Sound, Mixtures, and Learning

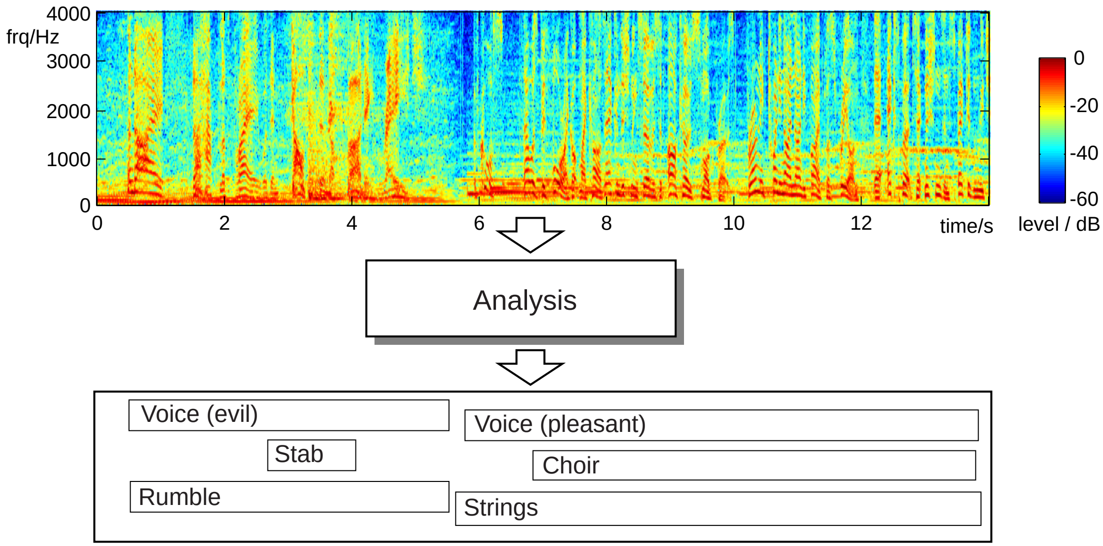
Dan Ellis
<dpwe@ee.columbia.edu>

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

Outline

1. **Human sound organization**
2. **Computational Auditory Scene Analysis**
3. **Speech models and knowledge**
4. **Sound mixture recognition**
5. **Learning opportunities**

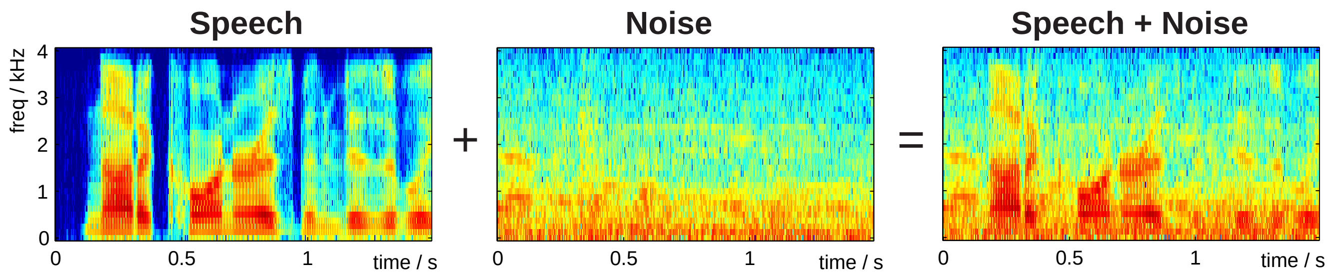
1 Human sound organization



- **Analyzing and describing complex sounds:**
  - continuous sound mixture $\rightarrow$ distinct events
- **Hearing is *ecologically* grounded**
  - reflects ‘natural scene’ properties
  - subjective *not* canonical (ambiguity)
  - *mixture* analysis as primary goal

Sound mixtures

- **Sound 'scene' is almost always a mixture**
  - always stuff going on
  - sound is 'transparent' - but big energy range

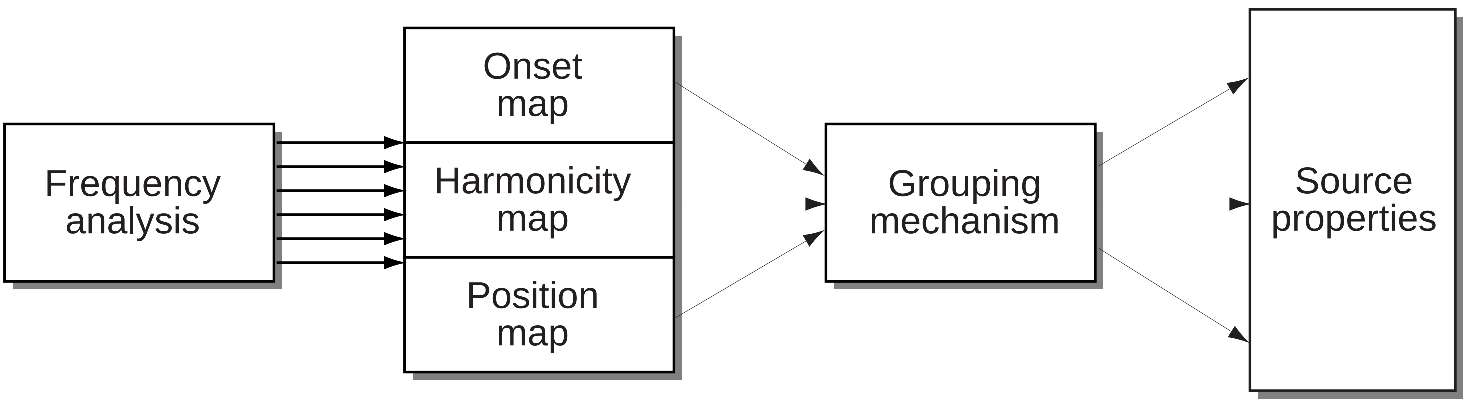


- **Need information related to our 'world model'**
  - i.e. separate objects
  - a wolf howling in a blizzard is the same as a wolf howling in a rainstorm
  - whole-signal statistics won't do this
- **'Separateness' is similar to independence**
  - objects/sounds that change in isolation
  - but: depends on the situation e.g. passing car vs. mechanic's diagnosis

Auditory scene analysis

(Bregman 1990)

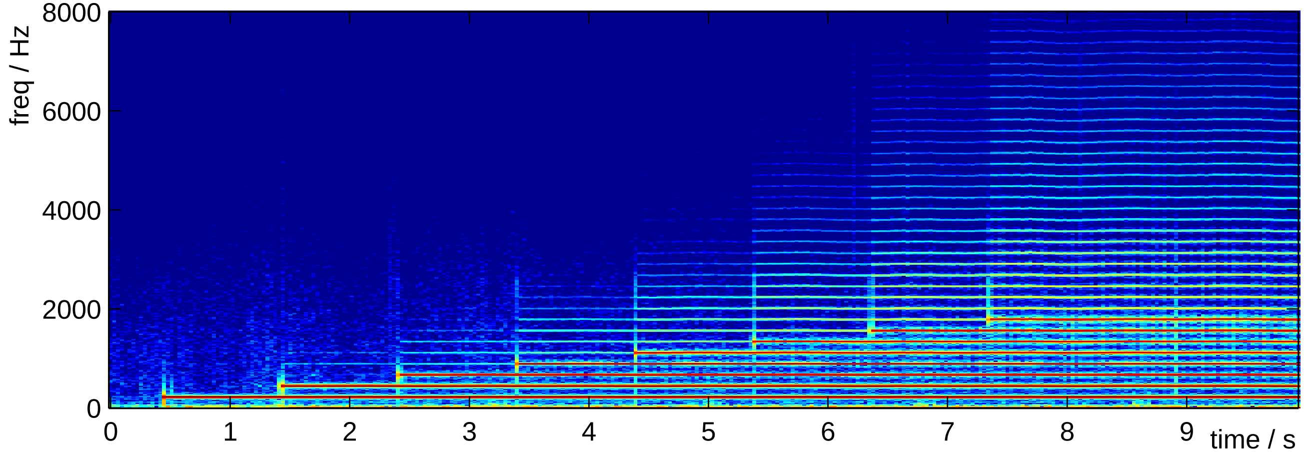
- **How do people analyze sound mixtures?**
  - break mixture into small *elements* (in time-freq)
  - elements are *grouped* in to sources using *cues*
  - sources have aggregate *attributes*
- **Grouping ‘rules’ (Darwin, Carlyon, ...):**
  - cues: common onset/offset/modulation, harmonicity, spatial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



