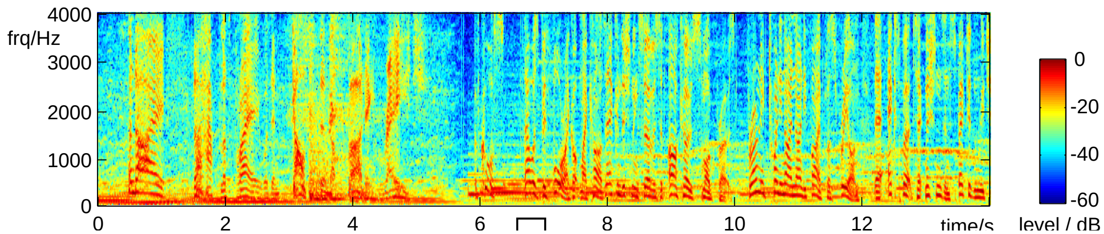
Scene Analysis for Speech and Audio Recognition

1. **Sound, Mixtures & Learning**
2. **Computational Auditory Scene Analysis**
3. **Recognizing Speech in Noise**
4. **Using Models in Parallel**
5. **The Listening Machine**

Dan Ellis <dpwe@ee.columbia.edu>

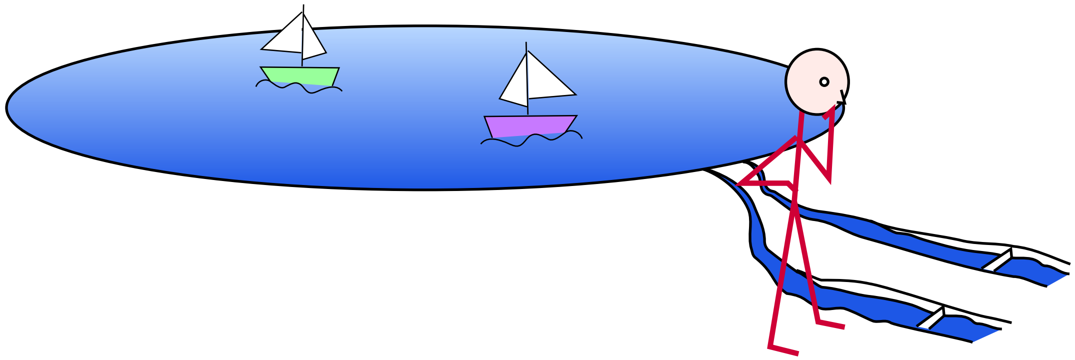
Laboratory for Recognition and Organization of Speech and Audio
(Lab**ROSA**)
Columbia University, New York
http://labrosa.ee.columbia.edu/

1 Sound, Mixtures & Learning



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

The problem with recognizing mixtures



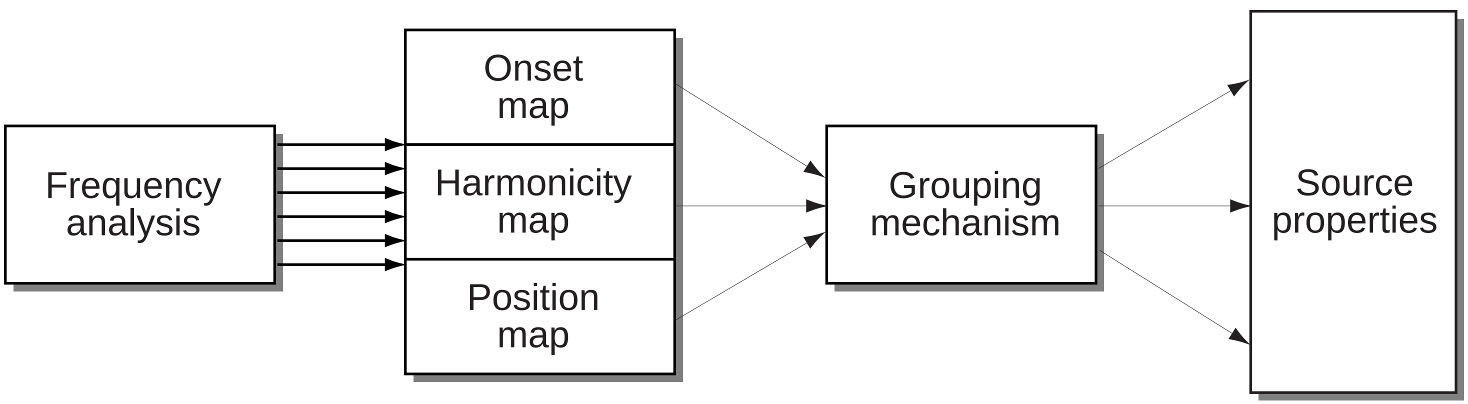
"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- **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)

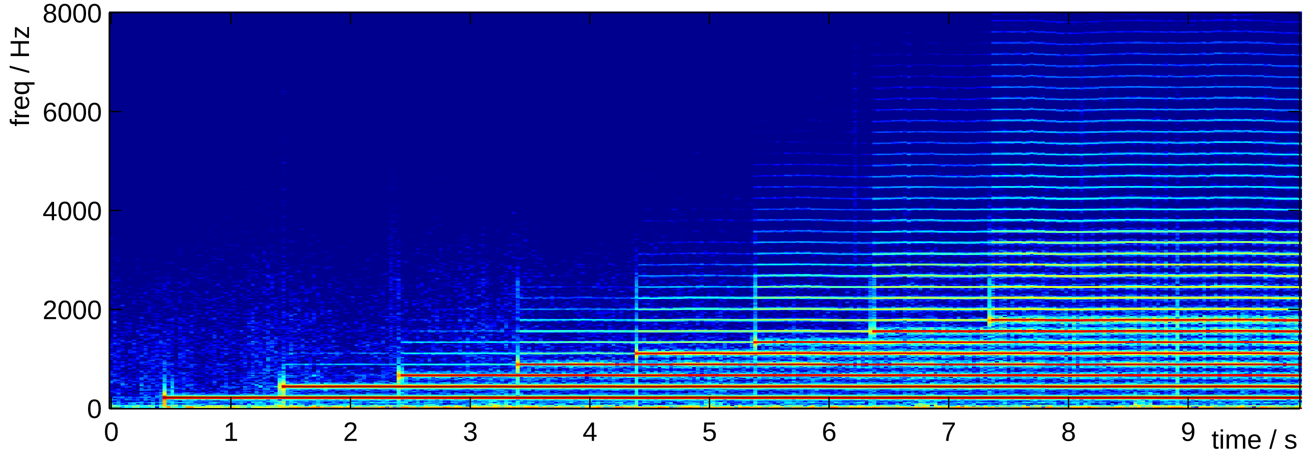
- **How do people analyze sound mixtures?**
  - break mixture into small *elements* (in time-freq)
  - elements are *grouped* in to sources using *cues*
  - sources have aggregate *attributes*
- **Grouping 'rules' (Darwin, Carlyon, ...):**
  - cues: common onset/offset/modulation, harmonicity, spatial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



(after Darwin, 1996)

