EE E6820: Speech & Audio Processing & Recognition

Lecture 11:
Signal Separation

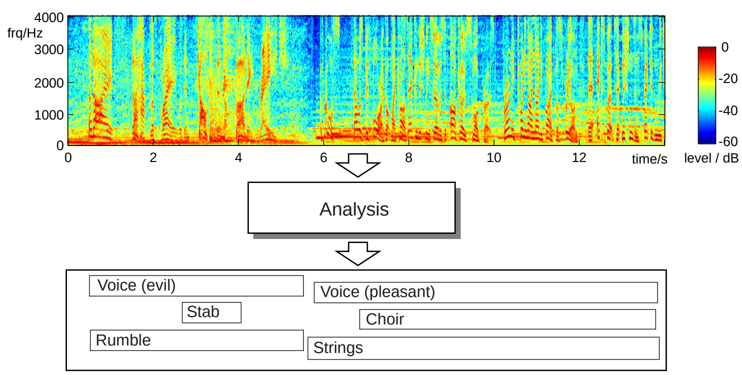
1. **Sound Mixture Organization**
2. **Computational Auditory Scene Analysis**
3. **Independent Component Analysis**
4. **Model-Based Separation**

Dan Ellis <dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/e6820/

Columbia University Dept. of Electrical Engineering

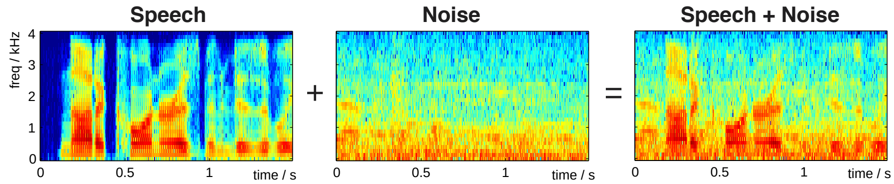
Spring 2006

1 Sound Mixture Organization



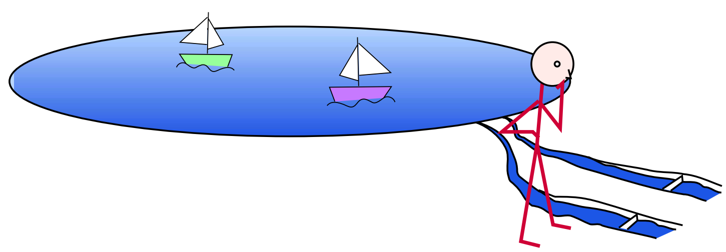
- **Auditory Scene Analysis: describing a complex sound in terms of high-level sources / events**
  - ... like listeners do
- **Hearing is ecologically grounded**
  - reflects ‘natural scene’ properties
  - subjective, not absolute

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



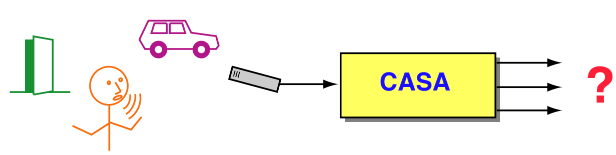
“Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?” (after Bregman’90)

- **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

Approaches to sound mixture recognition

- **Separate signals, then recognize**
  - e.g. Computational Auditory Scene Analysis (CASA), Independent Component Analysis (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

What is the goal of CASA?



- **Separate signals?**
  - output is unmixed waveforms
  - underconstrained, very hard ...
  - too hard? not required?
- **Source classification?**
  - output is set of event-names
  - listeners do more than this...
- **Something in-between?**
  Identify independent sources + characteristics
  - standard task, results?

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

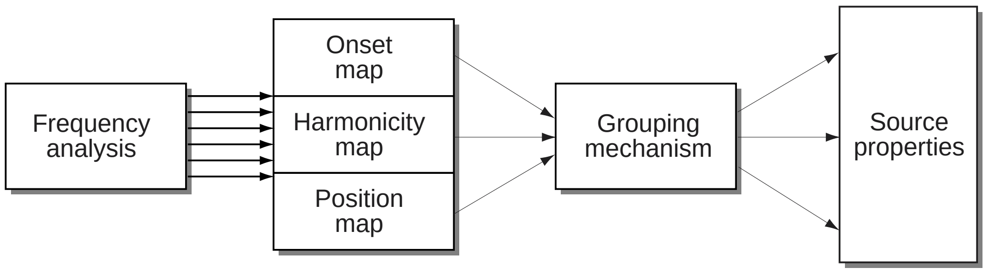
Outline

1. Sound Mixture Organization
2. **Computational Auditory Scene Analysis**
   - Human Auditory Scene Analysis
   - Bottom-up and Top-down models
   - Evaluation
3. Independent Component Analysis
4. Model-Based Separation

2 Auditory Scene Analysis

(Bregman 1990)