(after Darwin, 1996)

Cues to simultaneous grouping

- **Elements + attributes**

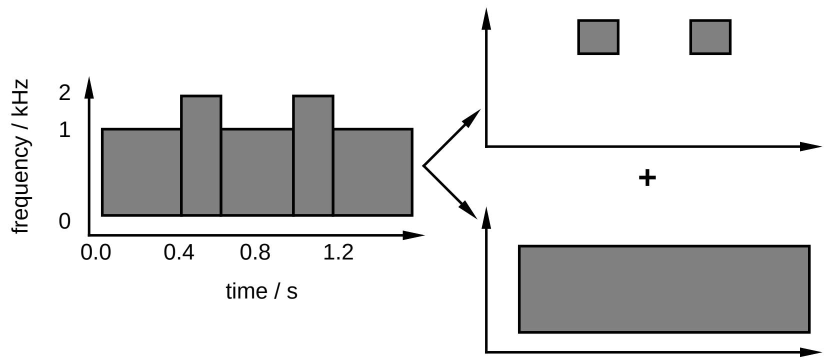


- **Common onset**
  - simultaneous energy has common source
- **Periodicity**
  - energy in different bands with same cycle
- **Other cues**
  - spatial (ITD/IID), familiarity, ...

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

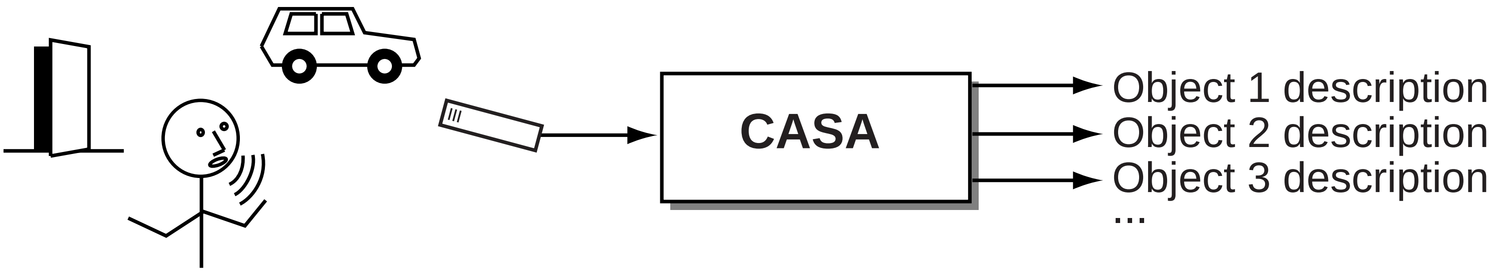


- a different division of the same energy depending on what preceded it

Outline

1. **Human sound organization**
2. **Computational Auditory Scene Analysis**
   - sound source separation
   - bottom-up models
   - top-down constraints
3. **Speech models and knowledge**
4. **Sound mixture recognition**
5. **Learning opportunities**

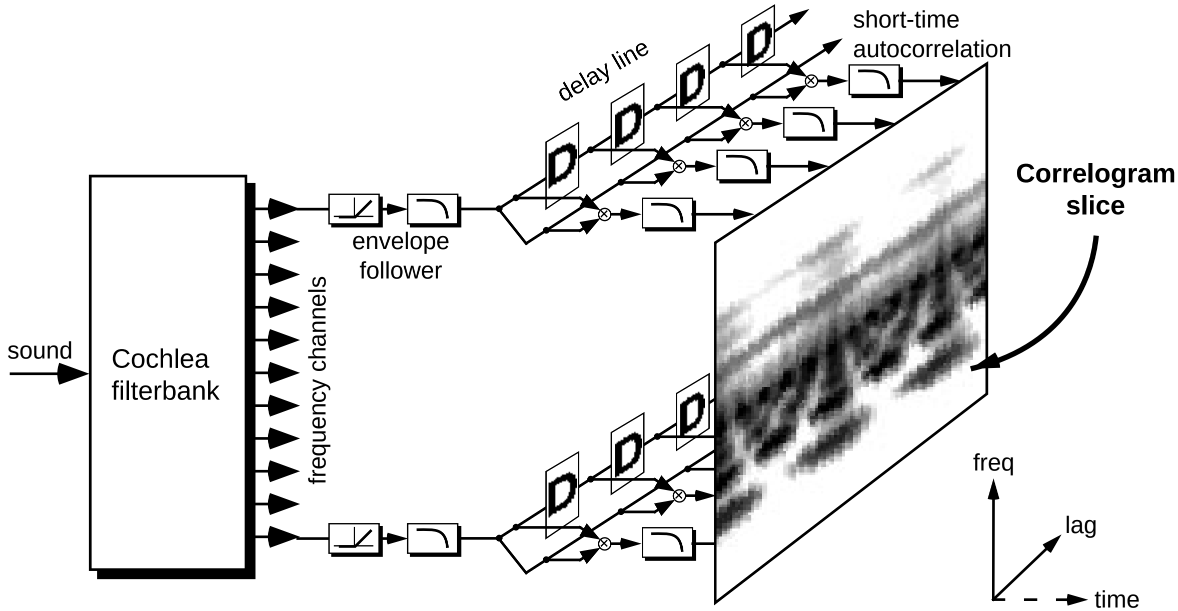
2 Computational Auditory Scene Analysis (CASA)



- **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?

CASA front-end processing

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

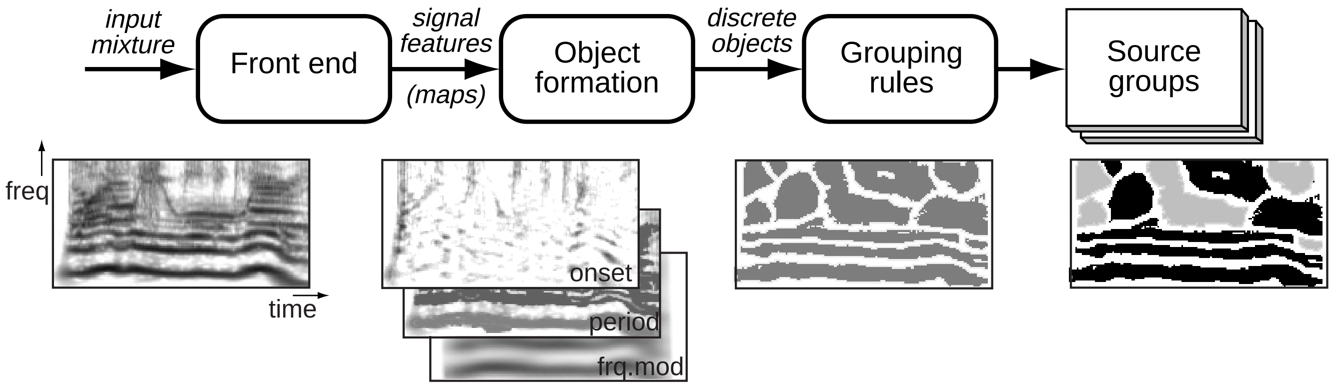


- linear filterbank cochlear approximation
- static nonlinearity
- zero-delay slice is like spectrogram
- periodicity from delay-and-multiply detectors