Cues to simultaneous grouping

- **Elements + attributes**

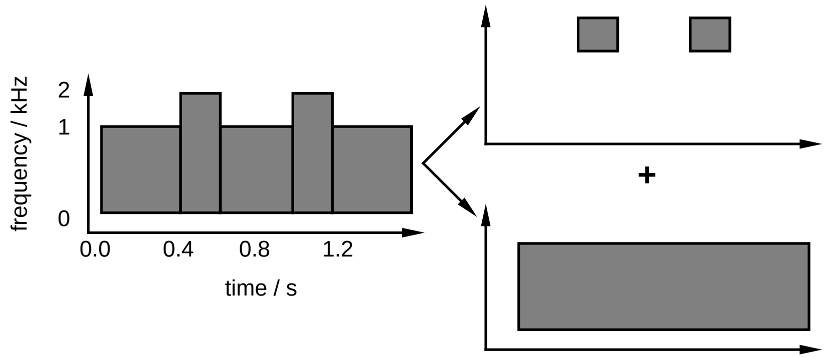


- **Common onset**
  - simultaneous energy has common source
- **Periodicity**
  - energy in different bands with same cycle
- **Other cues**
  - spatial (ITD/IID), familiarity, ...

The effect of context

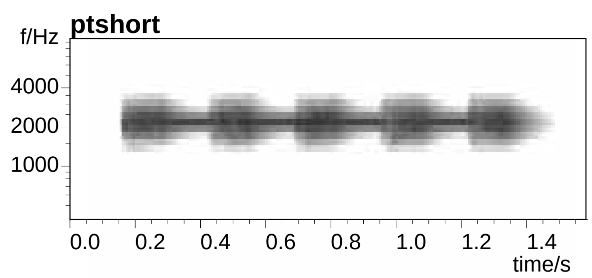
- **Context can create an 'expectation': i.e. a bias towards a particular interpretation**

- **Bregman's old-plus-new principle:**



  - a change is preferably interpreted as addition

- **E.g. the continuity illusion**



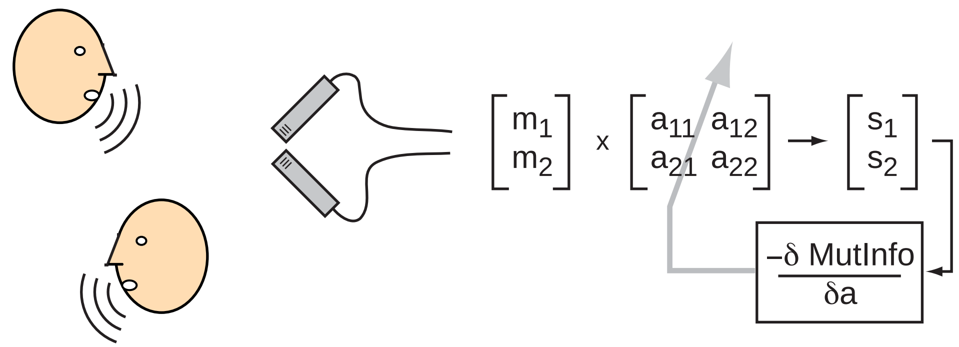
Approaches to sound mixture recognition

- **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?
  - divide signal into fragments, then use missing-data recognition

Independent Component Analysis (ICA)

(Bell & Sejnowski 1995 etc.)

- **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

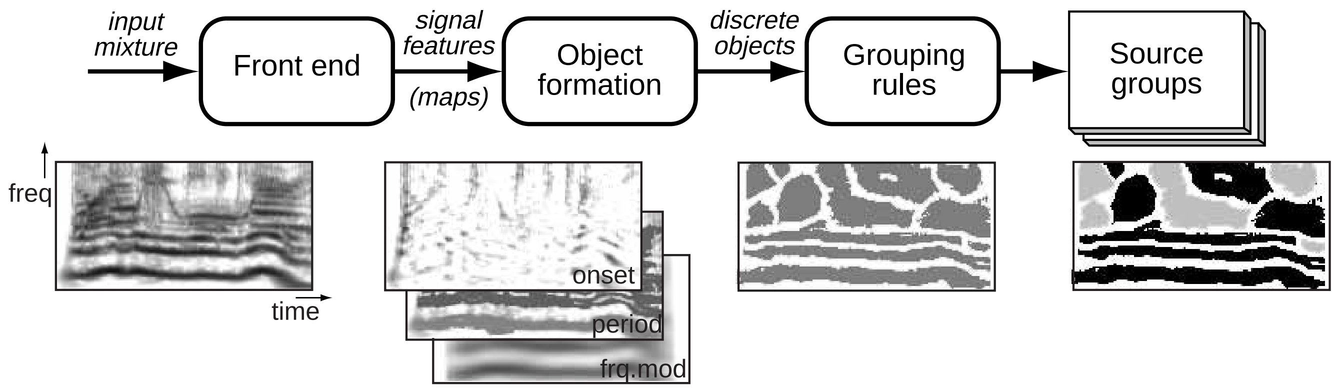
Outline

1. Sound, Mixtures & Learning
2. **Computational Auditory Scene Analysis**
   - Data-driven
   - Top-down constraints
3. Recognizing Speech in Noise
4. Using Models in Parallel
5. The Listening Machine

Computational Auditory Scene Analysis: The Representational Approach

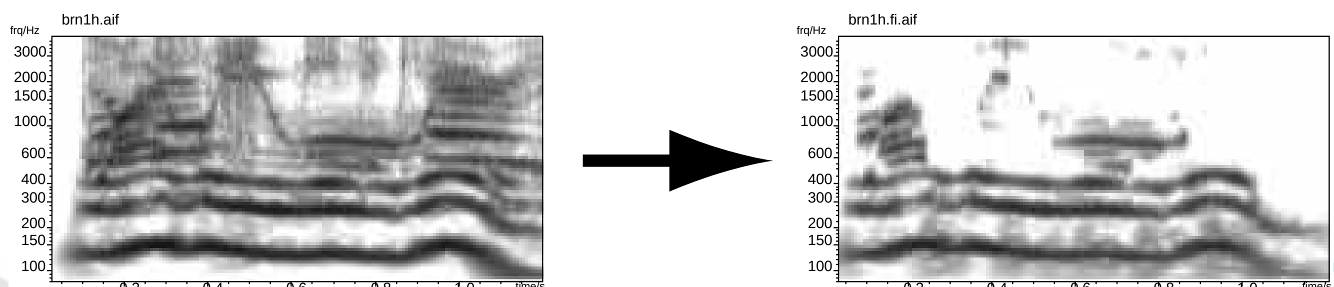
(Cooke & Brown 1993)