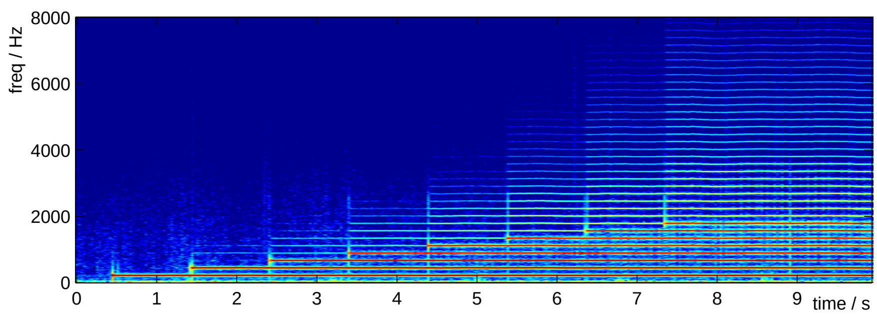
- **How do people analyze sound mixtures?**
  - break mixture into small elements (in time-freq)
  - elements are grouped in to sources using cues
  - sources have aggregate attributes
- **Grouping 'rules' (Darwin, Carlyon, ...):**
  - cues: common onset/offset/modulation, harmonicity, spatial location,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



(after Darwin, 1996)

Cues to simultaneous grouping

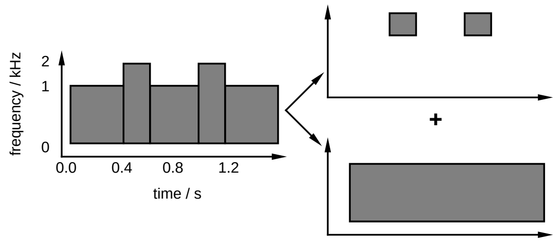
- **Elements + attributes**



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

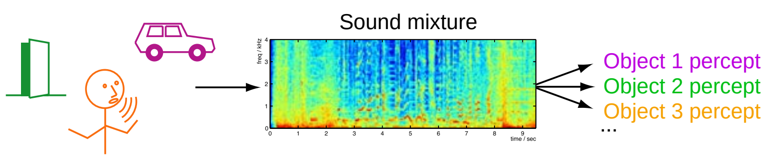
The effect of context

- **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

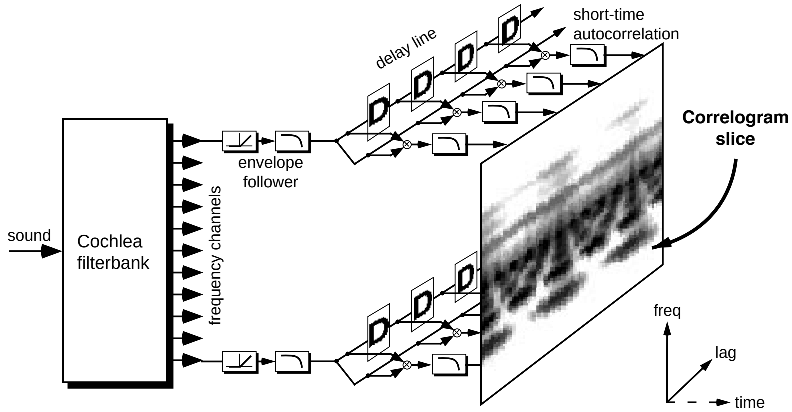
Computational Auditory Scene Analysis (CASA)



- **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?

CASA front-end processing

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

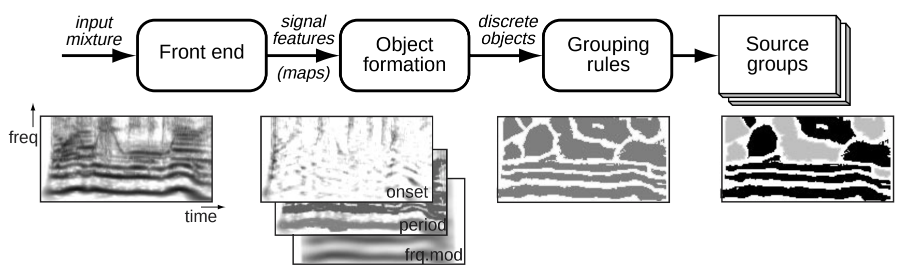


- linear filterbank cochlear approximation
- static nonlinearity
- zero-delay slice is like spectrogram
- periodicity from delay-and-multiply detectors

The Representational Approach

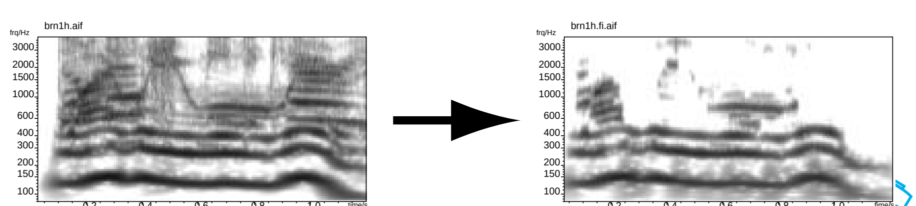
(Brown & Cooke 1993)