The Representational Approach

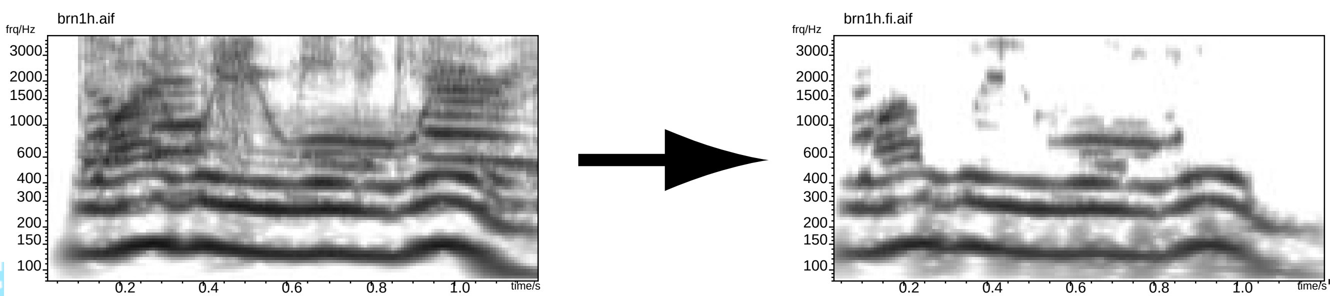
(Brown & Cooke 1993)

- **Implement psychoacoustic theory**

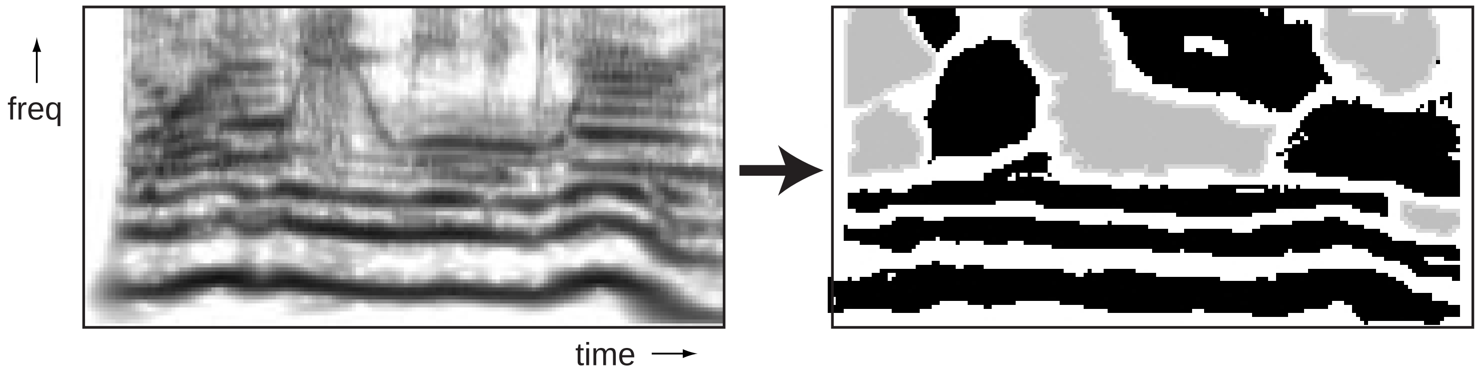


  - ‘bottom-up’ processing
  - uses common onset & periodicity cues

- **Able to extract voiced speech:**



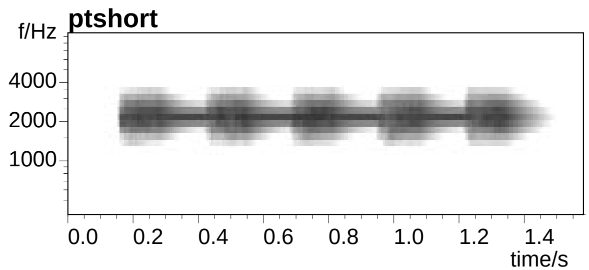
Problems with ‘bottom-up’ CASA



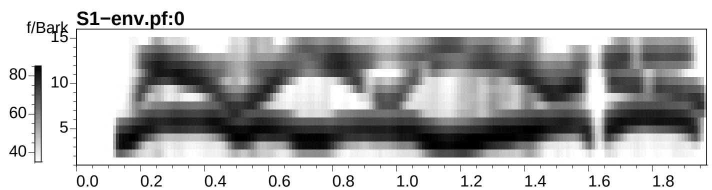
- **Circumscribing time-frequency elements**
  - need to have ‘regions’, but hard to find
- **Periodicity is the primary cue**
  - how to handle aperiodic energy?
- **Resynthesis via masked filtering**
  - cannot separate within a single t-f element
- **Bottom-up leaves no ambiguity or context**
  - how to model illusions?

Restoration in sound perception

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



- **SWS**

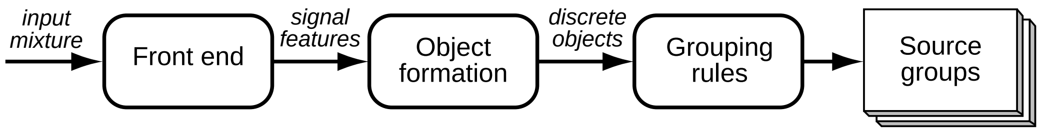


  - duplex perception

Adding top-down constraints

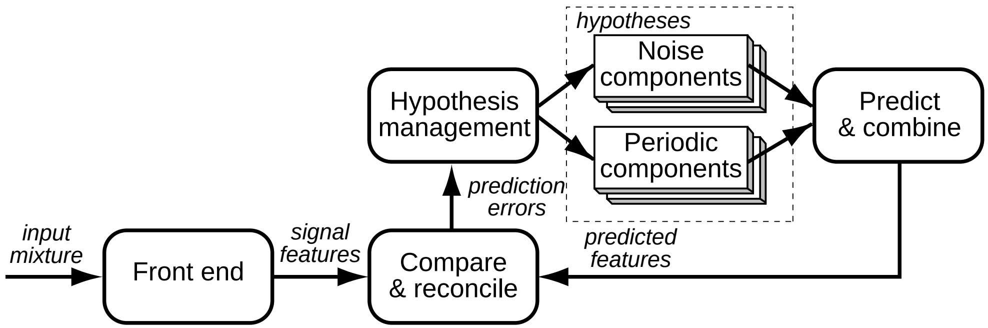
Perception is not *direct*
but a *search* for *plausible hypotheses*

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



  - objects irresistibly appear

vs. Prediction-driven (top-down)

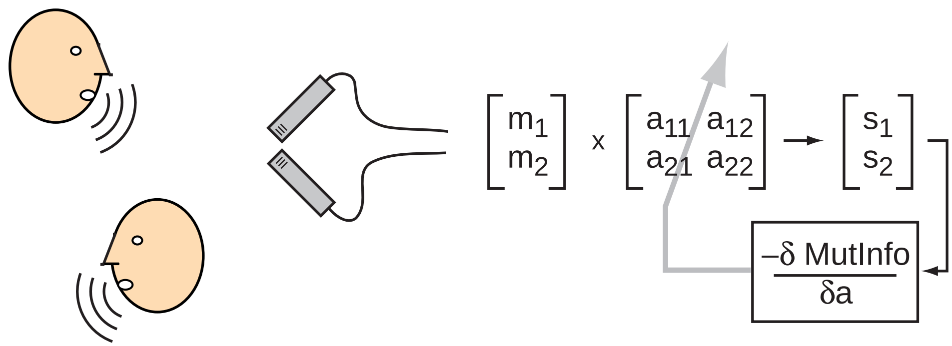


  - match observations with parameters of a world-model
  - need world-model constraints...