- **Direct implementation of psych. theory**



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

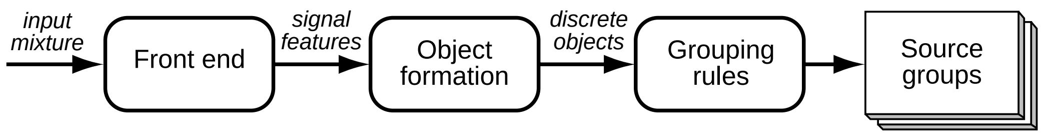
- **Able to extract voiced speech:**



Adding top-down constraints

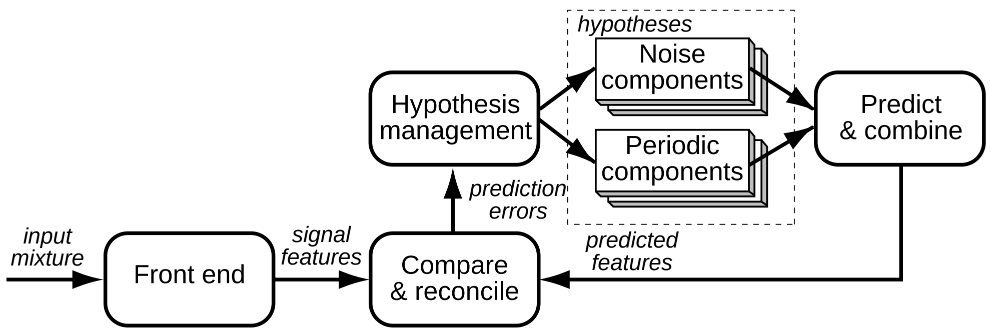
Perception is not direct
but a search for plausible hypotheses

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



  - objects irresistibly appear

vs. Prediction-driven (top-down)

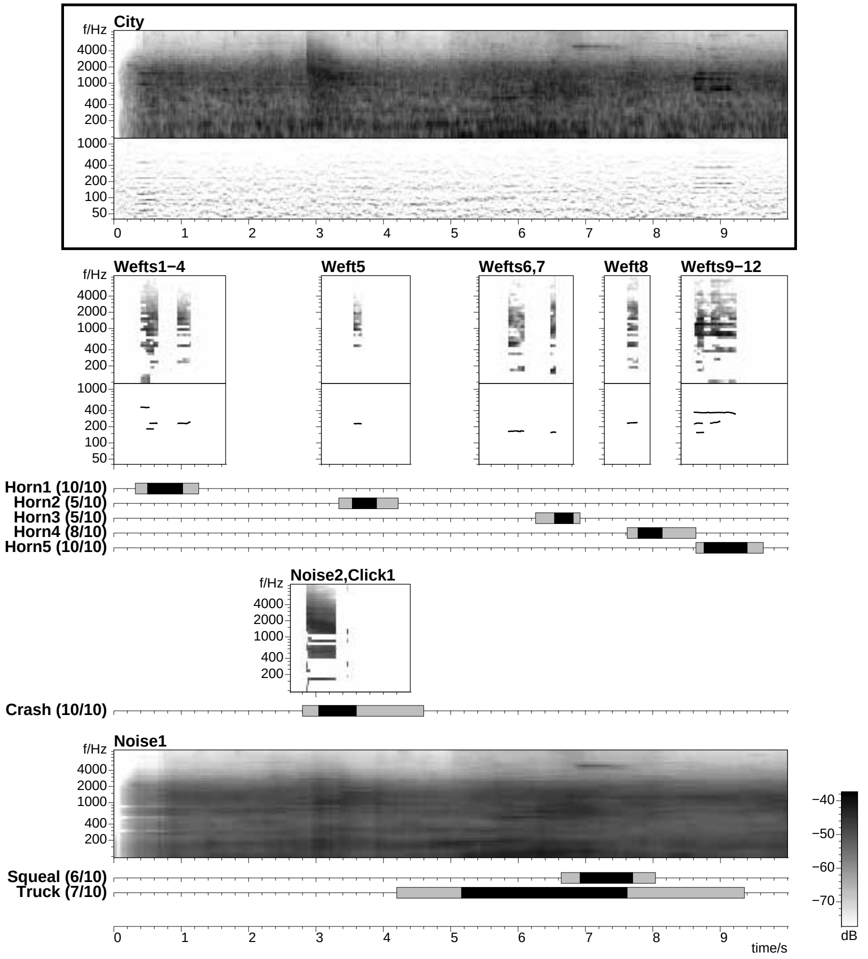


  - match observations
    with parameters of a world-model
  - need world-model constraints...

Prediction-Driven CASA

(Ellis 1996)

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



Aside: Evaluation

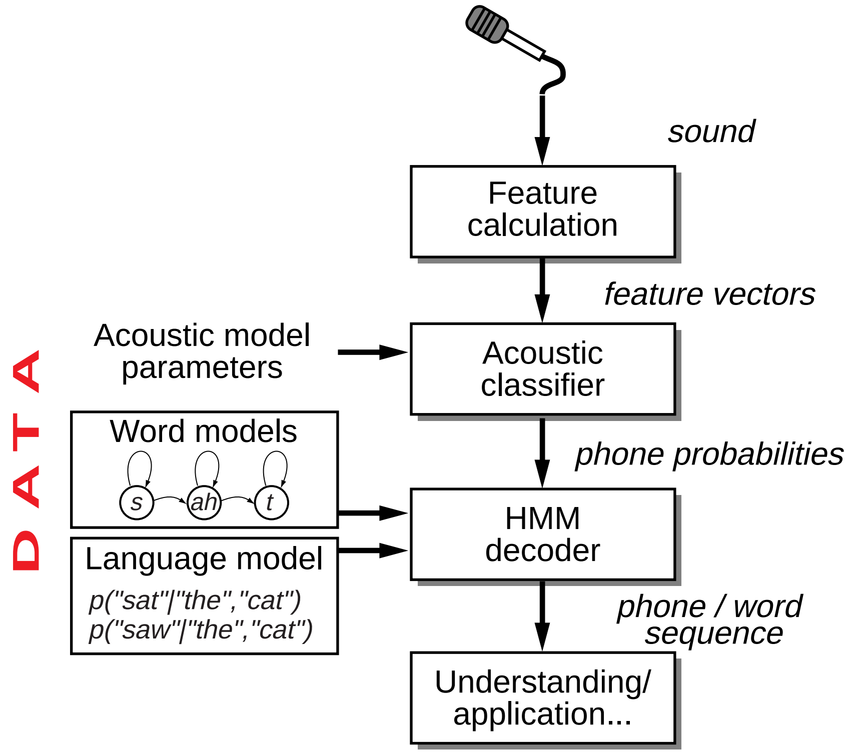
- **Evaluation is a big problem for CASA**
  - what is the goal, really?
  - what is a good test domain?
  - how do you measure performance?
- **SNR improvement**
  - tricky to derive from before/after signals: correspondence problem
  - can do with fixed filtering mask; but rewards removing signal as well as noise
- **Speech Recognition (ASR) improvement**
  - recognizers typically very sensitive to artefacts
- **'Real' task?**
  - mixture corpus with specific sound events...