- **Implement psychoacoustic theory**

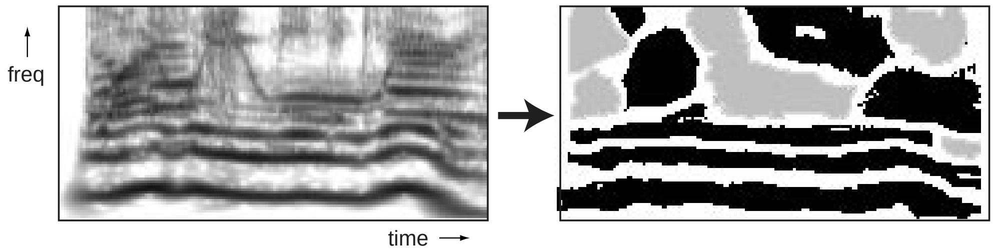


- 'bottom-up' processing
- uses common onset & periodicity cues

- **Able to extract voiced speech:**



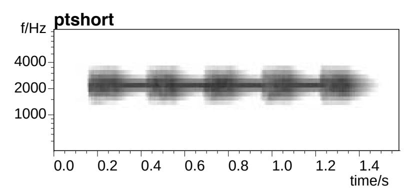
Problems with ‘bottom-up’ CASA



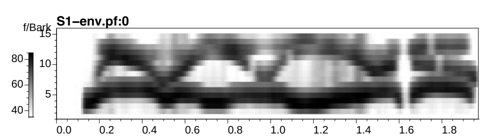
- **Circumscribing time-frequency elements**
  - need to have ‘regions’, but hard to find
- **Periodicity is the primary cue**
  - how to handle aperiodic energy?
- **Resynthesis via masked filtering**
  - cannot separate within a single t-f element
- **Bottom-up leaves no ambiguity or context**
  -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

Restoration in sound perception

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



- **Sinewave Speech (SWS)**

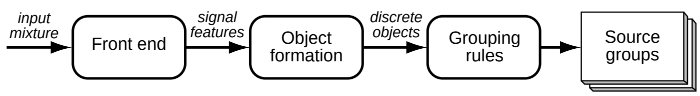


  - duplex perception
- **What kind of model accounts for this?**
  - is it an important part of hearing?

Adding top-down constraints: Prediction-Driven CASA (PDCASA)

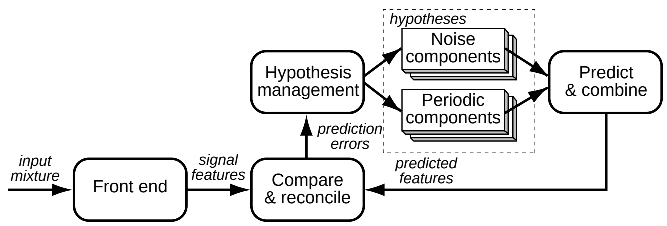
Perception is not direct
but a search for plausible hypotheses

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



  - objects irresistibly appear

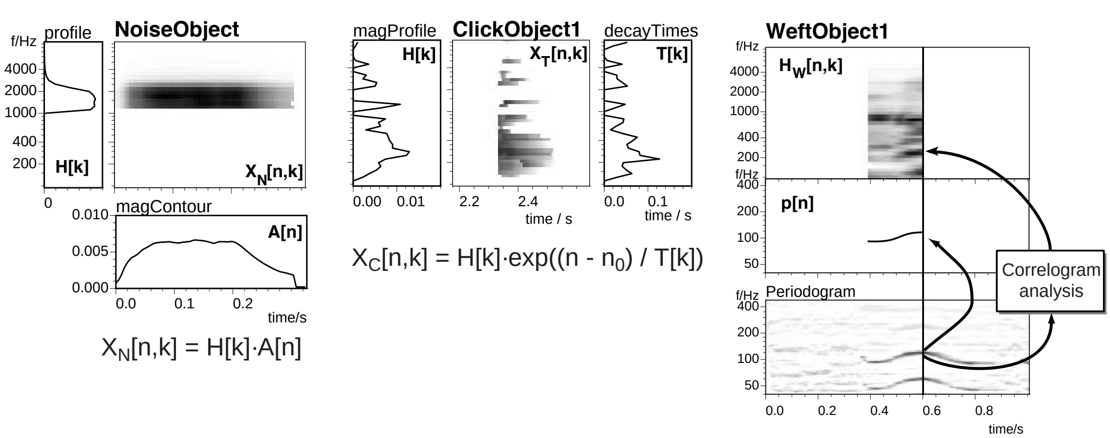
vs. Prediction-driven (top-down)



  - match observations with a 'world-model'
  - need world-model constraints...