Aside: Optimal techniques (ICA, ABF)

(Bell & Sejnowski etc.)

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



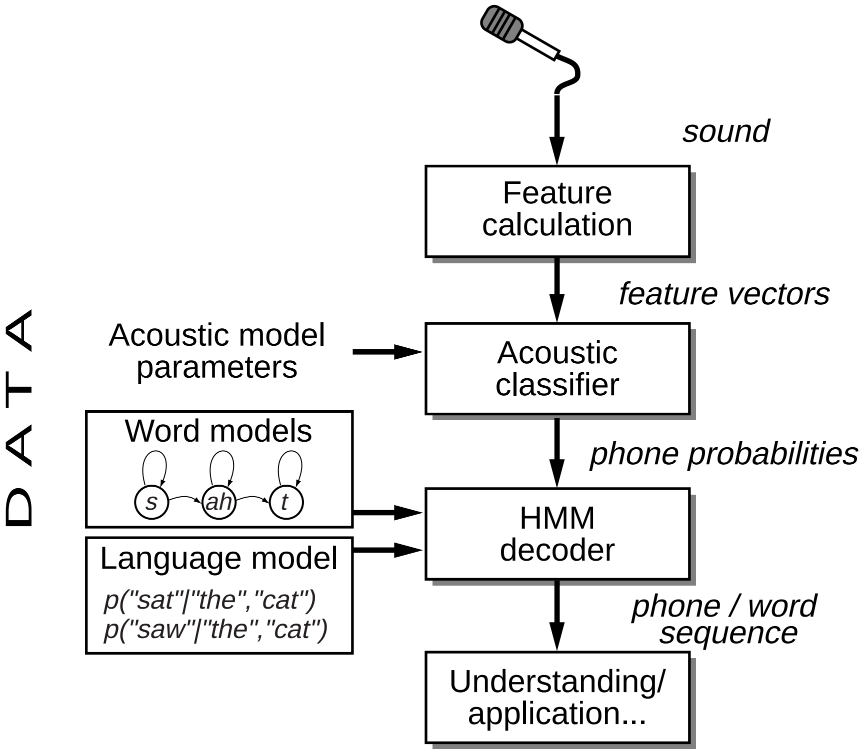
- **Attractions:**
  - mathematically rigorous, minimal assumptions
- **Problems:**
  - limitations of separation algorithm (N x N)
  - essentially bottom-up

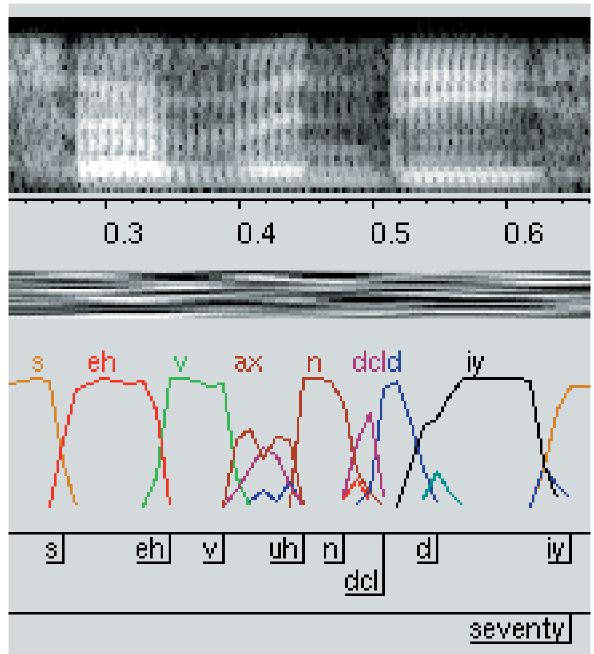
Outline

1. Human sound organization
2. Computational Auditory Scene Analysis
3. **Speech models and knowledge**
   - automatic speech recognition
   - subword states
   - cepstral coefficients
4. Sound mixture recognition
5. Learning opportunities

3 Speech models & knowledge

- **Standard speech recognition**



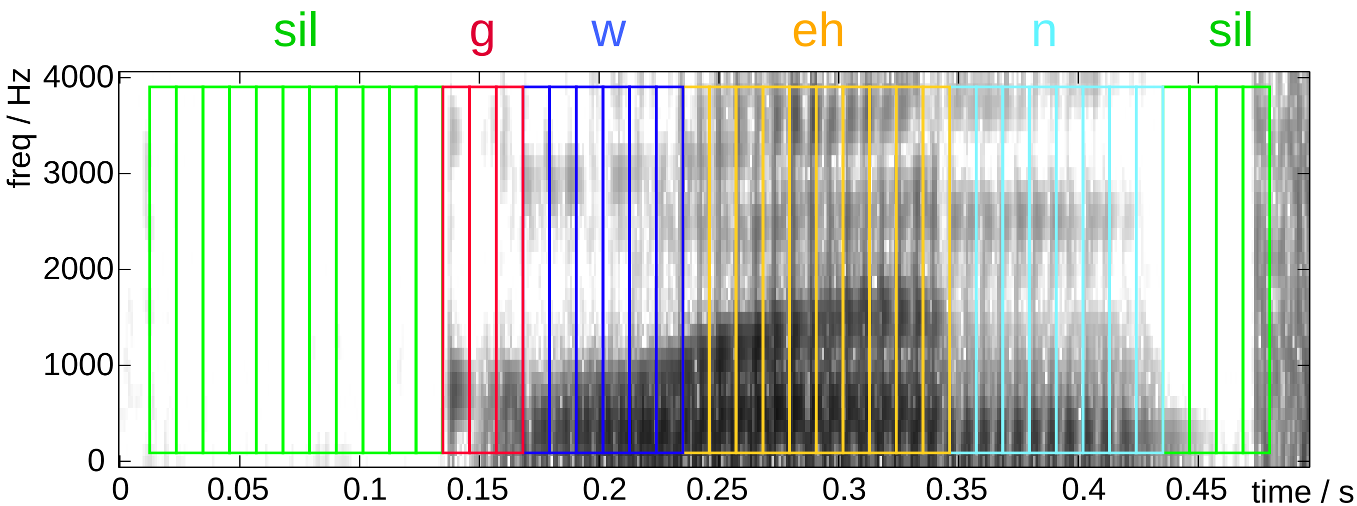


- **'State of the art' word-error rates (WERs):**
  - 2% (dictation) - 30% (telephone conversations)

Speech units

- **Speech is highly variable**
  - simple templates won't do
  - spectral variation (voice quality)
  - *time-warp* problems

- **Match short segments (states), allow repeats**
  - model with pseudo-stationary slices of ~ 10 ms



