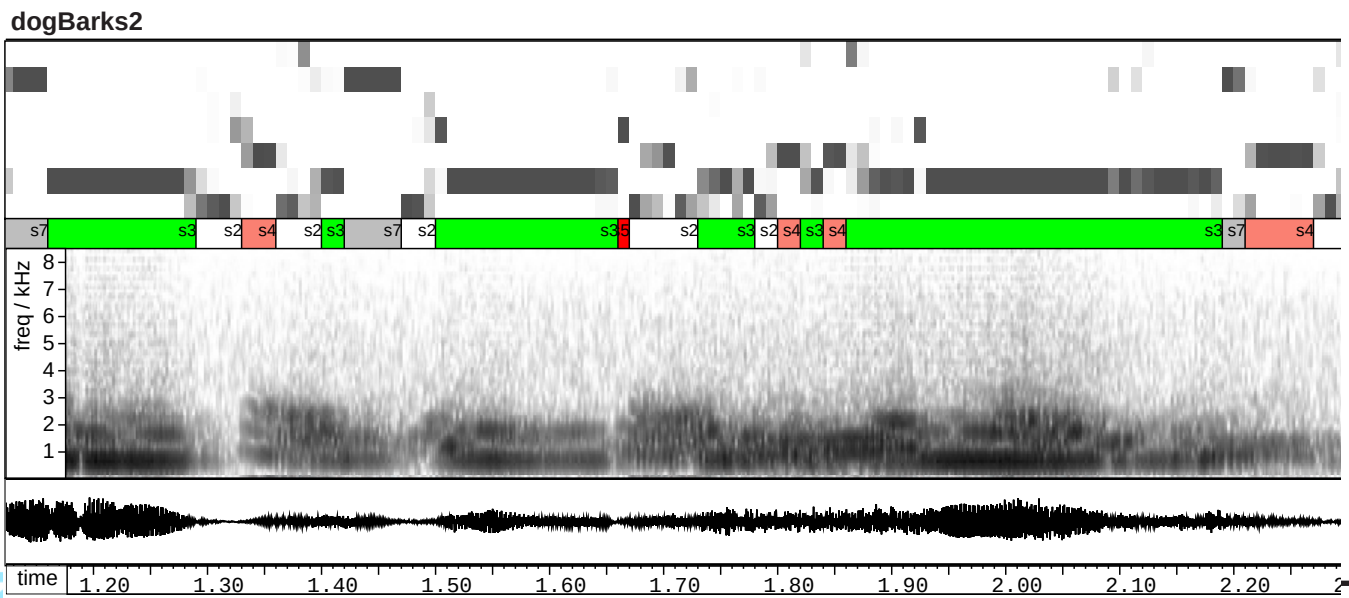
- **Musclefish corpus**
  - most commonly reported set
- **Features**
  - mfcc, brightness, bandwidth, pitch ...
  - no temporal sequence structure
- **Results:**
  - 208 examples, 16 classes

Global features: 41% corr *HMM models: 81% corr.*

	Mu	*Sp*	*Env*	*An*	*Mec*	*Mu*	*Sp*	*Env*	*An*	*Mec*
Musical	59/ **46**		**24**	**2**	**19**	136/ **6**		**2**	**1**	**5**
Speech		11/ **6**	**4**	**5**		**1**	14/ **2**	**5**	**3**	**1**
Eviron.			7/ **2**			**1**		7/	**1**	
Animals			**2**	1/	**2**				4/	**1**
Mechan	**1**		**4**	**1**	8/ **4**	**3**		**3**		12/

