Sound, Mixtures, and Learning

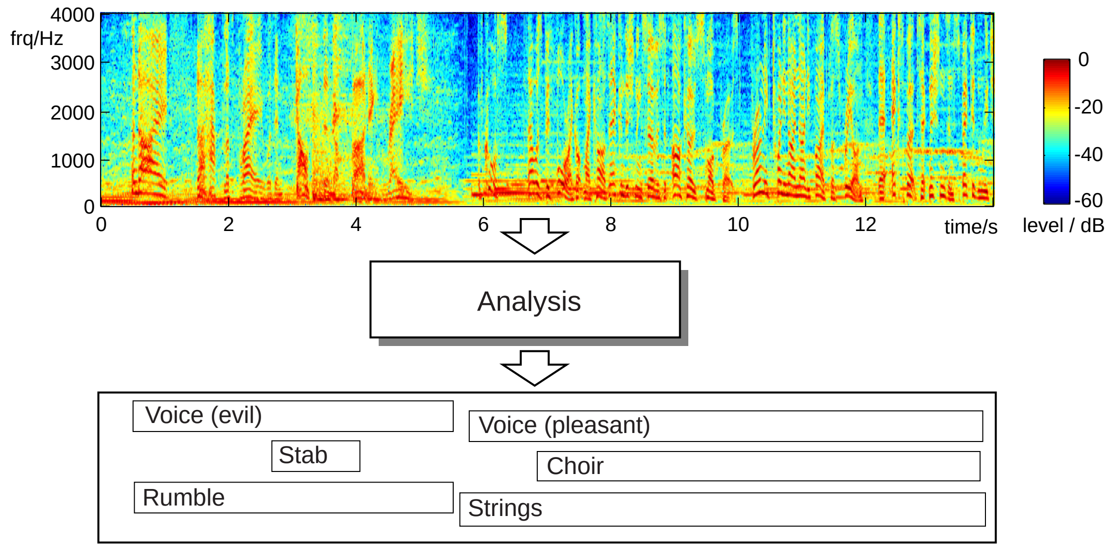
Dan Ellis
<dpwe@ee.columbia.edu>

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

Outline

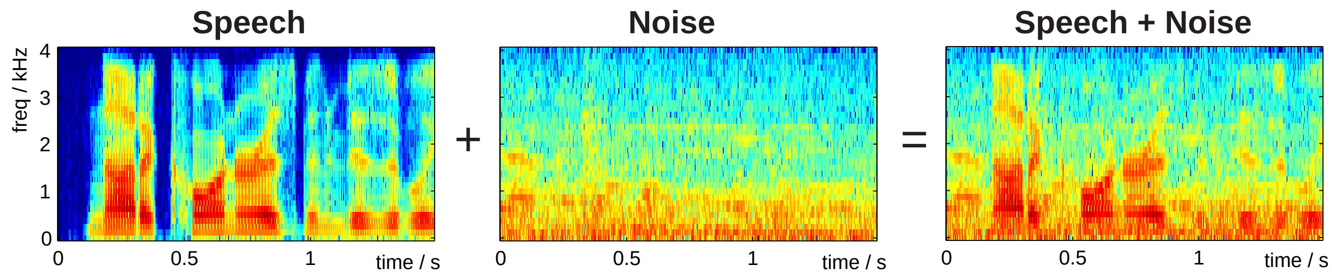
1. **Auditory Scene Analysis**
2. **Speech Recognition & Mixtures**
3. **Fragment Recognition**
4. **Alarm Sound Detection**
5. **Future Work**

1 Auditory Scene Analysis



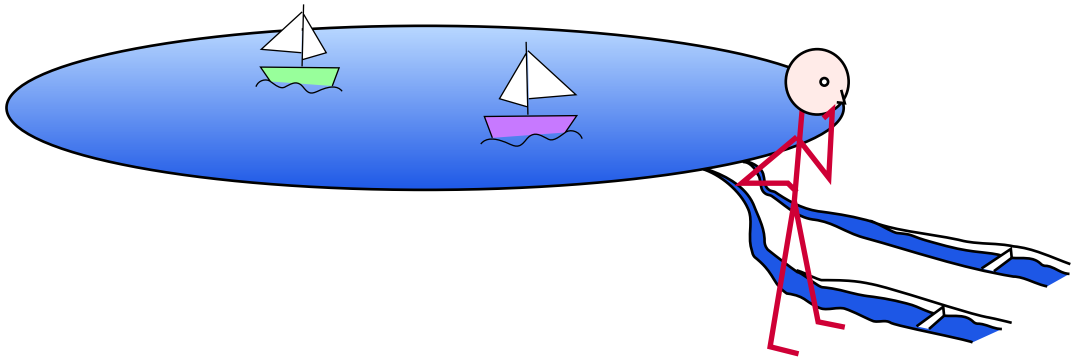
- ***Auditory Scene Analysis*: describing a complex sound in terms of high-level sources/events**
  - ... like listeners do
- **Hearing is *ecologically* grounded**
  - reflects ‘natural scene’ properties
  - subjective, not absolute

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

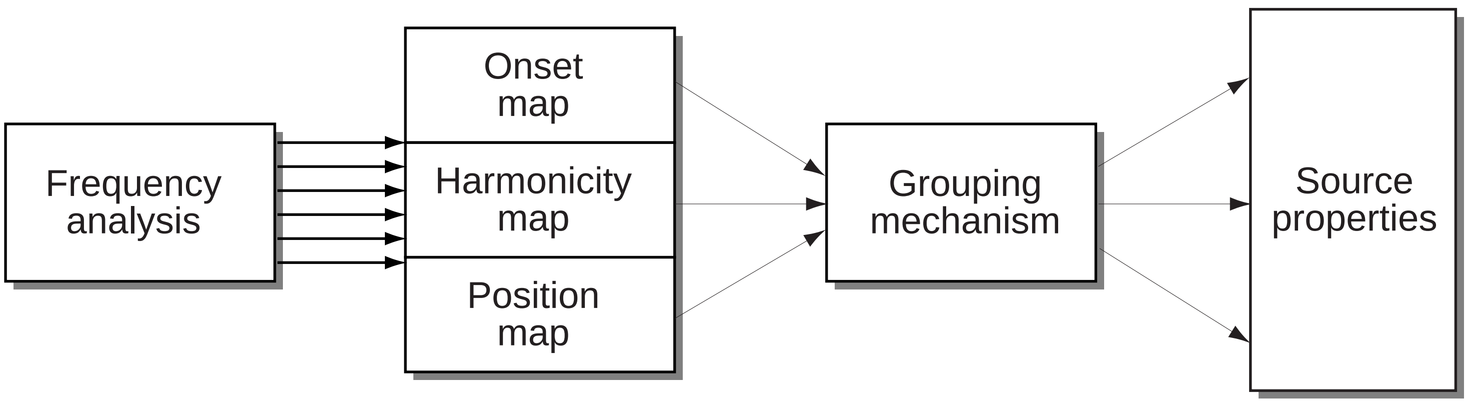


"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- **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)