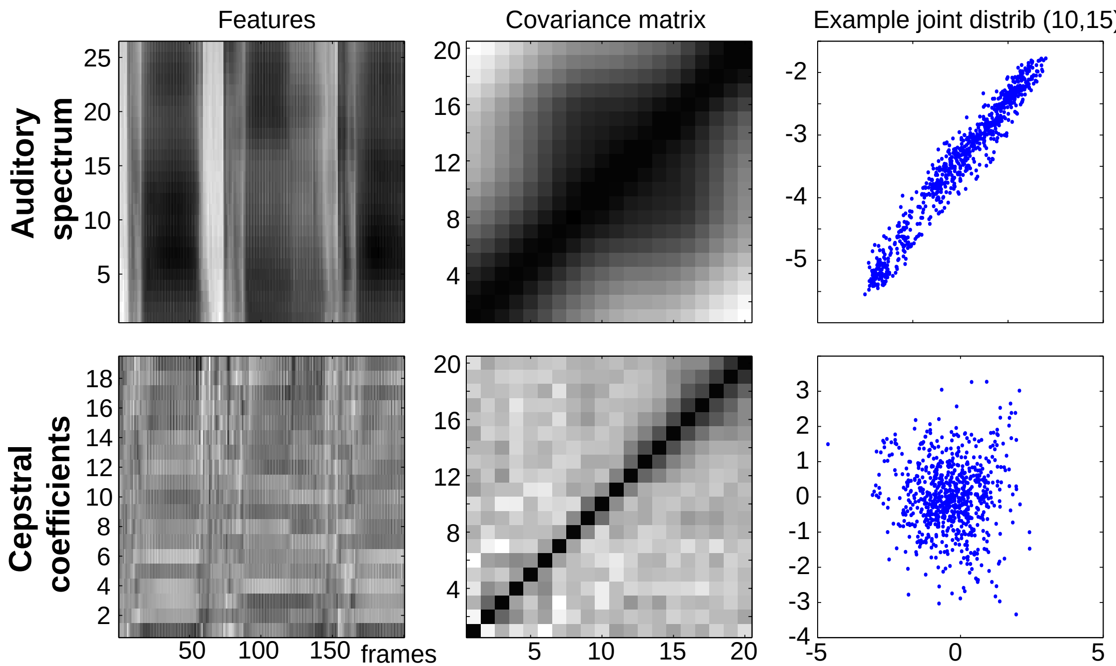
- **Speech models are distributions** $p(X|q)$

Speech features: Cepstra

- **Idea: Decorrelate & summarize spectral slices:**

$$X_m[l] = IDFT\{\log|S[mH,k]|\}$$

  - easier to model:

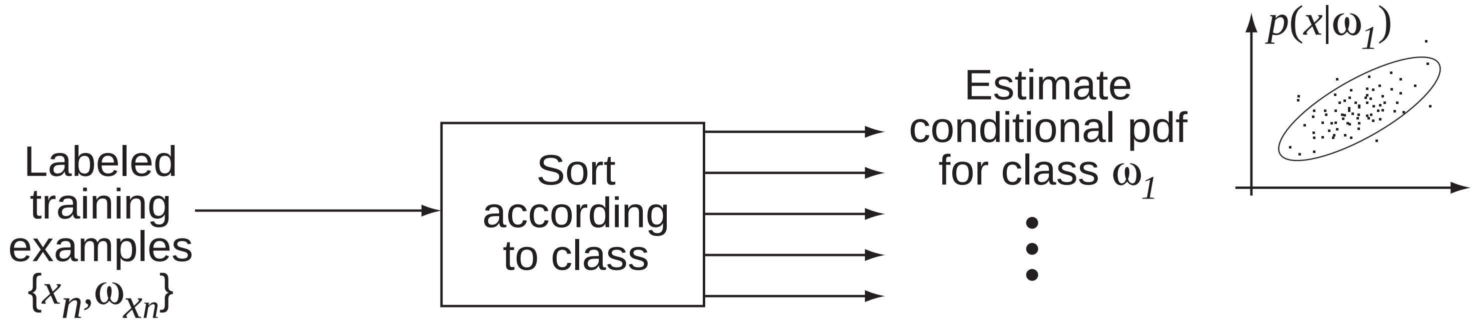


  - C_0 'normalizes out' average log energy

- **Decorrelated pdfs fit diagonal Gaussians**
  - DCT is close to PCA for log spectra

Acoustic model training

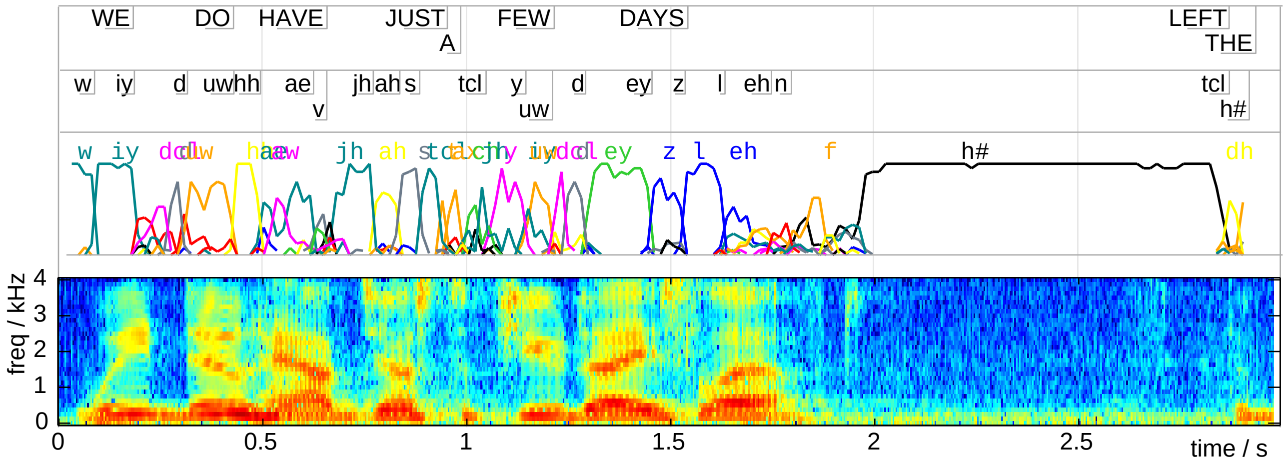
- **Goal: describe $p(X|q)$ with e.g. GMMs**



- **Training data labels from:**
  - manual phonetic annotation
  - ‘best path’ from earlier classifier (Viterbi)
  - EM: joint estimation of labels & pdfs

HMM decoding

- **Feature vectors cannot be reliably classified into phonemes**



- **Use top-down constraints to get good results**
  - allowable phonemes
  - dictionary of known words
  - grammar of possible sentences
- **Decoder searches all possible state sequences**
  - at least notionally; pruning makes it possible

Outline

1. Human sound organization
2. Computational Auditory Scene Analysis
3. Speech models and knowledge
4. **Sound mixture recognition**
    - feature invariance
    - mixtures including
    - general mixtures
5. Learning opportunities

4 Sound mixture recognition

- **Biggest problem in speech recognition is background noise interference**
- **Feature invariance approach**
  - use features that reflect only speech
  - e.g. normalization, mean subtraction
  - but: non-static noise?
- **Or: more complex models of the signal**
  - HMM decomposition
  - missing-data recognition
- **Generalize to other, multiple sounds**

Feature normalization

- **Idea: feature *variations*, not absolute level**
- **Hence: calculate average level & subtract it:**