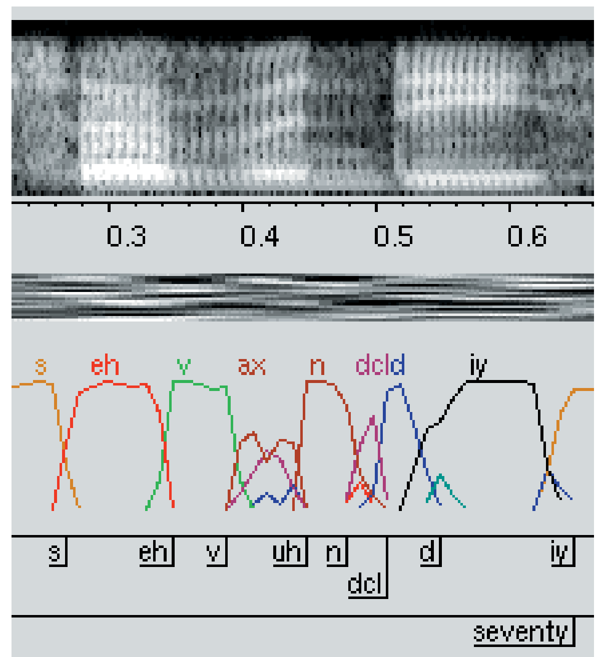
Outline

1. Sound, Mixtures & Learning
2. Computational Auditory Scene Analysis
3. **Recognizing Speech in Noise**
   - Conventional ASR
   - Tandem modeling
4. Using Models in Parallel
5. The Listening Machine

3 Recognizing Speech in Noise

- **Standard speech recognition structure:**





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

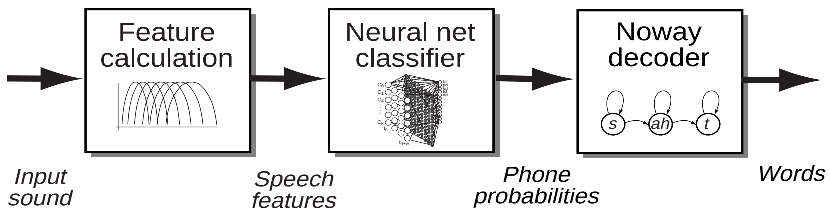
Tandem speech recognition

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

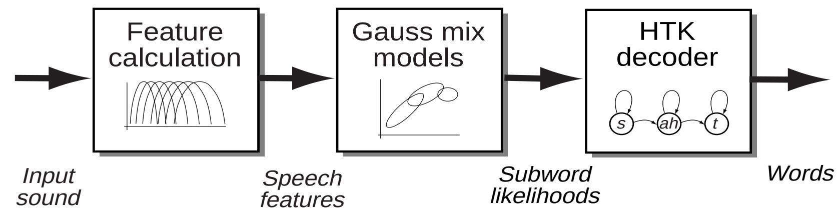
- **Neural net estimates phone posteriors; but Gaussian mixtures model finer detail**

- **Combine them!**

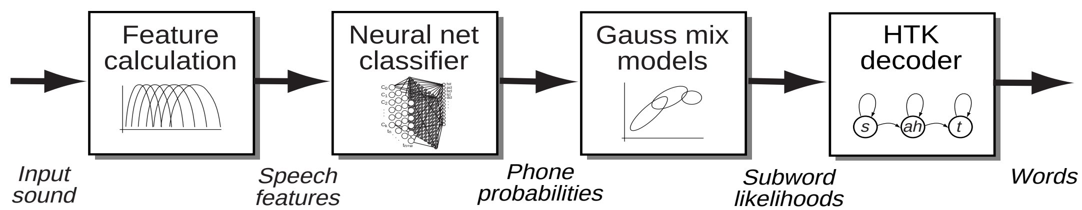
Hybrid Connectionist-HMM ASR



Conventional ASR (HTK)