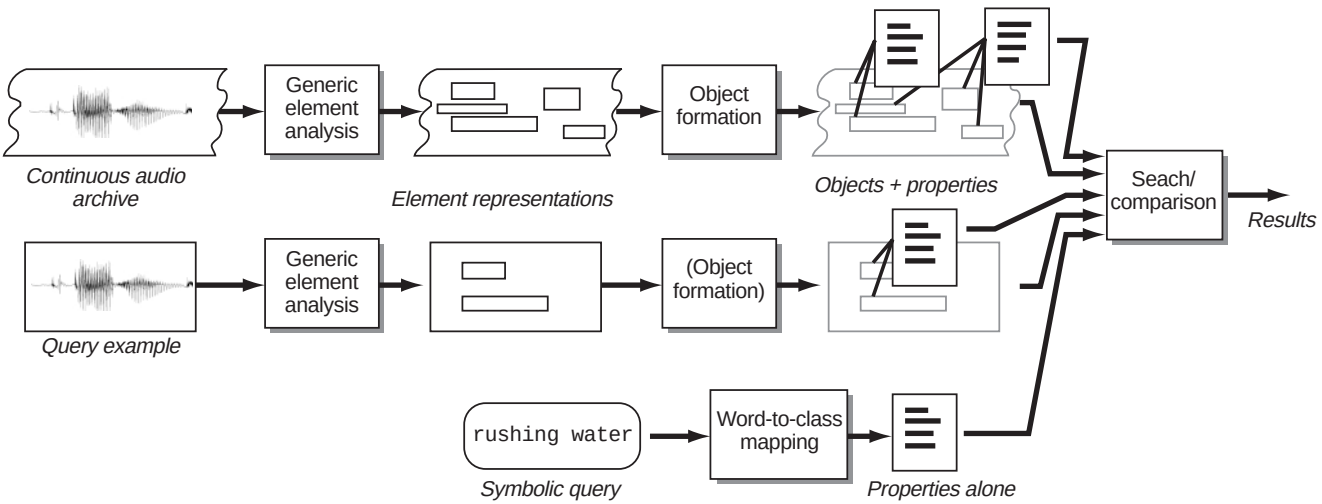
What are the HMM states?

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



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

Outline

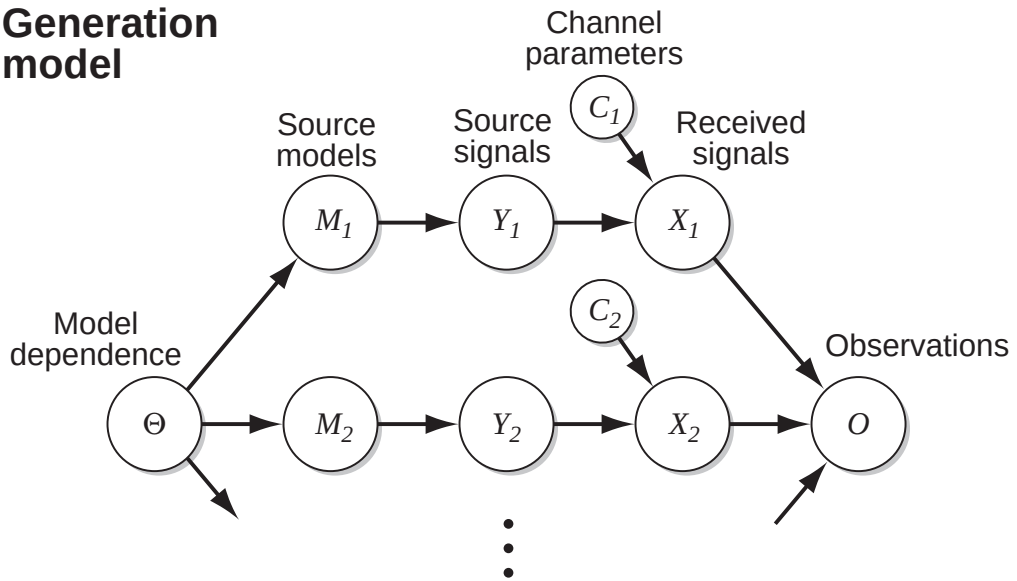
1. Sound organization
2. Speech, music, and other
3. General sound organization
4. **Future work & summary**