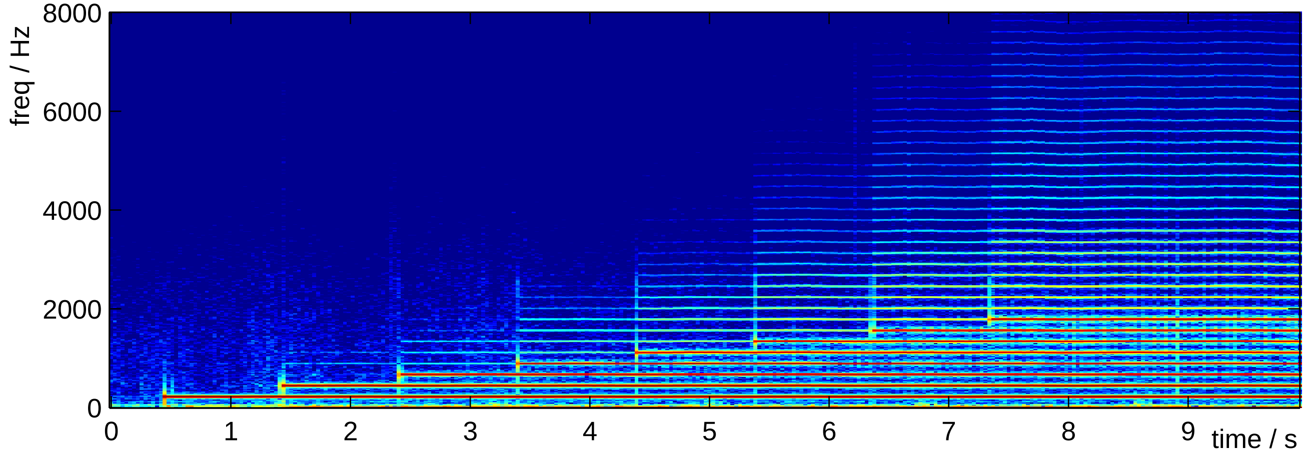
- **How do people analyze sound mixtures?**
  - break mixture into small *elements* (in time-freq)
  - elements are *grouped* in to sources using *cues*
  - sources have aggregate *attributes*
- **Grouping 'rules' (Darwin, Carlyon, ...):**
  - cues: common onset/offset/modulation, harmonicity, spatial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



(after Darwin, 1996)

Cues to simultaneous grouping

- **Elements + attributes**

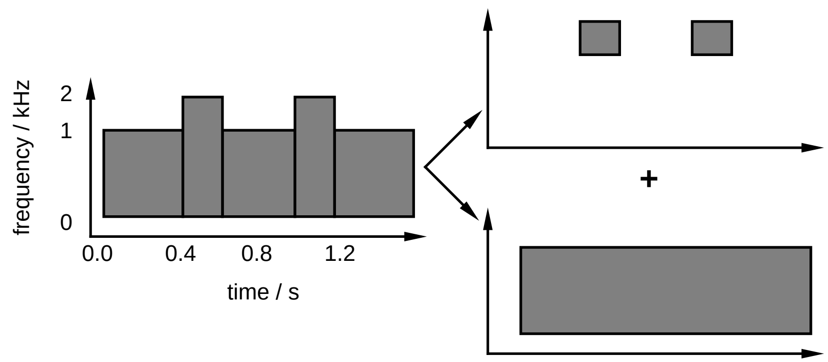


- **Common onset**
  - simultaneous energy has common source
- **Periodicity**
  - energy in different bands with same cycle
- **Other cues**
  - spatial (ITD/IID), familiarity, ...

The effect of context

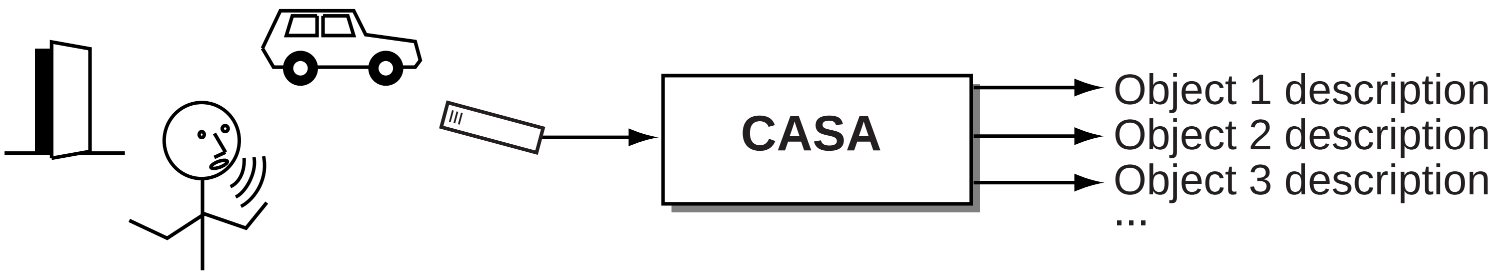
- **Context can create an 'expectation': i.e. a bias towards a particular interpretation**
- **e.g. Bregman's "old-plus-new" principle:**

  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)