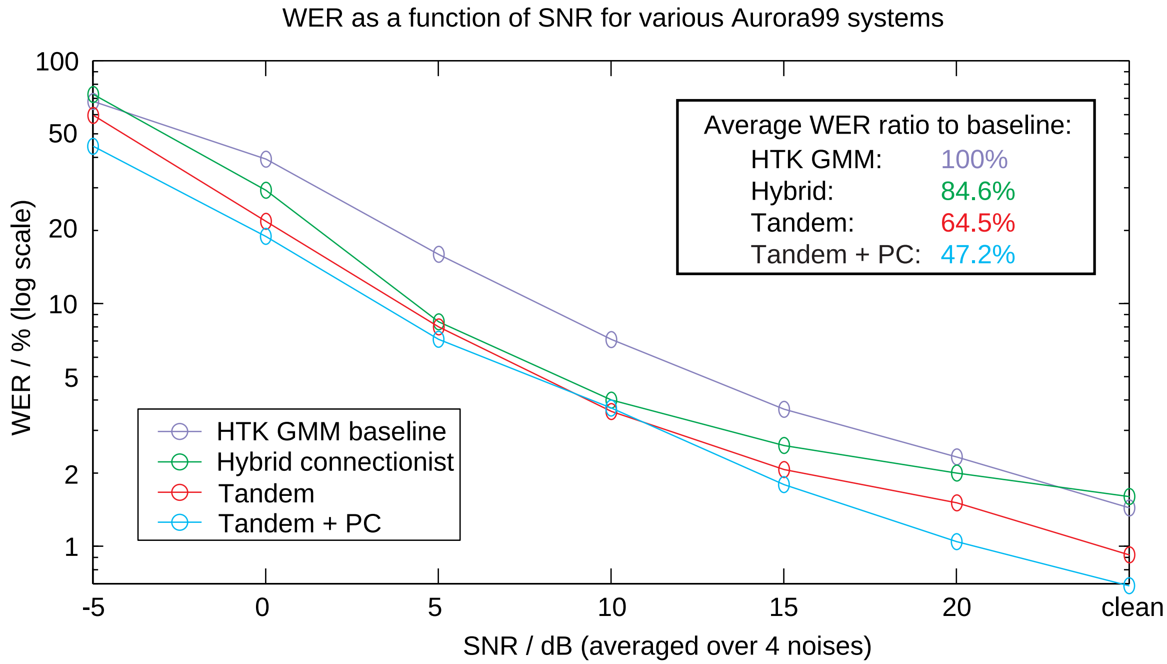
Tandem modeling



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

Tandem system results

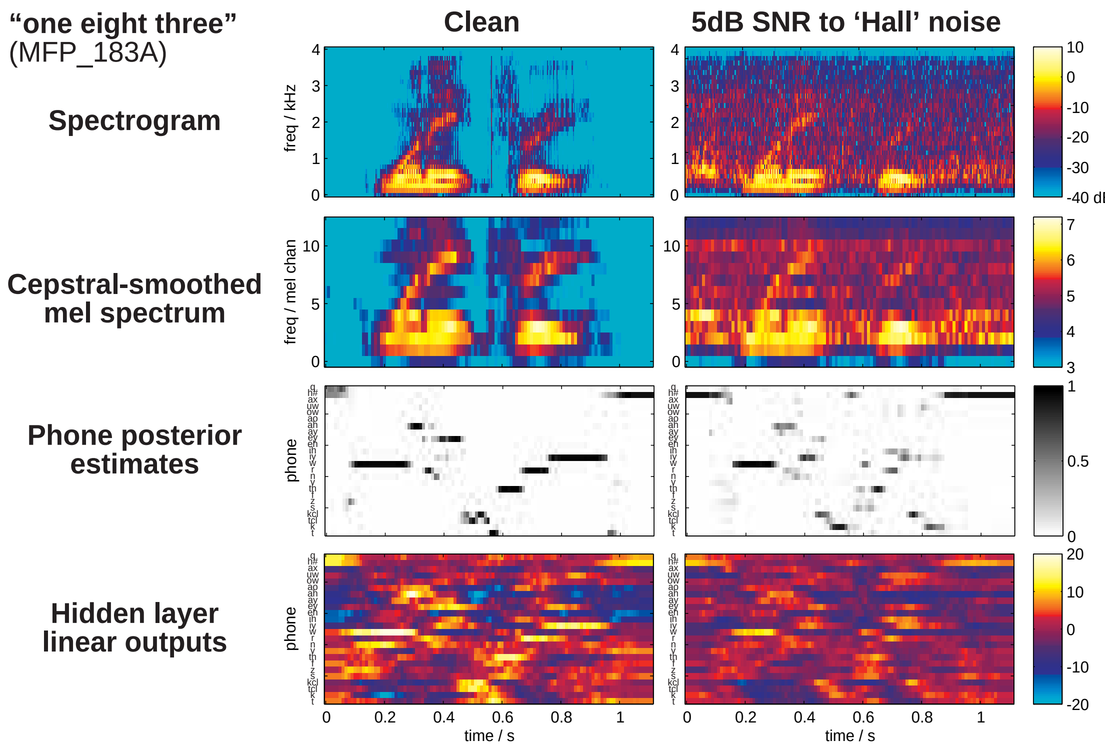
- **It works very well ('Aurora' noisy digits):**



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%**

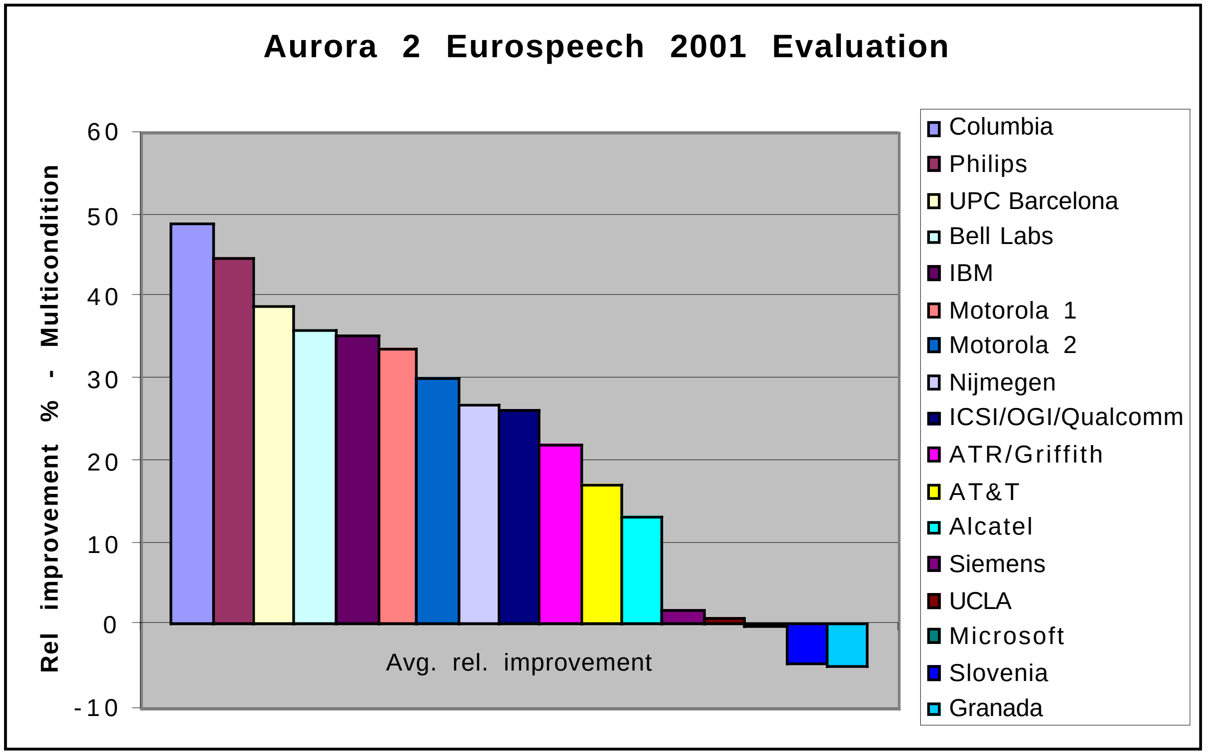
Inside Tandem systems: What's going on?

- **Visualizations of the net outputs**



- **Neural net normalizes away noise?**
  - ... just a successful way to build a classifier?

Tandem vs. other approaches



- **50% of word errors corrected over baseline**
- **Beat a ‘bells and whistles’ system that used many large-vocabulary techniques**

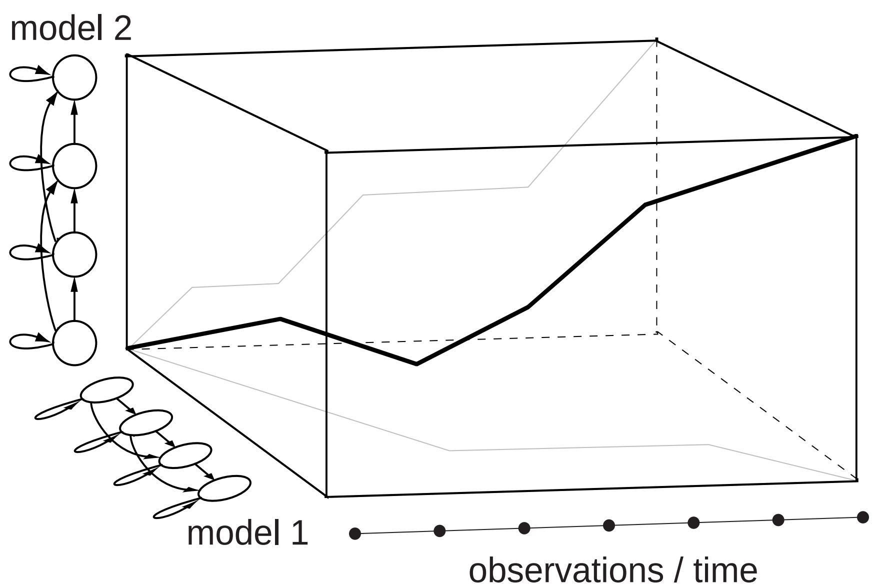
Outline

1. Sound, Mixtures & Learning
2. Computational Auditory Scene Analysis
3. Recognizing Speech in Noise
4. **Using Models in Parallel**
   - HMM decomposition/factoring
   - Speech fragment decoding
5. The Listening Machine