Lab ROSA

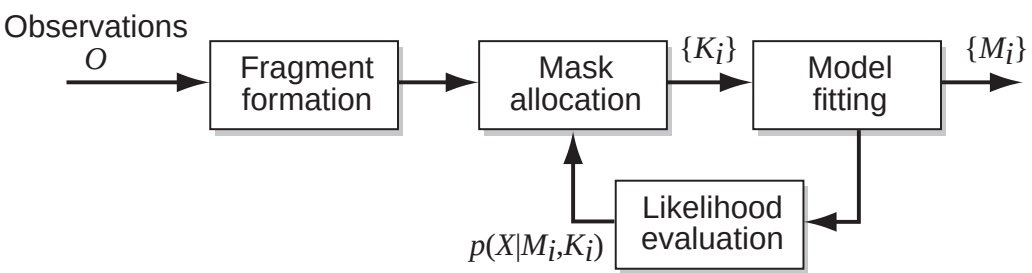
General sound mixtures

- **Search for generative explanation:**

Generation model



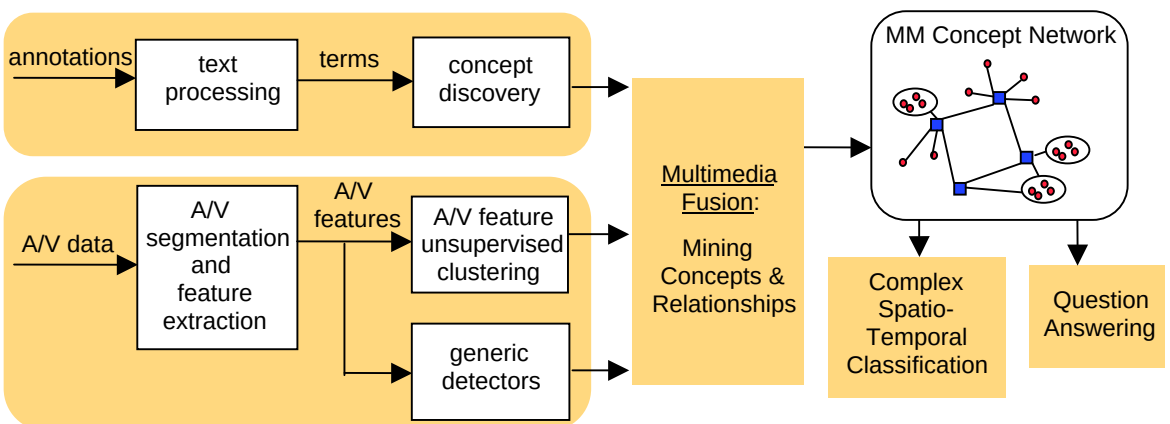
Analysis structure



Automatic audio-video analysis

(with Prof. Shih-Fu Chang, Prof. Kathy McKeown)

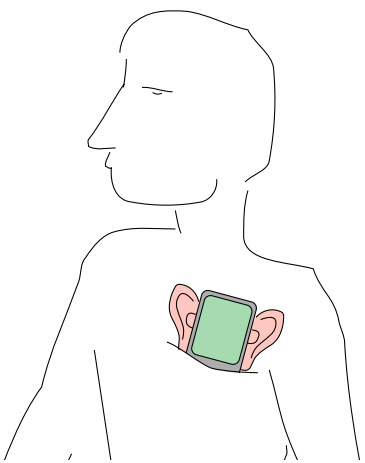
- **Documentary archive management**
  - huge ratio of raw-to-finished material
  - costly manual logging
  - missed opportunities for cross-fertilization
- **Problem: term ↔ signal mapping**
  - training corpus of past annotations
  - interactive semi-automatic learning
  - need object-related features



Lab ROSA

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

Summary: Audio Info Extraction

- **Sound carries information**
  - useful and detailed
  - often tangled in mixtures
- **Various important general classes**
  - Speech: activity, recognition
  - Music: segmentation, clustering
  - Other: detection, description
- **General processing framework**
  - Computational Auditory Scene Analysis
  - Audio Information Retrieval
- **Future applications**
  - Ubiquitous intelligent indexing
  - Intelligent monitoring & description