Generic sound elements for PDCASA

- **Goal is a representational space that**
  - covers real-world perceptual sounds
  - minimal parameterization (sparseness)
  - separate attributes in separate parameters



$X_N[n,k] = H[k]\cdot A[n]$

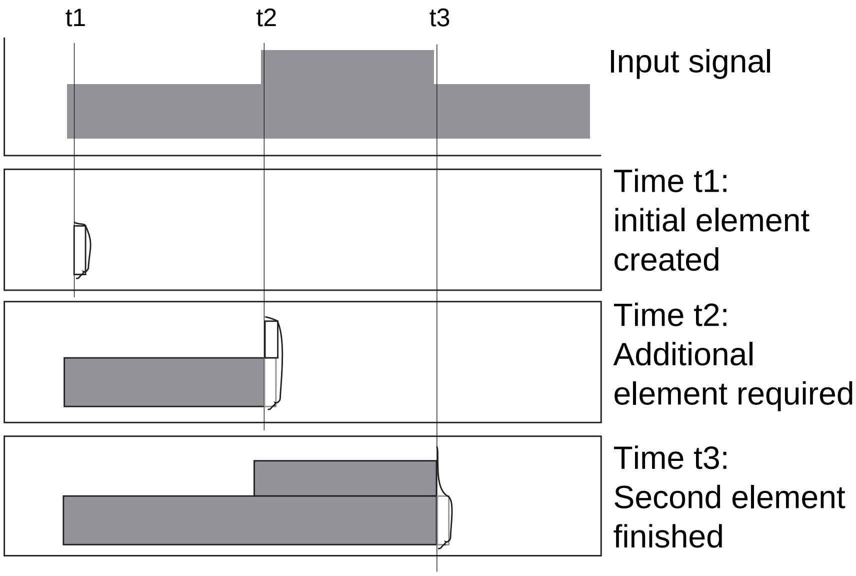
$X_C[n,k] = H[k]\cdot\exp((n - n_0) / T[k])$

$X_W[n,k] = H_W[n,k]\cdot P[n,k]$

- **Object hierarchies built on top...**

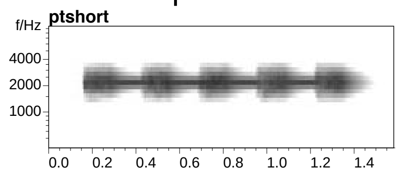
PDCASA for old-plus-new

- **Incremental analysis**

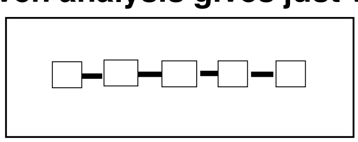


PDCASA for the continuity illusion

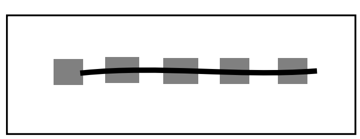
- **Subjects hear the tone as continuous**
  ... if the noise is a plausible masker



- **Data-driven analysis gives just visible portions:**

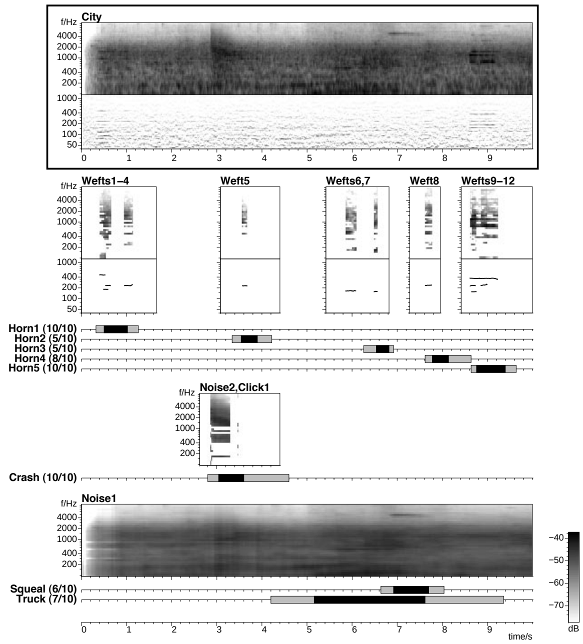


- **Prediction-driven can infer masking:**



Prediction-Driven CASA

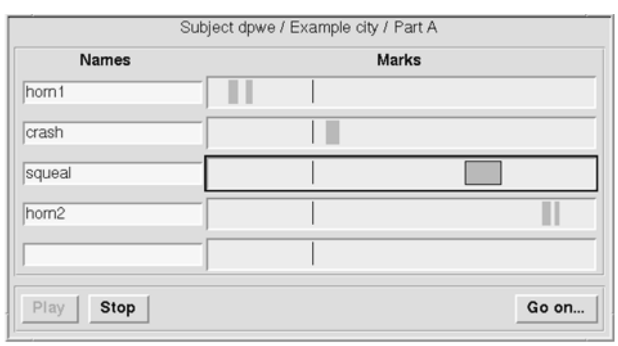
- **Explain a complex sound with basic elements**



Aside: Ground Truth

- **What do people hear in sound mixtures?**
  - do interpretations match?

→ **Listening tests to collect 'perceived events':**



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...

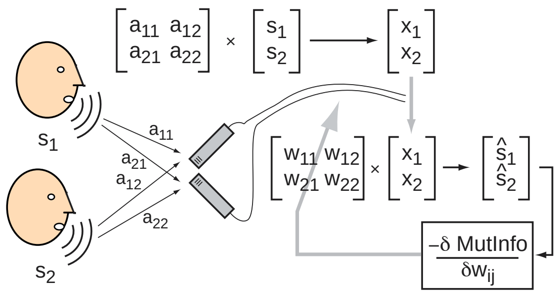
Outline

1. Sound Mixture Organization
2. Computational Auditory Scene Analysis
3. **Independent Component Analysis**
   - Blind source separation
   - Independence and kurtosis
   - Limits of the approach
4. Model-Based Separation

❸ Independent Component Analysis (ICA)

(Bell & Sejnowski 1995 etc.)