Using Models in Parallel: HMM decomposition

(e.g. Varga & Moore 1991, Gales & Young 1996)

- **Independent state sequences for 2+ component source models**

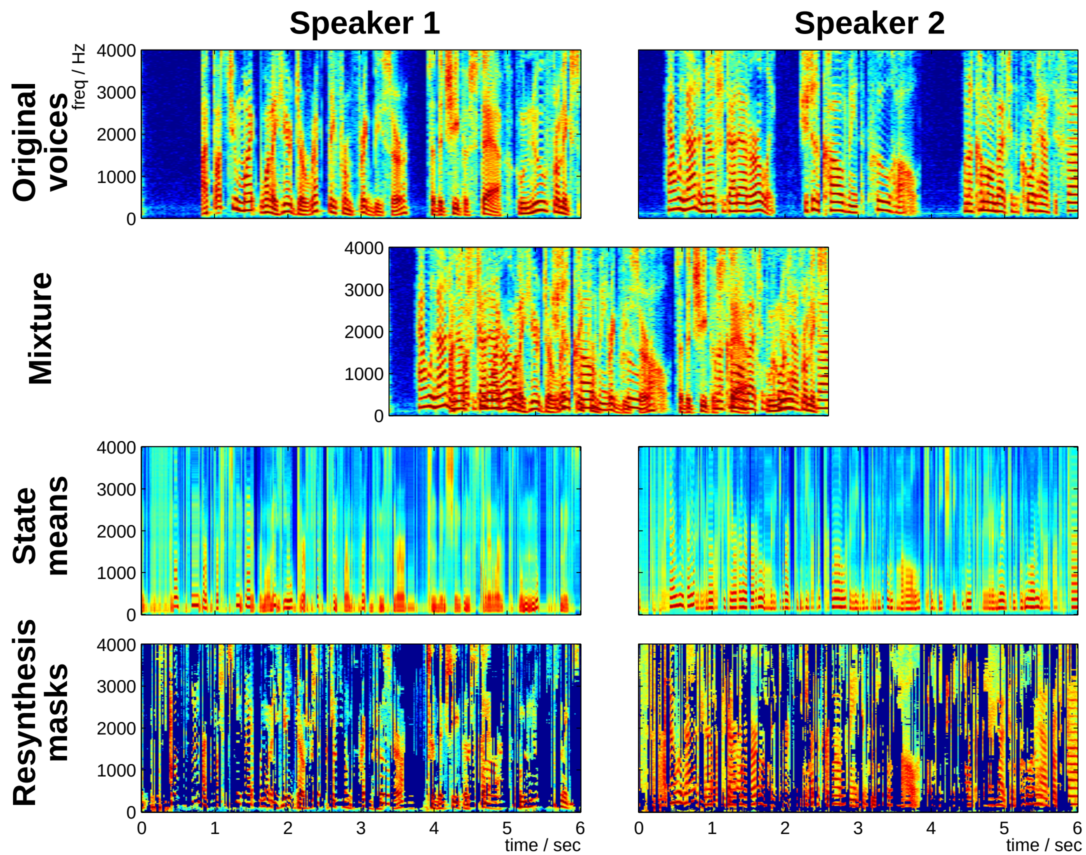


- **New combined state space** $q' = \{q_1\ q_2\}$
  - need pdfs for each combination $p(X|q_1, q_2)$

"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)?

Speech Fragment Recognition

(Jon Barker & Martin Cooke, Sheffield)

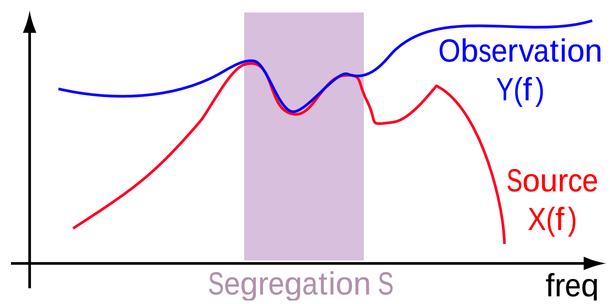
- **Signal separation is too hard!**
  Instead:
  - segregate features into partially-observed sources
  - then classify
- **Made possible by missing data recognition**
  - integrate over uncertainty in observations for optimal posterior distribution
- **Goal:**
  Relate clean speech models $P(X|M)$
  to speech-plus-noise mixture observations
  - .. and make it tractable

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:**

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

  - integral collapses in several cases...

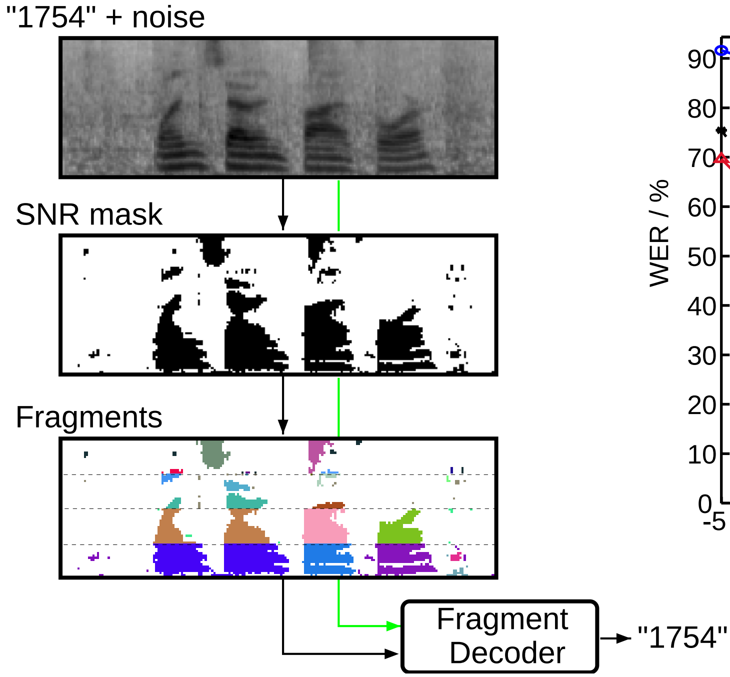
Calculating fragment matches

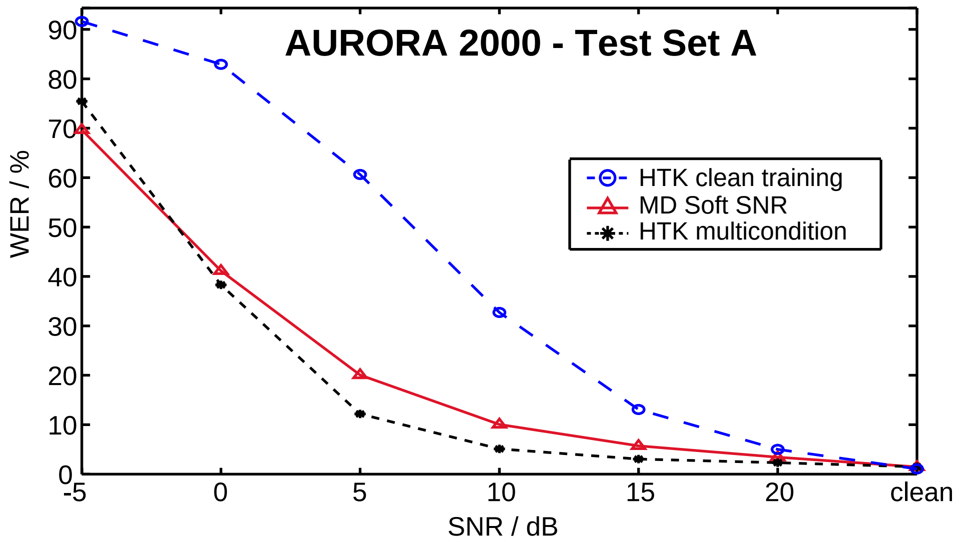
$$P(M, S|Y) = P(M)\int P(X|M) \cdot \frac{P(X|Y,S)}{P(X)} dX \cdot P(S|Y)$$