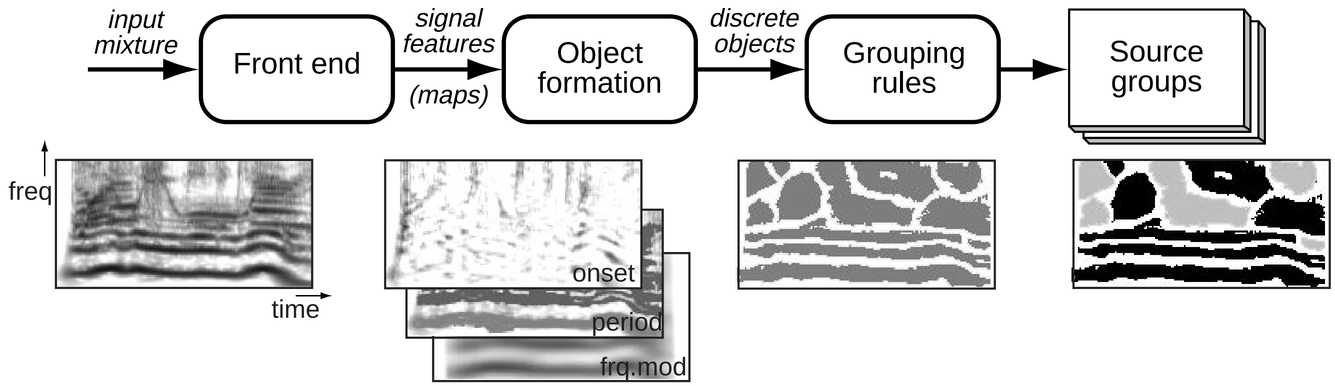


- **Goal: Automatic sound organization ; Systems to 'pick out' sounds in a mixture**
  - ... like people do
- **E.g. voice against a noisy background**
  - to improve speech recognition
- **Approach:**
  - psychoacoustics describes grouping 'rules'
  - ... just implement them?

The Representational Approach

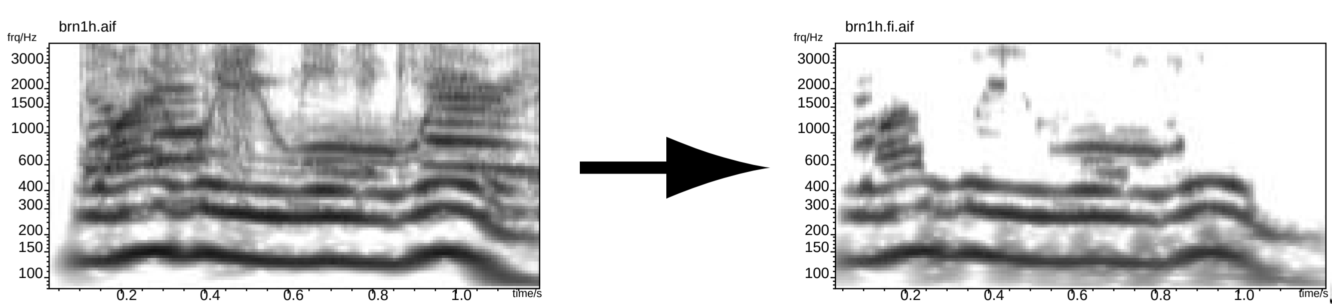
(Brown & Cooke 1993)

- **Implement psychoacoustic theory**



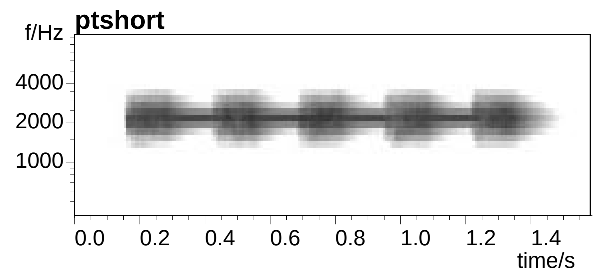
  - ‘bottom-up’ processing
  - uses common onset & periodicity cues

- **Able to extract voiced speech:**

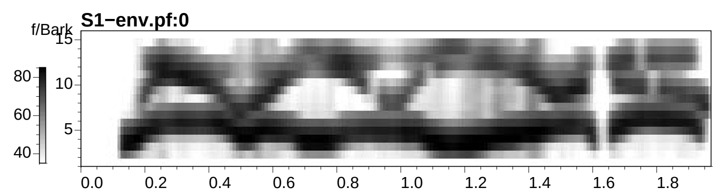


Restoration in sound perception

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



- **SWS**



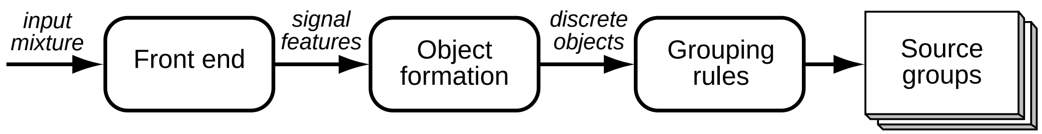
  - duplex perception

- **How to model in CASA?**

Adding top-down constraints

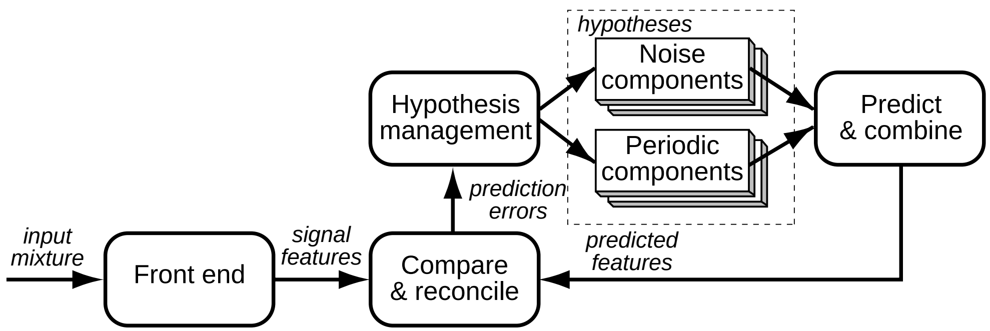
Perception is not *direct*
but a *search* for *plausible hypotheses*

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



  - objects irresistibly appear

vs. Prediction-driven (top-down)



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

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**
- standard ASR
- approaches to speech + noise