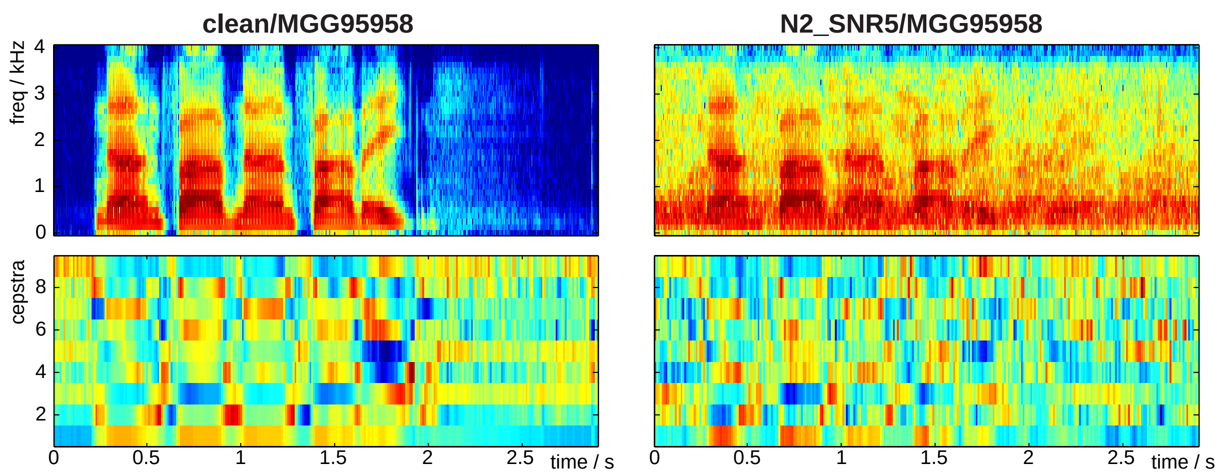
$$X[k] = S[k] - \mathrm{mean}\{S[k]\}$$

- **Factors out fixed channel frequency response:**

$$s[n] = h[n] * e[n]$$

$$\log|S[k]| = \log|H[k]| + \log|E[k]|$$

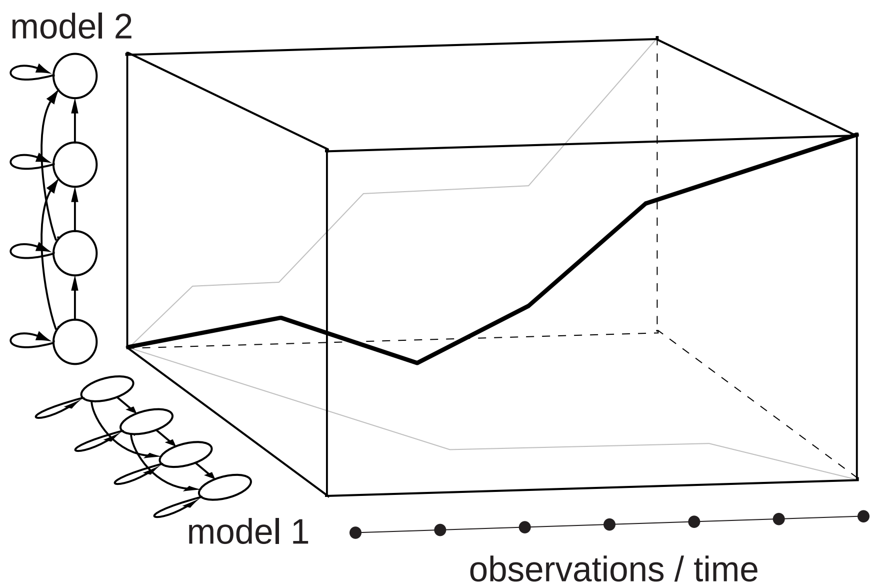
- **Normalize variance to handle added noise?**



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

$$p(X^i \mid q_1^i, q_2^i)$$

Problems with HMM decomposition

- $O(q_k)^N$ **is exponentially large...**
- ***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**

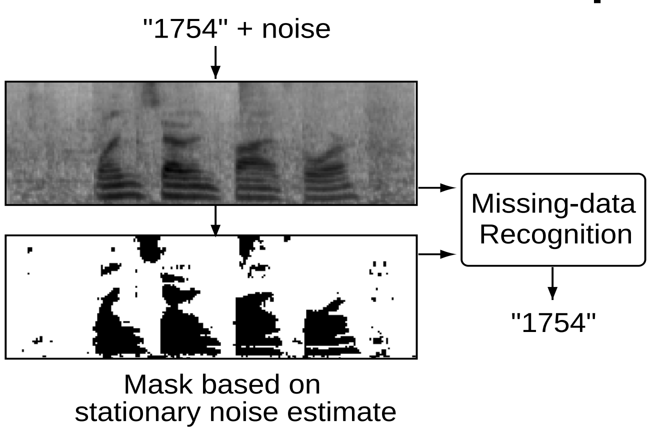
Missing data recognition

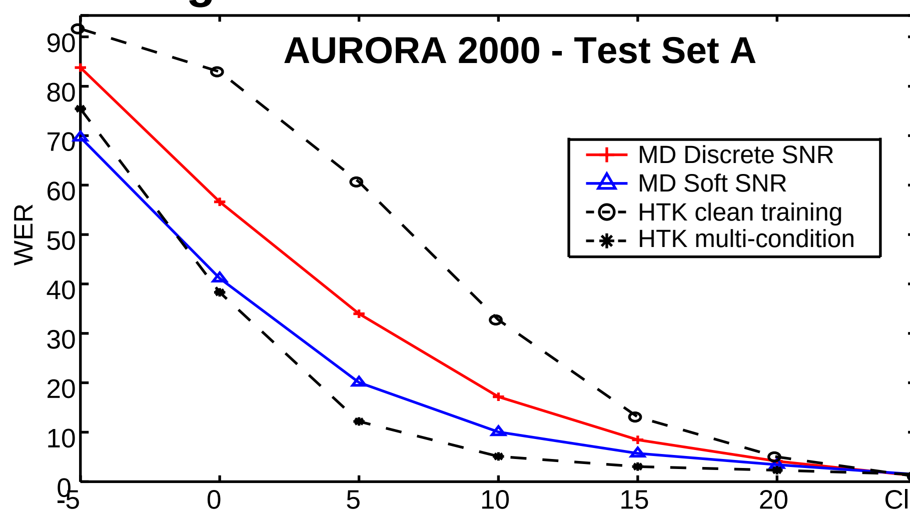
(Cooke, Green, Barker @ Sheffield)

- **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**



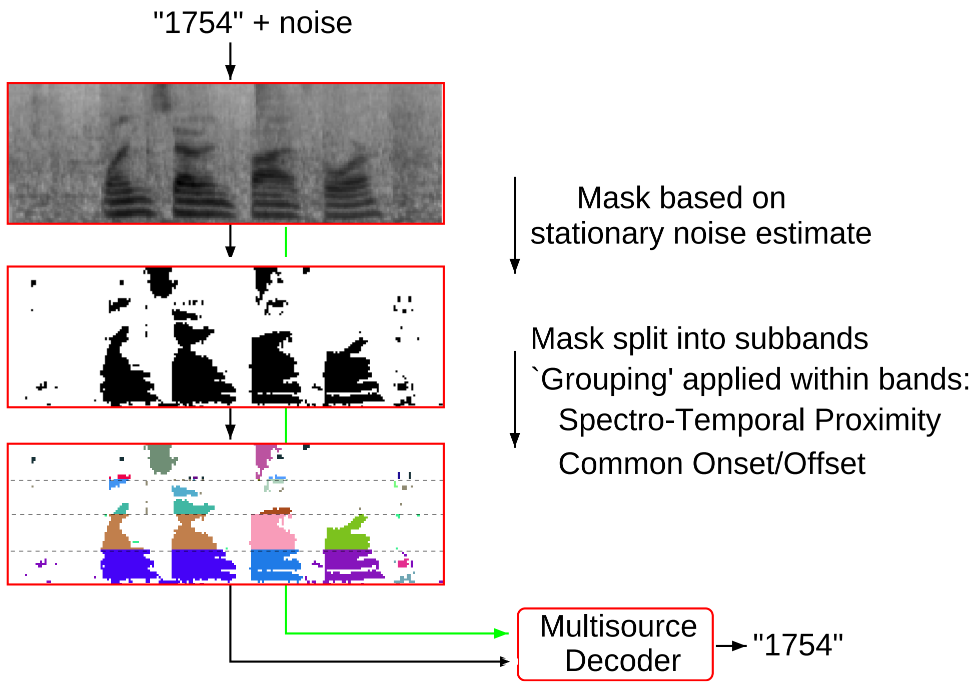


- **Problem: finding the missing data mask**

Maximum-likelihood data mask

(Jon Barker @ Sheffield)