- $P(X|M)$ **- the clean-signal feature model**
- $P(X|Y,S)/P(X)$ **- is X 'visible' given segregation?**
- **Integration collapses some bands...**
- $P(S|Y)$ **- segregation inferred from observation**
  - just assume uniform, find S for most likely M
  - or: use extra information in Y to distinguish S's e.g. harmonicity, onset grouping
- **Result:**
  - probabilistically-correct relation between clean-source models $P(X|M)$ and inferred, recognized source + segregation $P(M,S|Y)$

Speech fragment decoder results

- **Simple $P(S|Y)$ model forces contiguous regions to stay together**
  - big efficiency gain when searching S space

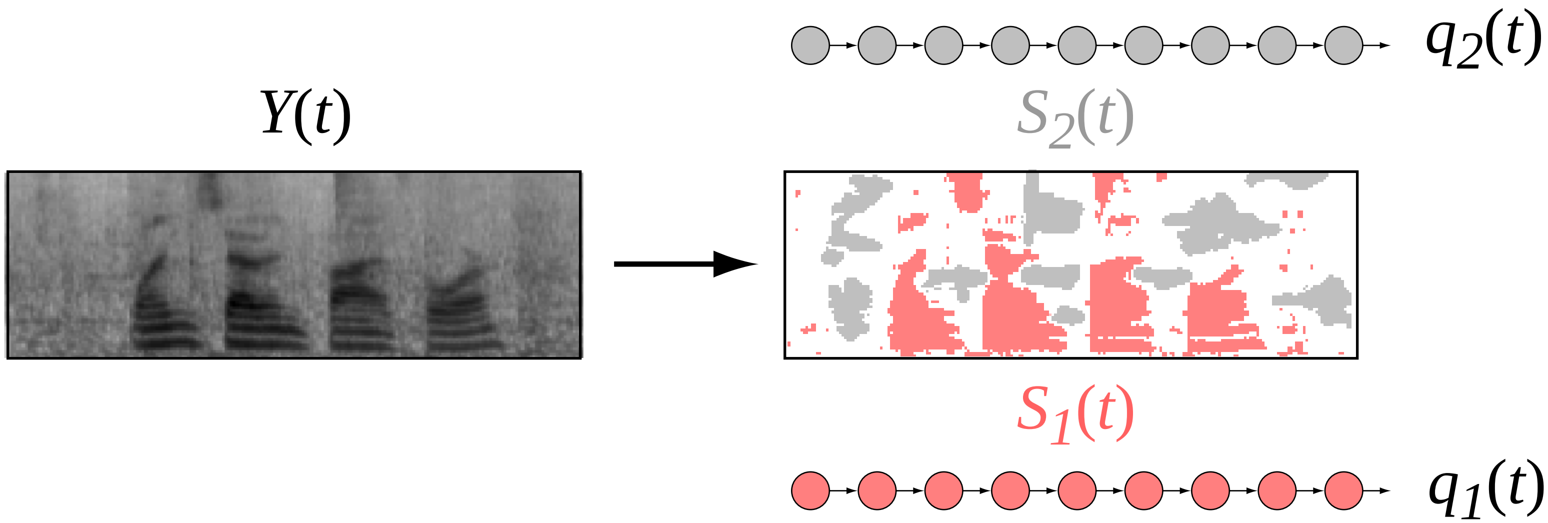




- **Clean-models-based recognition rivals trained-in-noise recognition**

Multi-source decoding

- **Search for more than one source**



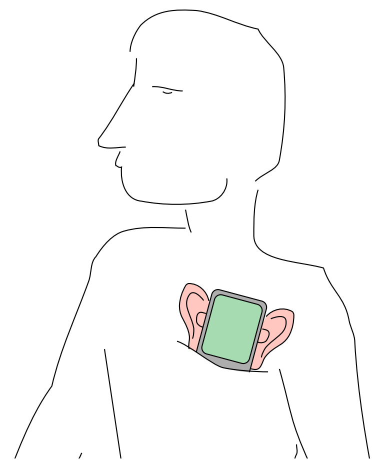
- **Mutually-dependent data masks**
- **Use e.g. CASA features to propose masks**
  - locally coherent regions
  - more powerful than Roweis masks
- **Huge practical advantage over full search**

Outline

1. Sound, Mixtures & Learning
2. Computational Auditory Scene Analysis
3. Recognizing Speech in Noise
4. Using Models in Parallel
5. **The Listening Machine**
   - Everyday sound
   - Alarms
   - Music

5 The Listening Machine

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

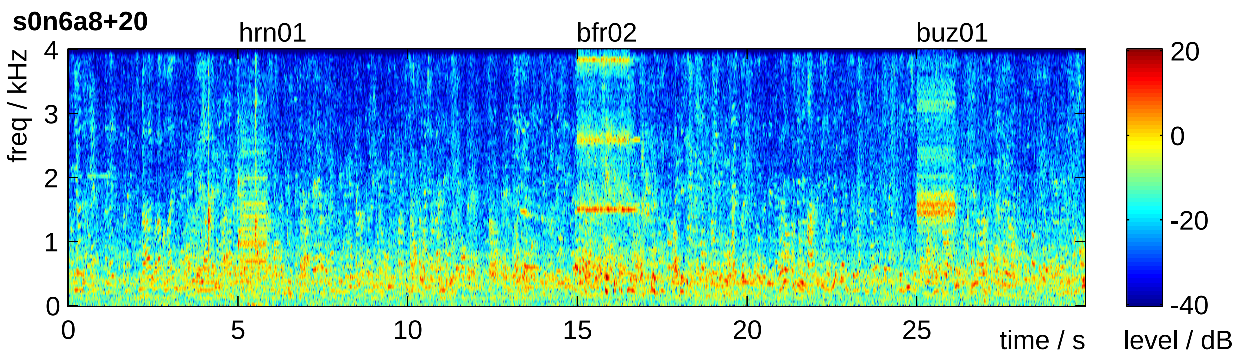


  - personal listener → summary of your day
  - future prosthetic hearing device
  - autonomous robots
- **Meeting data, ambulatory audio**

Alarm sound detection

(Ellis 2001)