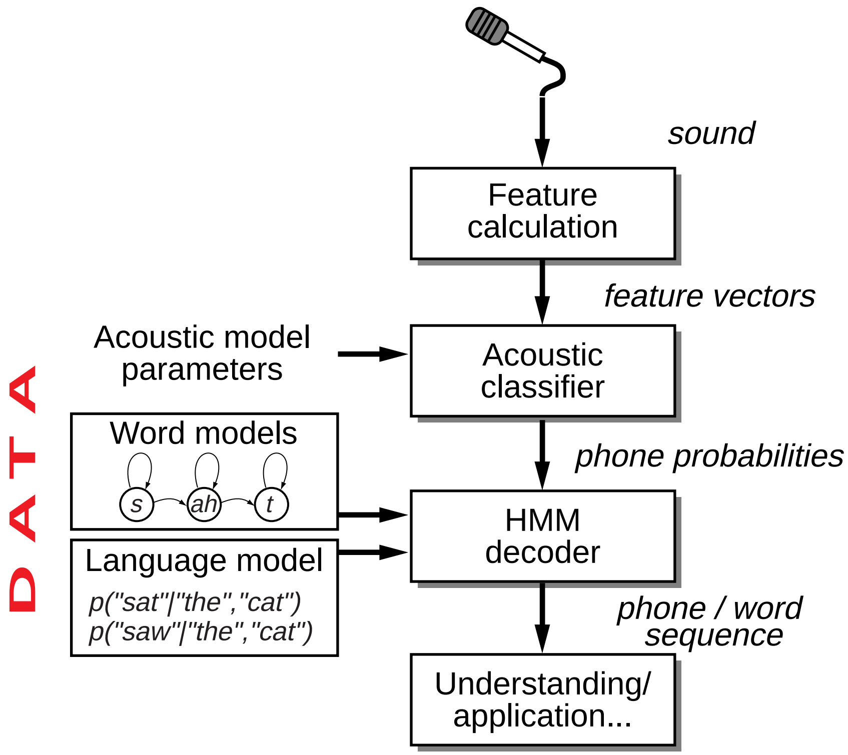
3 Fragment Recognition

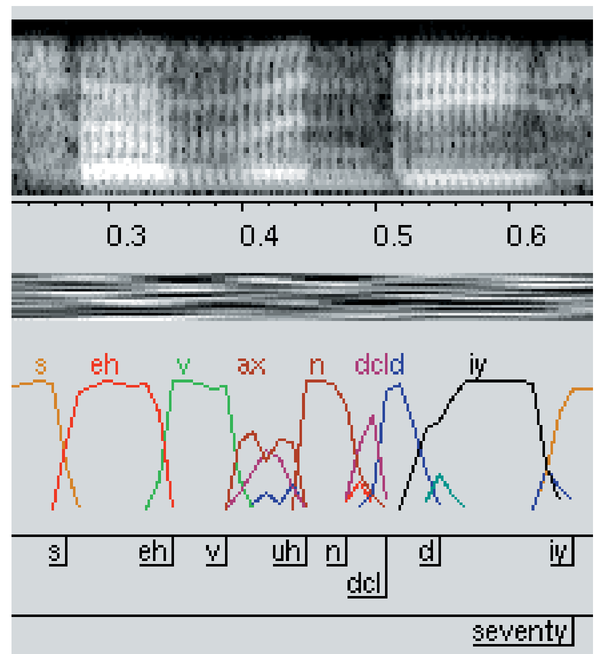
4 Alarm Sound Detection

5 Future Work

2 Speech recognition & mixtures

- **Speech recognizers are the most successful and sophisticated acoustic recognizers to date**

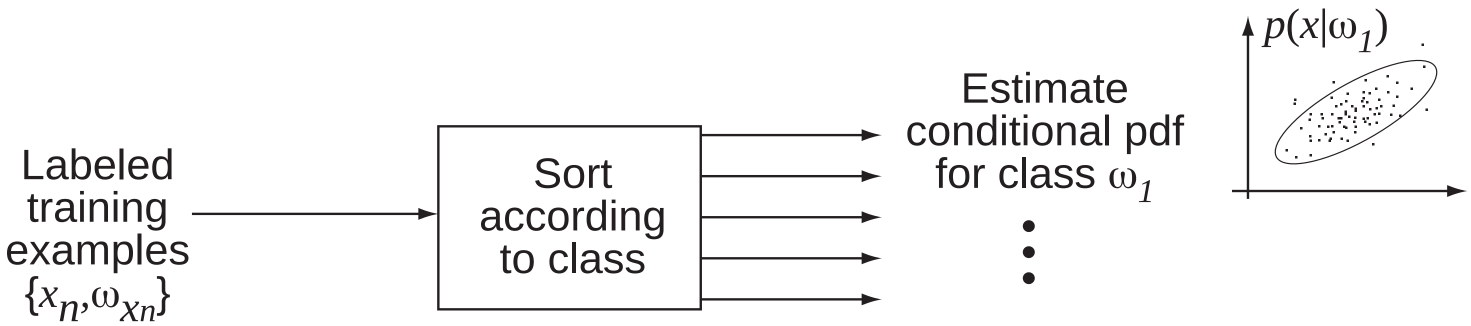




- **'State of the art' word-error rates (WERs):**
  - 2% (dictation) - 30% (phone conv'ns)

Learning acoustic models

- **Goal: describe $p(X|M)$ with e.g. GMMs**



- **Separate models for each class**
  - generalization as blurring
- **Training data labels from:**
  - manual annotation
  - 'best path' from earlier classifier (Viterbi)
  - EM: joint estimation of labels & pdfs

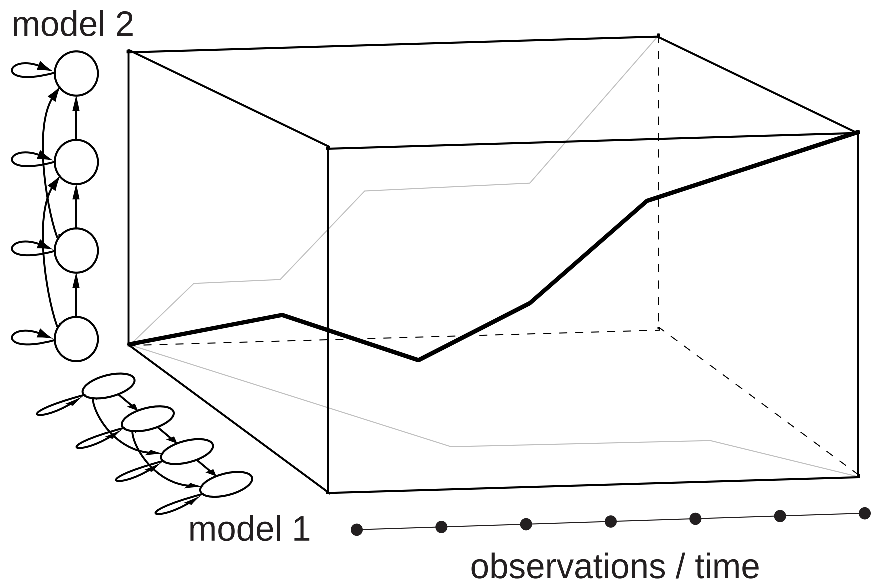
Speech + noise mixture recognition

- **Background noise is biggest (?) problem facing current ASR**
- **Feature invariance approach: Design features to reflect only speech**
  - e.g. normalization, mean subtraction
- **Ideally, models of clean speech will match speech in noise**
  - .. although training on noisy examples can't hurt
- **Static noise is relatively easy**
  - but: non-static noise?
- **Alternative: More complex models of the signal**
  - separate models for speech and 'rest'