- **If mixing is like matrix multiplication, then separation is searching for the inverse matrix**

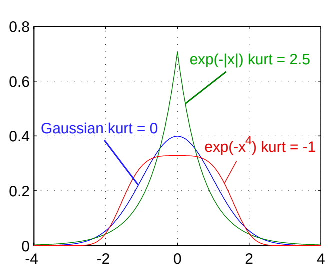


  - i.e. $W \approx A^{-1}$
  - with *N* different versions of the mixed signals (microphones), we can find *N* different input contributions (sources)
  - how to rate quality of outputs?
    i.e. when do outputs look separate?

Gaussianity, Kurtosis & Independence

- **A signal can be characterized by its PDF p(x)**
  - i.e. as if successive time values are drawn from a random variable (RV)
  - Gaussian PDF is 'least interesting'
  - Sums of independent RVs (PDFs convolved) tend to Gaussian PDF (Weak law of large nums)
- **Measures of deviations from Gaussianity: 4th moment is Kurtosis ("bulging")**

$$kurt(y) = E\left[\left(\frac{y-\mu}{\sigma}\right)^4\right] - 3$$



-kurtosis of Gaussian is zero (this def.)

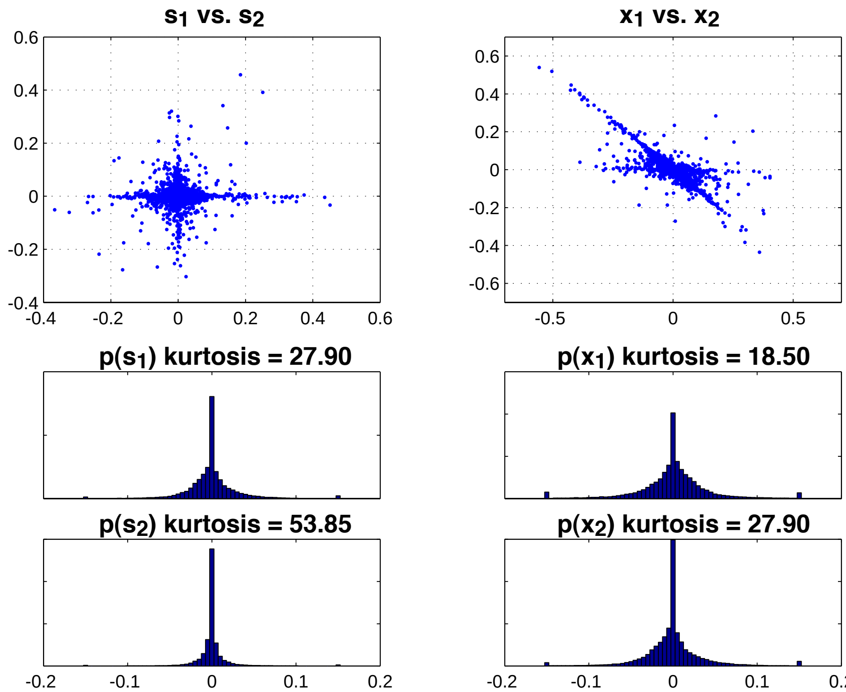
-'heavy tails' → *kurt* > 0

-closer to uniform dist. → *kurt* < 0

•**Directly related to KL divergence from Gaussian PDF**

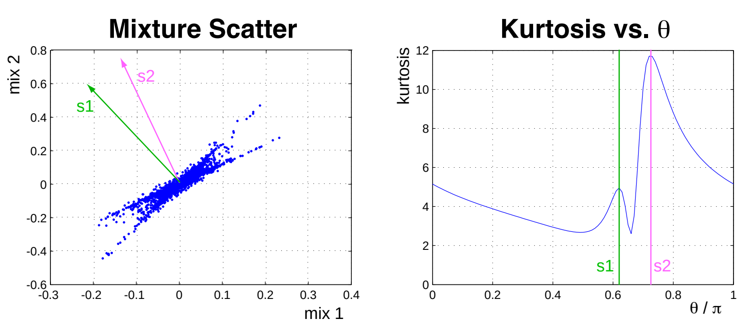
Independence in Mixtures

- **Scatter plots & Kurtosis values**



Finding Independent Components

- **Sums of independent RVs are more Gaussian**
  → minimize Gaussianity to undo sums
  - i.e. search over w_{ij} terms in inverse matrix



- **Solve by Gradient descent or Newton-Raphson:**

$$\mathbf{w}^+ = E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$$

$$\mathbf{w} = \mathbf{w}^+ / \|\mathbf{w}^+\|$$

"Fast ICA", http://www.cis.hut.fi/projects/ica/fastica/

Limitations of ICA

- **Assumes instantaneous mixing**
  - real world mixtures have delays & reflections
  - STFT domain?

$$x_1(t) = a_{11}(t) \otimes s_1(t) + a_{12}(t) \otimes s_2(t)$$

$$\Rightarrow X_1(\omega) = A_{11}(\omega)S_1(\omega) + A_{12}(\omega)S_2(\omega)$$

  Solve ω subbands separately, match up answers

- **Searching for best possible inverse matrix**
  - cannot find more than *N* outputs from *N* inputs
    but: "projection pursuit" ideas
      + time-frequency masking...