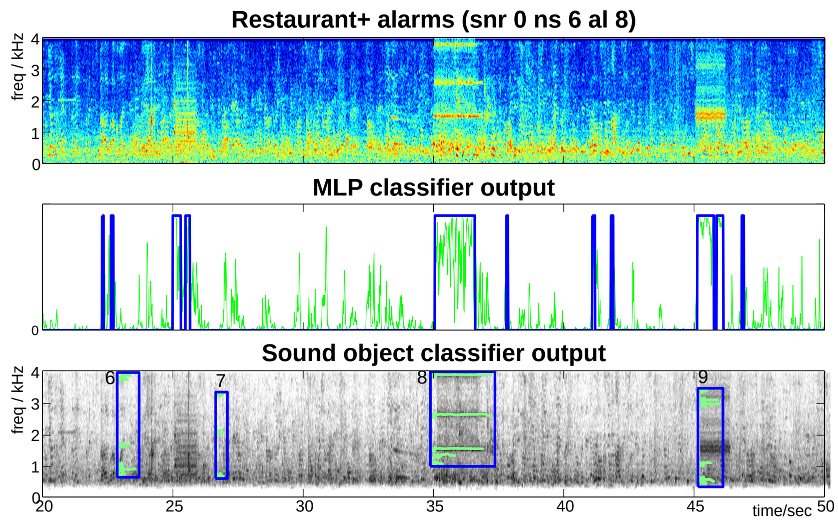
- **Alarm sounds have particular structure**
  - people 'know them when they hear them'
  - clear even at low SNRs



- **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

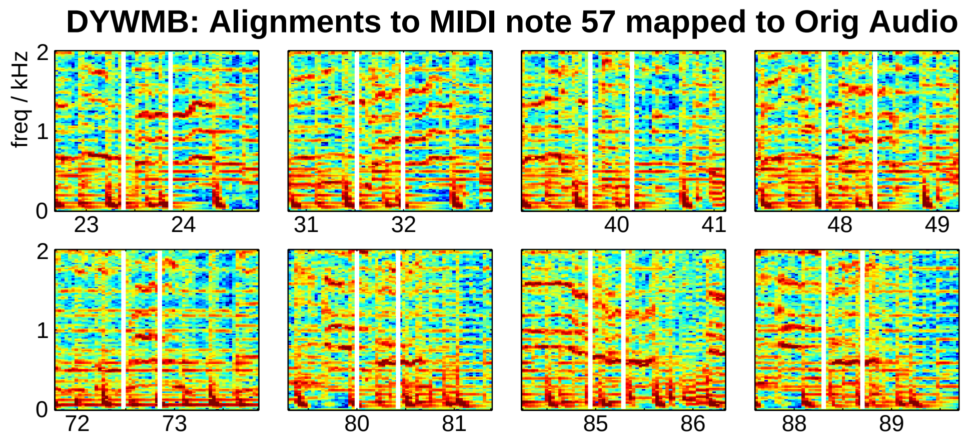


- **Both systems commit many insertions at 0dB SNR, but in different circumstances:**

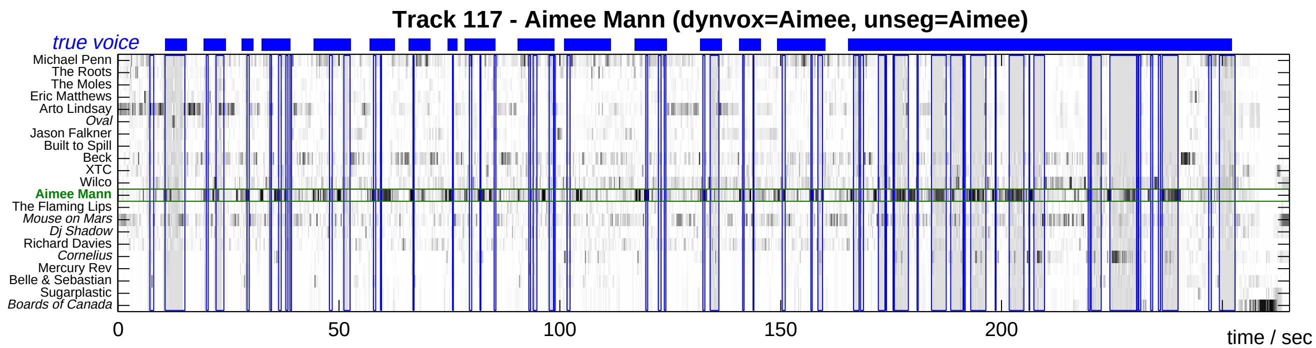
Noise	*Neural net system* *Del*	*Ins*	*Tot*	*Sinusoid model system* *Del*	*Ins*	*Tot*
1 (amb)	7 / 25	2	36%	14 / 25	1	60%
2 (bab)	5 / 25	63	272%	15 / 25	2	68%
3 (spe)	2 / 25	68	280%	12 / 25	9	84%
4 (mus)	8 / 25	37	180%	9 / 25	135	576%
Overall	**22 / 100**	170	**192%**	**50 / 100**	147	**197%**

Music Applications

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



  - singing detection (→ genre identification ...)



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

References

A. Bregman. *Auditory Scene Analysis*, MIT Press, 1990.

A. Bell and T. Sejnowski. “An information-maximization approach to blind separation and blind deconvolution,” *Neural Computation*, 7: 1129-1159, 1995.

http://citeseer.nj.nec.com/bell95informationmaximization.html

A. Berenzweig, D. Ellis, S. Lawrence (2002). “Using Voice Segments to Improve Artist Classification of Music “, Proc. AES-22 Intl. Conf. on Virt., Synth., and Ent. Audio. Espoo, Finland, June 2002.

http://www.ee.columbia.edu/~dpwe/pubs/aes02-aclass.pdf

A. Berenzweig, D. Ellis, S. Lawrence (2002). “Anchor Space for Classification and Similarity Measurement of Music“, Proc. ICME-03, Baltimore, July 2003.

http://www.ee.columbia.edu/~dpwe/pubs/icme03-anchor.pdf

M. Cooke and G. Brown. “Computational auditory scene analysis: Exploiting principles of perceived continuity”, *Speech Communication* 13, 391-399, 1993

D. Ellis. *Prediction-driven computational auditory scene analysis*, Ph.D. dissertation, MIT, 1996.

http://www.ee.columbia.edu/~dpwe/pubs/pdcasa.pdf

D. Ellis. “Detecting Alarm Sounds”, Proc. Workshop on Consistent & Reliable Acoustic Cues CRAC-01, Denmark, Sept. 2001.

http://www.ee.columbia.edu/~dpwe/pubs/crac01-alarms.pdf

M. Gales and S. Young. “Robust continuous speech recognition using parallel model combination”, *IEEE Tr. Speech and Audio Proc.*, 4(5):352--359, Sept. 1996.

http://citeseer.nj.nec.com/gales96robust.html

H. Hermansky,D. Ellis and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems,” *Proc. ICASSP*, Istanbul, June 2000.

http://citeseer.nj.nec.com/hermansky00tandem.html