HMM decomposition

(e.g. Varga & Moore 1991, Roweis 2000)

- **Total signal model has independent state sequences for 2+ component sources**



- **New combined state space** $q' = \{q_1\ q_2\}$
  - new observation pdfs for each combination

$$p(X|q_1, q_2)$$

Problems with HMM decomposition

- $O(q_k)^N$ **is exponentially large...**
- **Feature *normalization* no longer holds!**
  - each source has a different gain
    $\rightarrow$ model at various SNRs?
  - models typically don't use overall energy C_0
  - each source has a different *channel* $H[k]$
- **Modeling every possible sub-state combination is inefficient, inelegant and impractical**

Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. **Fragment Recognition**
   - separating signals vs. separating features
   - missing data recognition
   - recognizing multiple sources
4. Alarm Sound Detection
5. Future Work

3

Fragment Recognition

(Jon Barker & Martin Cooke, Sheffield)

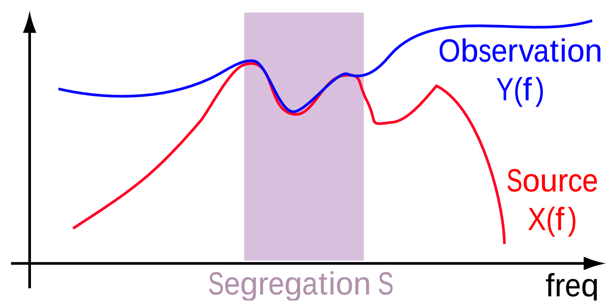
- ***Signal* separation is too hard!**
  Instead:
  - segregate *features* into partially-observed sources
  - then classify
- **Made possible by 'missing data' recognition**
  - integrate over uncertainty in observations for optimal posterior distribution
- **Goal:**
  Relating clean speech models $P(X|M)$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

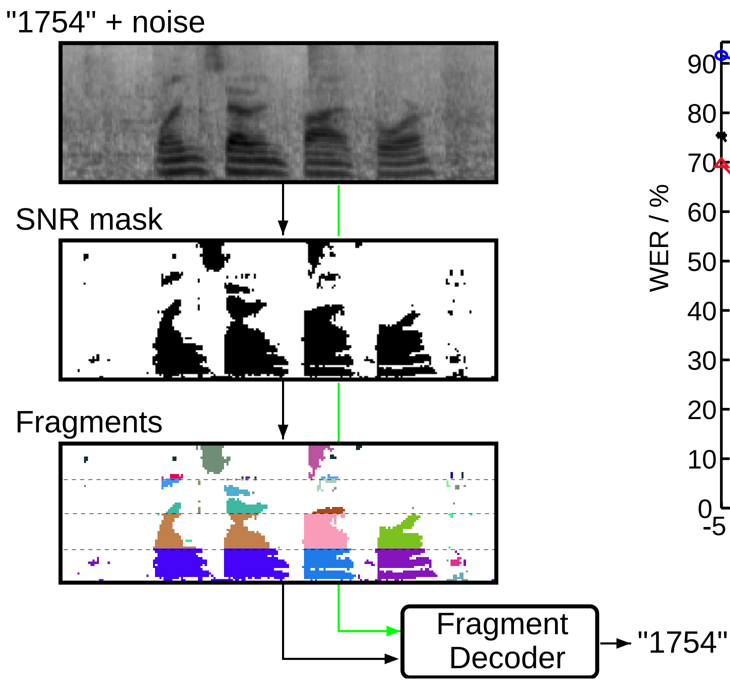
Calculating fragment matches

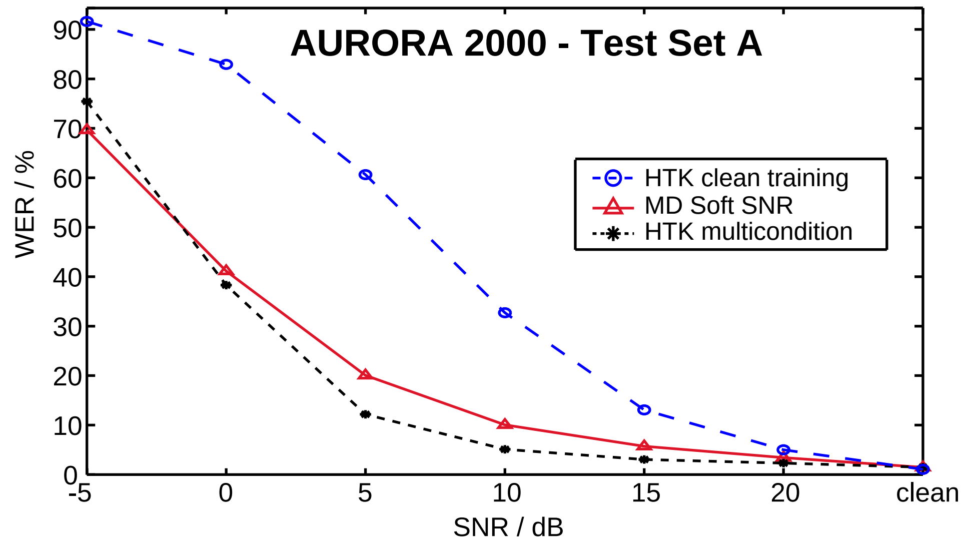
$$P(M,S|Y) = P(M)\int P(X|M)\cdot\frac{P(X|Y,S)}{P(X)}dX\cdot P(S|Y)$$

- $P(X|M)$ **- the clean-signal feature model**
- $P(X|Y,S)/P(X)$ **- is X 'visible' given segregation?**
- **Integration collapses some channels...**
- $P(S|Y)$ **- segregation inferred from observation**
  - just assume uniform, find S for most likely M
  - 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 contributory source $P(M,S|Y)$