- **Cancellation inherently fragile**
  - $\hat{s}_1 = w_{11} \cdot x_1 + w_{12} \cdot x_2$ to cancel out s_2
  - sensitive to noise in x channels
  - time-varying mixtures are a problem

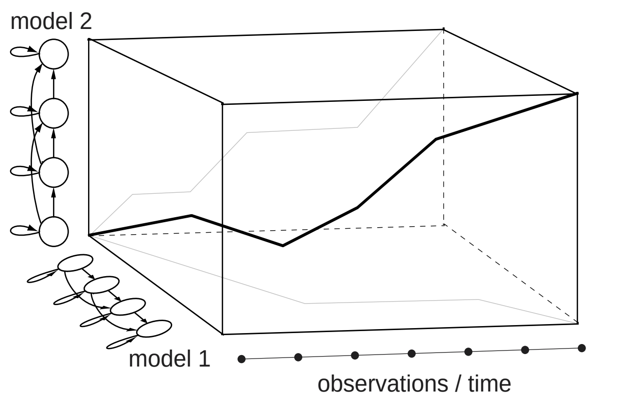
Outline

1. Sound Mixture Organization
2. Computational Auditory Scene Analysis
3. Independent Component Analysis
4. **Model-Based Separation**
   - Fitting models to mixtures
   - Missing-data recognition
   - Speech Fragment Decoding

4 Model-Based Separation: HMM decomposition

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

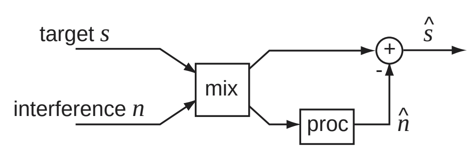
- Independent state sequences for 2+ component source models



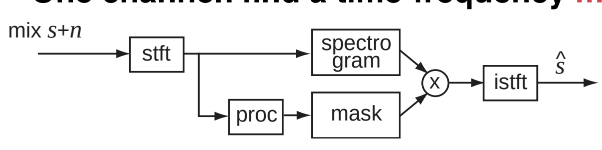
- New combined state space $q' = \{q_1\ q_2\}$
  - need pdfs for combinations $p(X|q_1, q_2)$

One-channel Separation: Masked Filtering

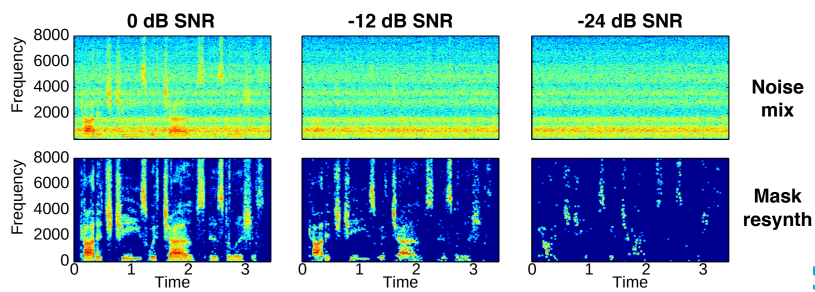
- Multichannel → ICA: Inverse filter & cancel



- One channel: find a time-frequency mask



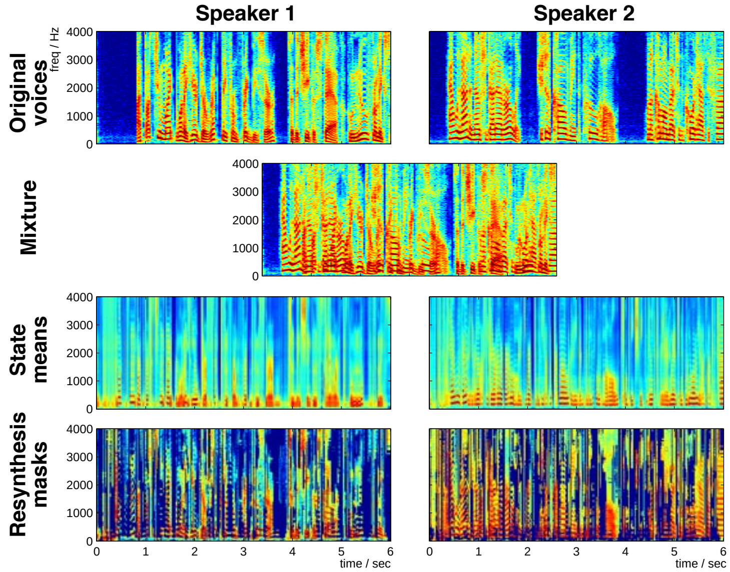
- Cannot remove overlapping noise in TF cells, but surprisingly effective (psy. masking?):



“One microphone source separation”

(Roweis 2000, Manuel Reyes)

- **State sequences → t-f estimates → mask**



- 1000 states/model (→ 10^6 transition probs.)
- simplify by subbands (coupled HMM)?