- **Search of sound-fragment interpretations**



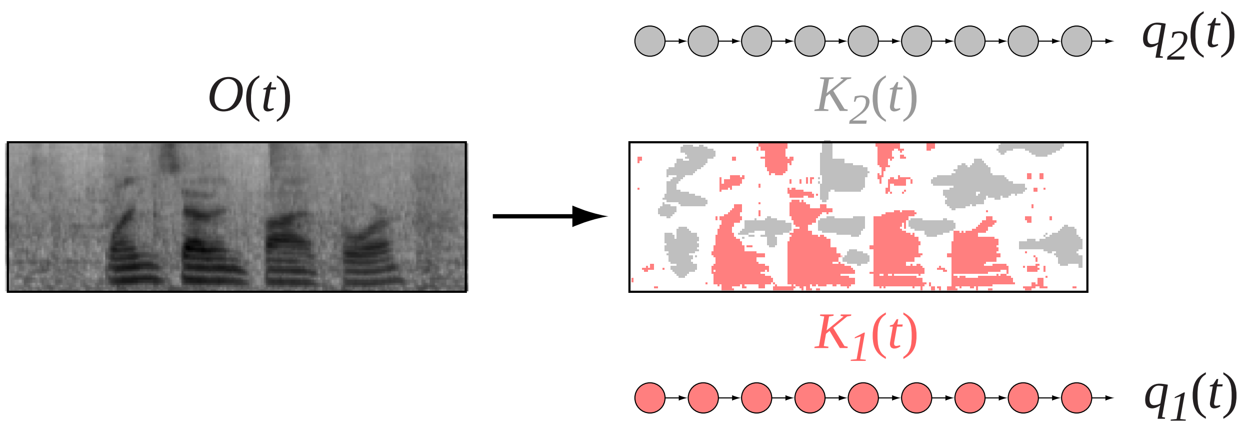
- **Decoder searches over data mask K:**

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

  - how to estimate $p(K)$

Multi-source decoding

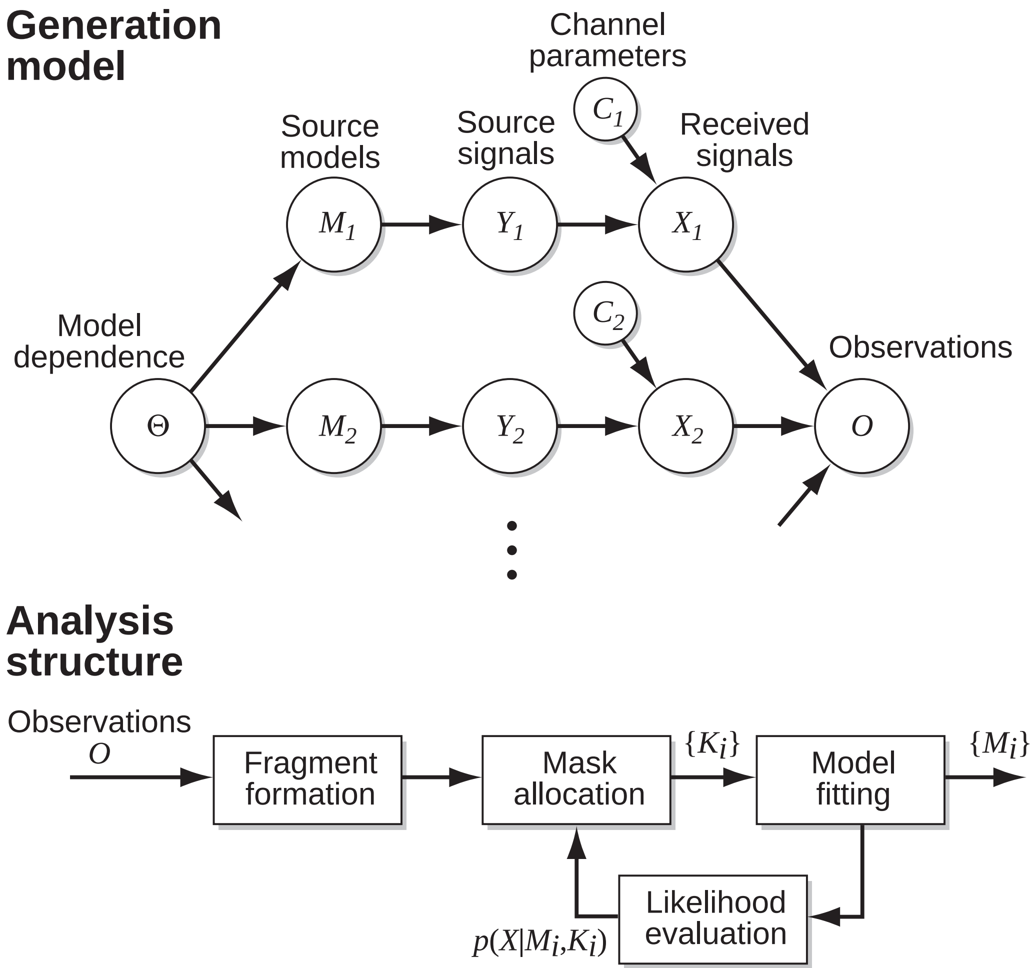
- **Search for more than one source**



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

General sound mixtures

- **Search for generative explanation:**



Outline

1 Human sound organization

2 Computational Auditory Scene Analysis

3 Speech models and knowledge

4 Sound mixture recognition

5 **Opportunities for learning**
   - learnable aspects of modeling
   - tractable decoding
   - some examples

5 Opportunities for learning

- **Per model feature distributions** $P(Y|M)$
  - e.g. analyzing isolated sound databases
- **Channel modifications** $P(X|Y)$
  - e.g. by comparing multi-mic recordings
- **Signal combinations** $P(O|\{X_i\})$
  - determined by acoustics
- **Patterns of model combinations** $P(\{M_i\})$
  - loose dependence between sources
- **Search for most likely explanations**

$$P(\{M_i\}|O) \propto P(O|\{X_i\})P(\{X_i\}|\{M_i\})P(\{M_i\})$$

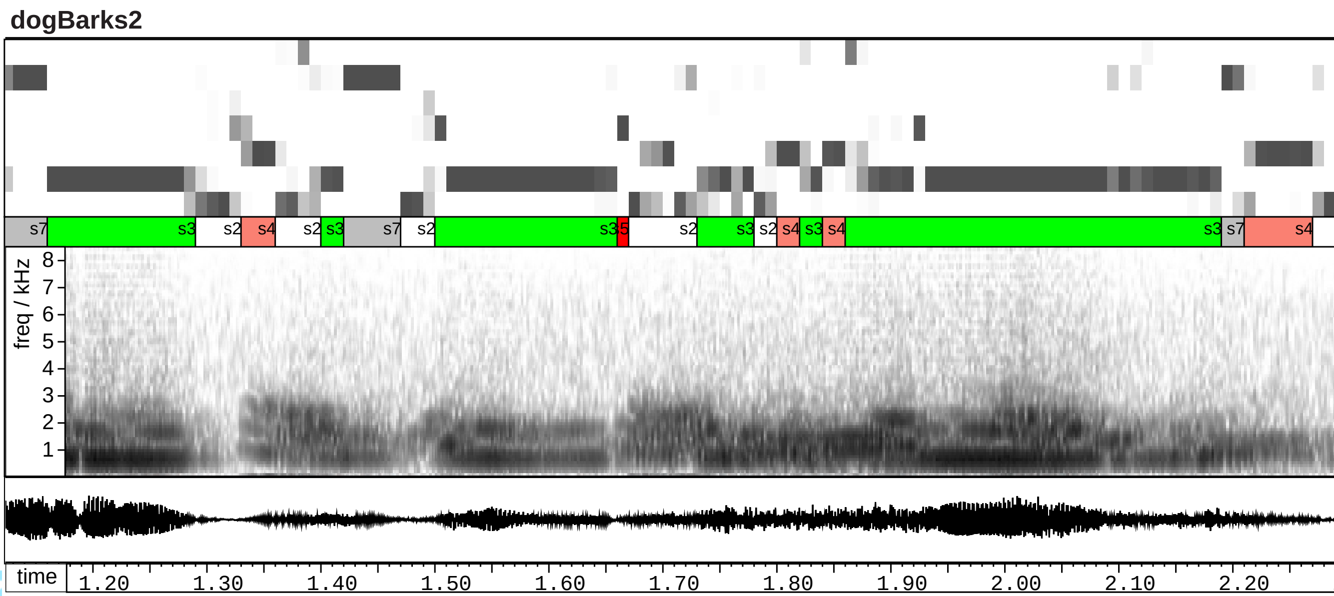
- **Short-term structure: repeating events**