Speech fragment decoder results

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

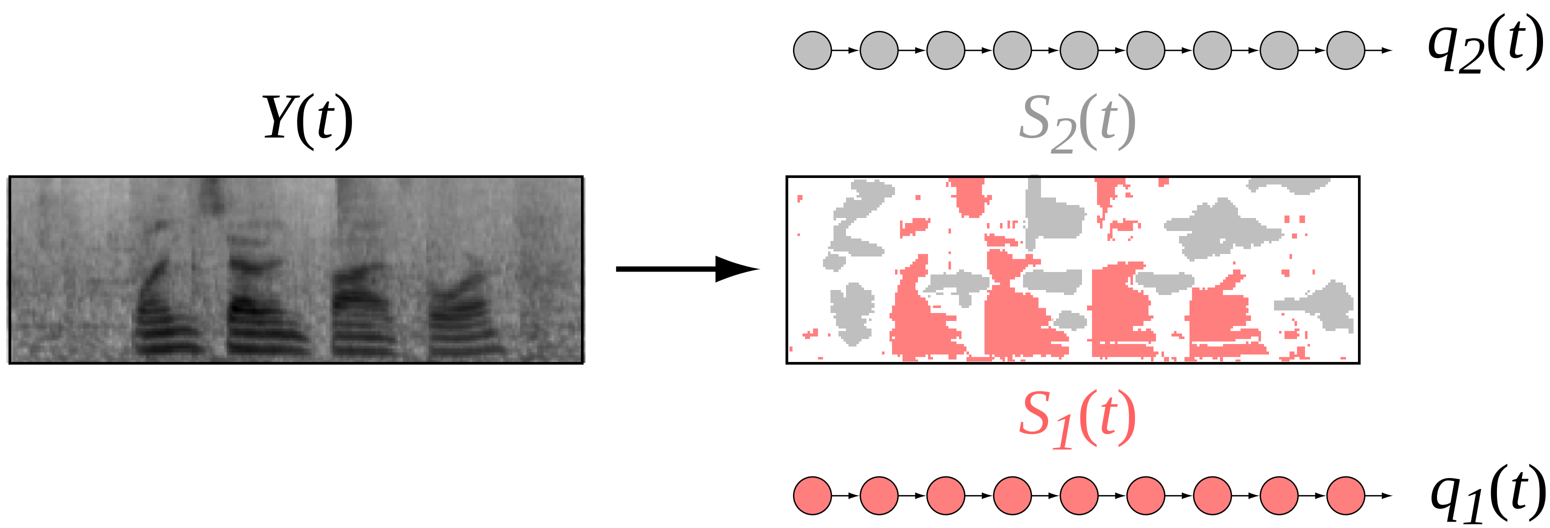




- **Clean-models-based recognition rivals trained-in-noise recognition**

Multi-source decoding

- **Search for more than one source**



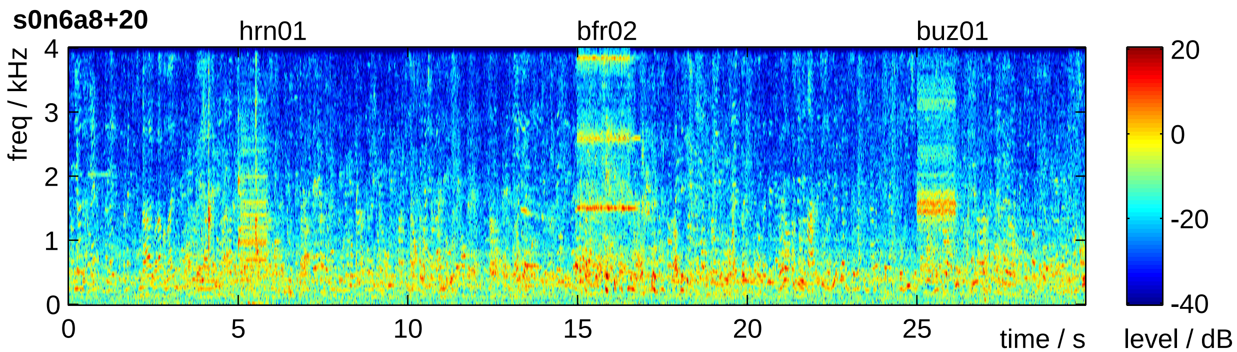
- **Mutually-dependent data masks**
- **Use e.g. CASA features to propose masks**
  - locally coherent regions
- **Theoretical vs. practical limits**

Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. Fragment Recognition
4. **Alarm Sound Detection**
   - sound
   - mixtures
   - learning
5. Future Work

4 Alarm sound detection

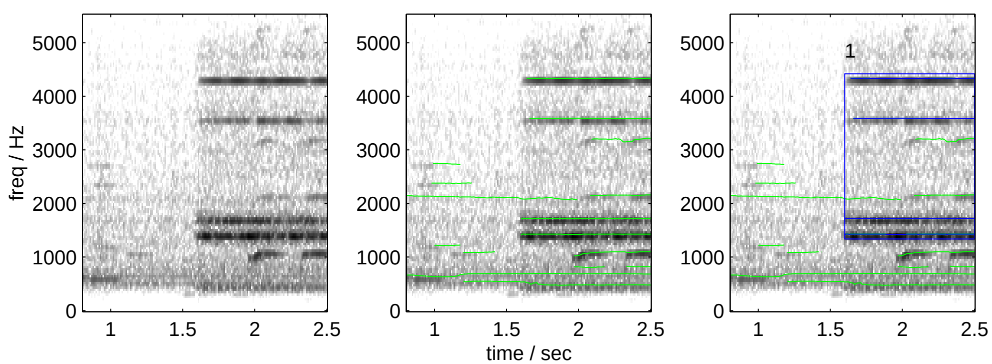
- **Alarm sounds have particular structure**
  - people ‘know them when they hear them’
  - clear even at low SNRs



- **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: Sound (representation)