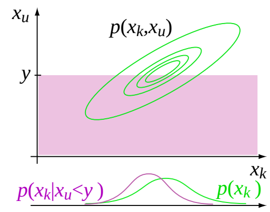
Speech Fragment Recognition

(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 true posterior distribution
- **Goal:**
  Relate clean speech models $P(X|M)$
  to speech-plus-noise mixture observations
  - .. and make it tractable

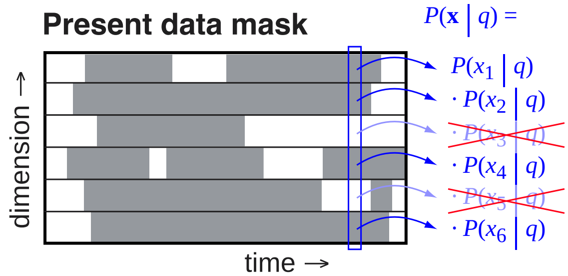
Missing Data Recognition

- **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**



$$p(\mathbf{x}_k|m) = \int p(\mathbf{x}_k, \mathbf{x}_u|m)d\mathbf{x}_u$$

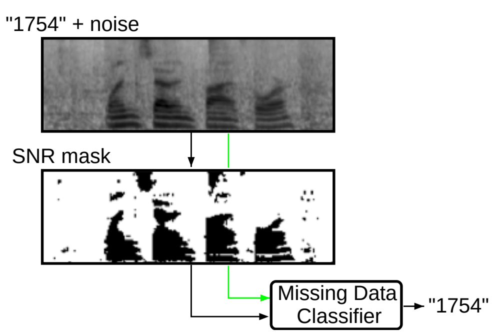
- **Hence, missing data recognition:**

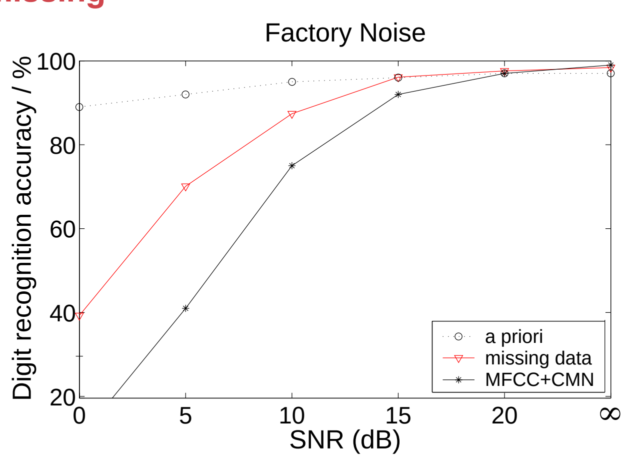


  - hard part is finding the mask (segregation)

Missing Data Results

- **Estimate static background noise level $N(f)$**
- **Cells with energy close to background are considered "missing"**





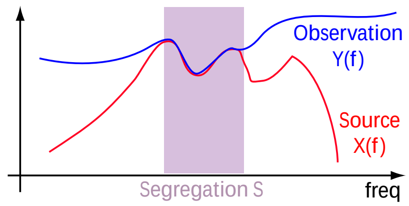
  - must use spectral features!
- **But: nonstationary noise → spurious mask bits**
  - can we try removing parts of mask?

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 $P(X|Y,S)$**



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

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

  - $P(X)$ no longer constant

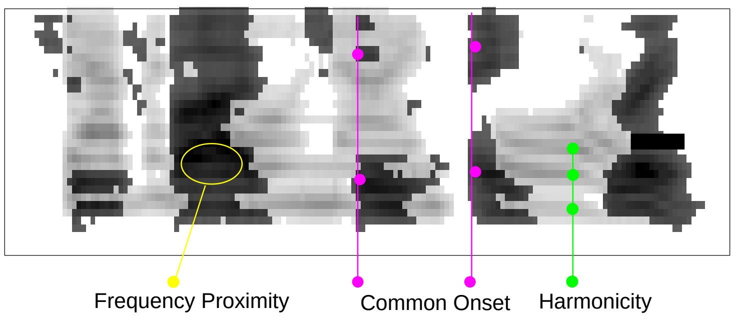
Calculating fragment matches

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

- $P(X|M)$ **- 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...
- **Result:**
  - probabilistically-correct relation between clean-source models $P(X|M)$ and inferred, recognized source + segregation $P(M,S|Y)$