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
- **Model types**
  - templates
  - parametric/constraint models
  - HMMs

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

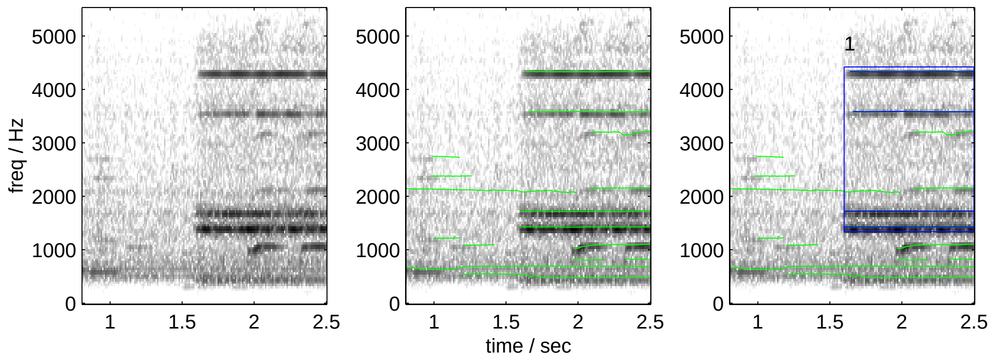


Tractable decoding

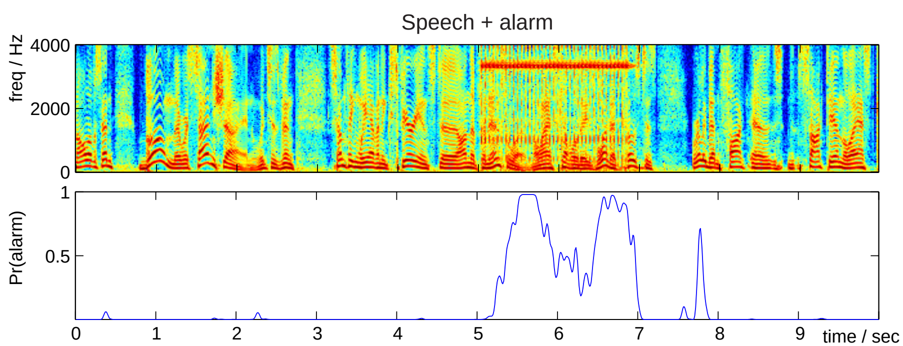
- **Speech decoder notionally searches all states**
- **Parametric models give infinite space**
  - need closed-form partial explanations
  - examine residual, iterate, converge
- **Need general cues to get started**
  - return to Auditory Scene Analysis:
    - onsets
    - harmonic patterns
  - then parametric fitting
- **Need multiple hypothesis search, pruning, efficiency tricks**
- **Learning?**
  Parameters for new source events
  - e.g. from artificial (hence labeled) mixtures

Example: Alarm sound detection

- **Alarm sounds have particular structure**
  - people ‘know them when they hear them’
- **Isolate alarms in sound mixtures**



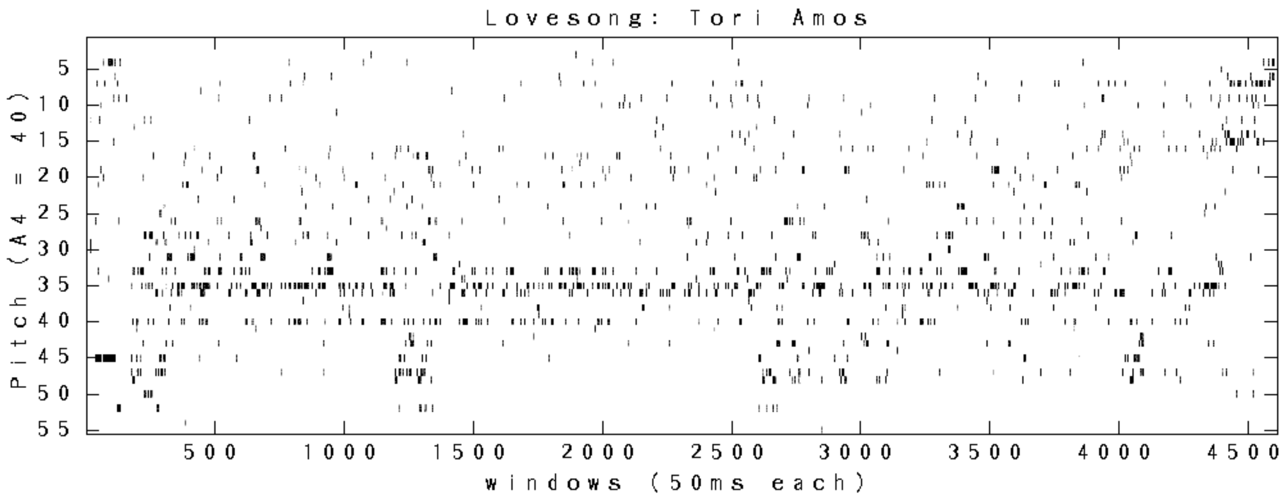
  - sinusoid peaks have invariant properties



- **Learn model parameters from examples**

Example: Music transcription
(e.g. Masataka Goto)

- **High-quality training material: Synthesizer sample kits**
- **Ground truth available: Musical scores**
- **Find ML explanations for scores**
  - guide by multiple pitch tracking (hyp. search)



- **Applications in similarity matching**

Summary

- **Sound contains lots of information**

  ... but it’s always mixed up

- **Psychologists describe ASA**

  ... but bottom-up computer models don’t work

- **Speech recognition works for isolated speech**

  ... by exploiting top-down, context constraints

- **Speech in mixtures via multiple-source models**

  ... practical combinatorics are the main problem

- **Generalize this idea for all sounds**

  ... need models of ‘all sounds’

  ... plus models of channel modification

  ... plus ways to propose segmentations

  ... plus missing-data recognition

Further reading

[BarkCE00] J. Barker, M.P. Cooke & D. Ellis (2000). "Decoding speech in the presence of other sound sources," *Proc. ICSLP-2000*, Beijing.
ftp://ftp.icsi.berkeley.edu/pub/speech/papers/icslp00-msd.pdf

[Breg90] A.S. Bregman (1990). *Auditory Scene Analysis: the perceptual organization of sound*, MIT Press.

[Chev00] A. de Cheveigné (2000). "The Auditory System as a Separation Machine," Proc. Intl. Symposium on Hearing.
http://www.ircam.fr/pcm/cheveign/sh/ps/ATReats98.pdf

[CookE01] M. Cooke, D. Ellis (2001). "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication* (accepted for publication).
http://www.ee.columbia.edu/~dpwe/pubs/tcfkas.pdf

[Ellis99] D.P.W. Ellis (1999). "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis...," *Speech Communications* 27.
http://www.ee.columbia.edu/~dpwe/pubs/spcomcasa98.pdf

[Roweis00] S. Roweis (2000). "One microphone source separation.," *Proc. NIPS 2000*.
http://www.ee.columbia.edu/~dpwe/papers/roweis-nips2000.pdf