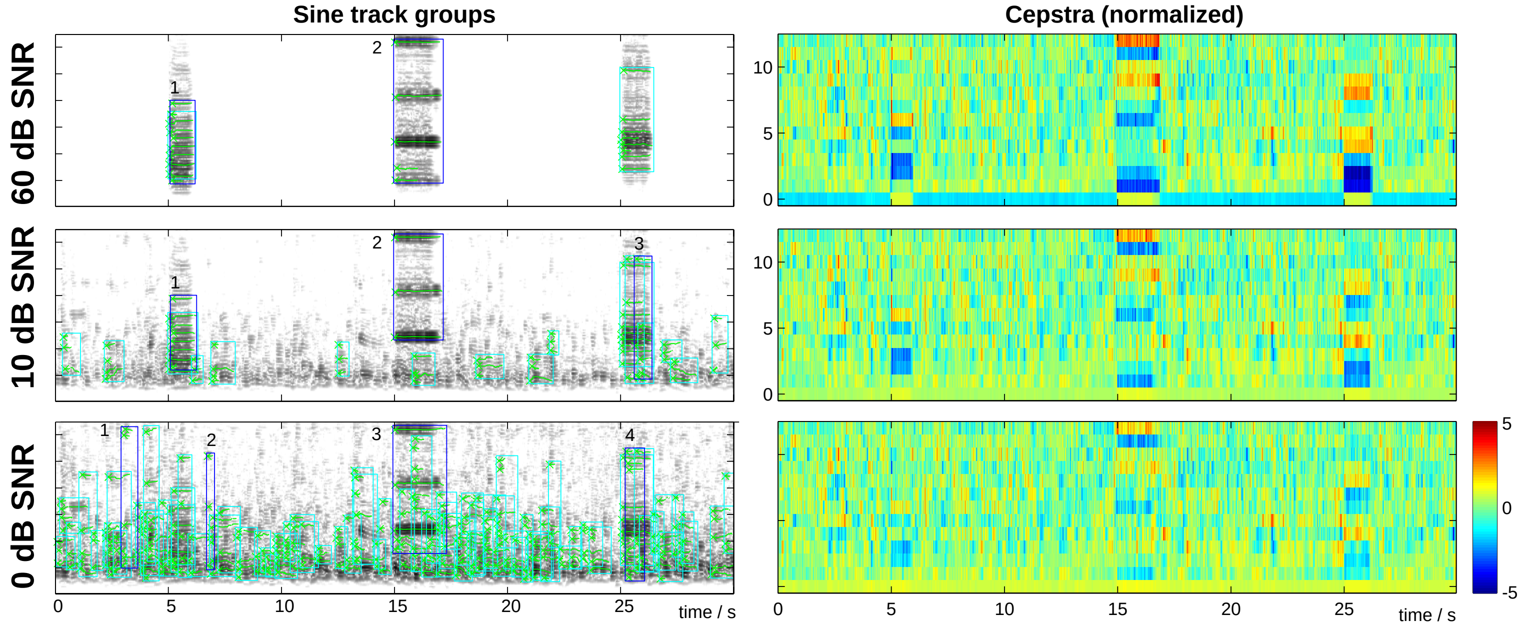
- **Standard system: Mel Cepstra**
  - have to model alarms in noise context: each cepstral element depends on whole signal
- **Contrast system: Sinusoid groups**
  - exploit sparse, stable nature of alarm sounds
  - 2D-filter spectrogram to enhance harmonics
  - simple magnitude threshold, track growing
  - form groups based on common onset



- **Sinusoid representation is already *fragmentary***
  - does not record non-peak energies

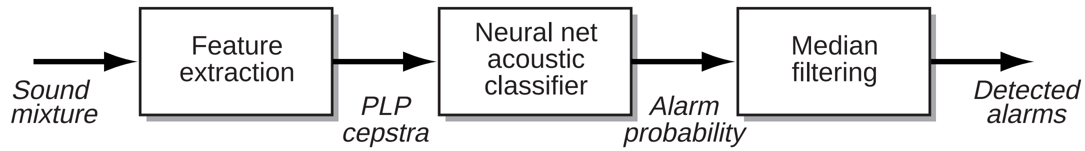
Alarms: Mixtures

- **Effect of varying SNR on representations:**
  - sinusoid peaks have ~ invariant properties

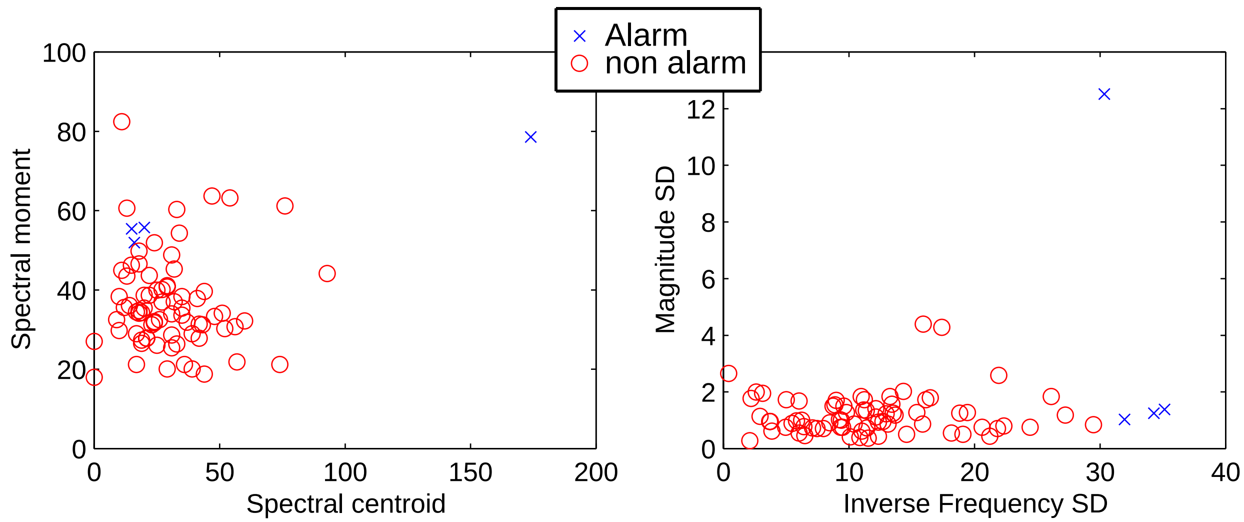


Alarms: Learning

- **Standard: train MLP on noisy examples**

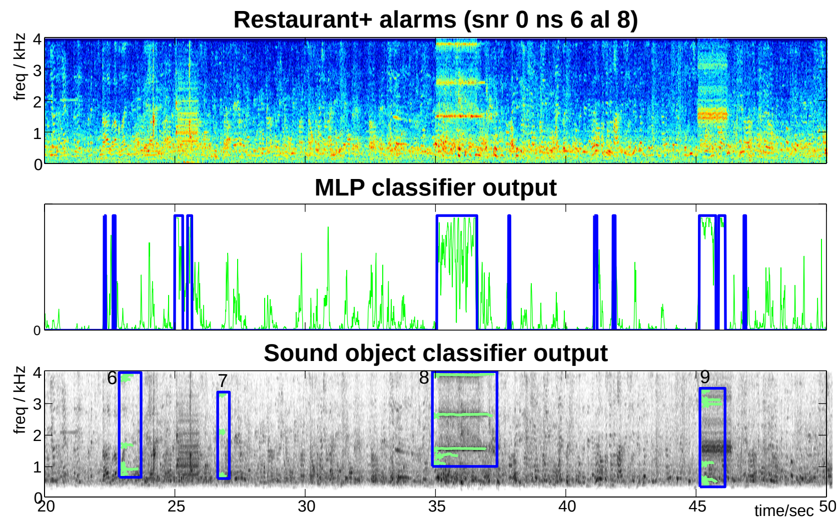


- **Alternate: learn distributions of group features**
  - duration, frequency deviation, amp. modulation...



  - underlying models are clean (isolated)
  - recognize in different contexts...

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%**

Alarms: Summary

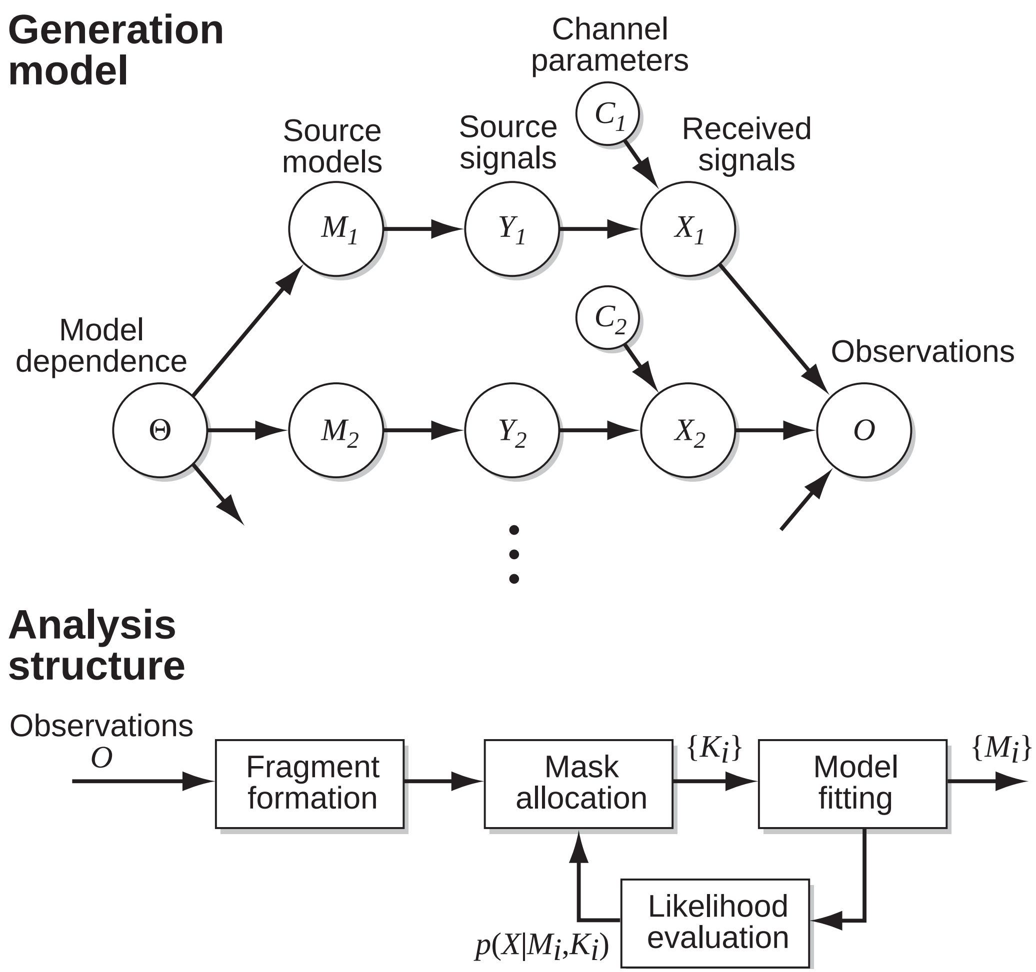
- **Sinusoid domain**
  - feature components belong to 1 source
  - simple 'segregation' (grouping) model
  - alarm model as properties of group
  - robust to partial feature observation
- **Future improvements**
  - more complex alarm class models
  - exploit repetitive structure of alarms

Outline

1. Auditory Scene Analysis
2. Speech Recognition & Mixtures
3. Fragment Recognition
4. Alarm Sound Detection
5. **Future Work**
    - generative models & inference
    - model acquisition
    - ambulatory audio

5 Future work

- **CASA as generative model parameterization:**

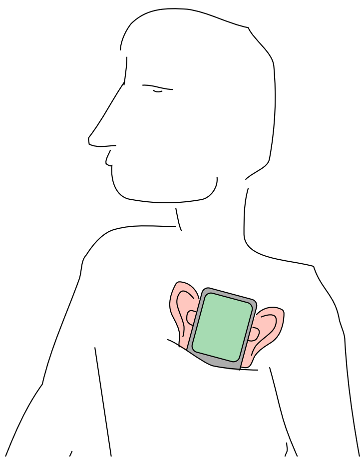


Learning source models

- **The speech recognition lesson: Use the data as much as possible**
  - what can we do with unlimited data feeds?
- **Data sources**
  - clean data corpora
  - identify near-clean segments in real sound
  - build up ‘clean’ views from partial observations?
- **Model types**
  - templates
  - parametric/constraint models
  - HMMs
- **Hierarchic classification vs. individual characterization...**

Personal Audio Applications

- **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**

Summary

- **Sound**
  - carries important information
- **Mixtures**
  - need to segregate different source properties
  - fragment-based recognition
- **Learning**
  - information extracted by classification
  - models guide segregation
- **Alarm sounds**
  - simple example of fragment recognition
- **General sounds**
  - recognize simultaneous components
  - acquire classes from training data
  - build index, summary of real-world sound