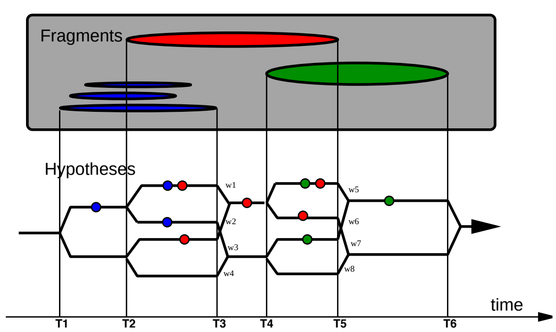
Using CASA features

- $P(S|Y)$ **links acoustic information to segregation**
  - is this segregation worth considering?
  - how likely is it?
- **Opportunity for CASA-style information to contribute**
  - periodicity/harmonicity: these different frequency bands belong together
  - onset/continuity: this time-frequency region must be whole



Fragment decoding

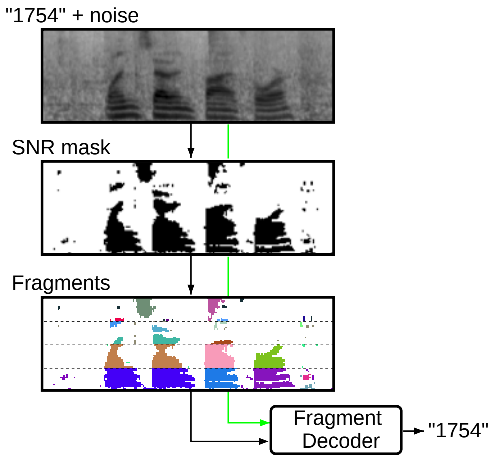
- **Limiting S to whole fragments makes hypothesis search tractable:**

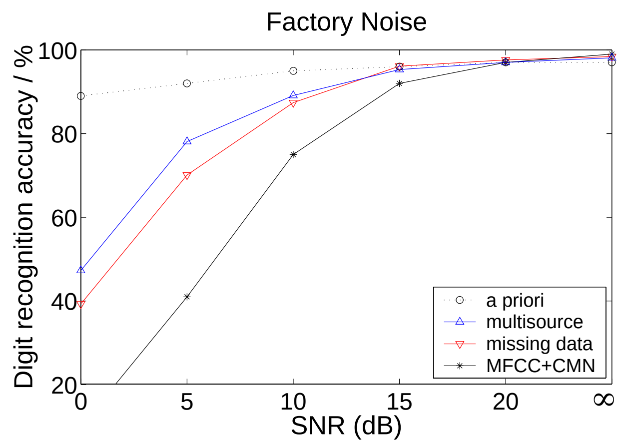


  - choice of fragments reflects $P(S|Y) \cdot P(X|M)$ i.e. best combination of segregation and match to speech models
- **Merging hypotheses limits space demands**
  - .. but erases specific history

Speech fragment decoder results

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

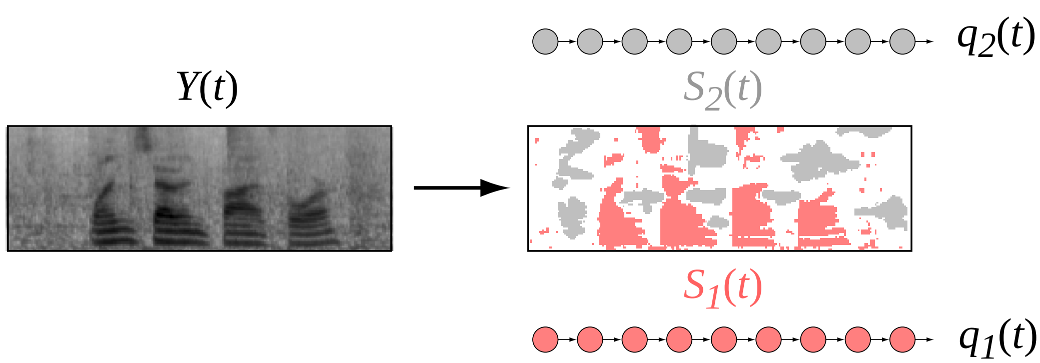




- **Clean-models-based recognition rivals trained-in-noise recognition**

Multi-source decoding

- **Search for more than one source**



- **Mutually-dependent data masks**
  - disjoint subsets of cells for each source
  - each model match $P(M_x|S_x,Y)$ is independent
  - masks are mutually dependent: $P(S_1,S_2|Y)$
- **Huge practical advantage over full search**

Summary

- **Auditory Scene Analysis**:
  Hearing: partially understood, very successful
- **Independent Component Analysis**:
  Simple and powerful, some practical limits
- **Model-based separation**:
  Real-world constraints, implementation tricks

Mixture separation the main obstacle in many applications e.g. soundtrack recognition

References

Aapo Hyvärinen, Errki Oja (2000) "Independent Component Analysis: Algorithms and Applications", *Neural Networks*.
http://www.ee.columbia.edu/~dpwe/e6820/papers/HyvO00-icatut.pdf

Martin Cooke, Dan Ellis (2001) "The auditory organization of speech and other sources in listeners and computational models", *Speech Communication*, vol. 35, no. 3-4, pp. 141-177.
http://www.ee.columbia.edu/~dpwe/pubs/CookeE01-audorg.pdf

Jon Barker, Martin Cooke, Dan Ellis (2004) "Decoding speech in the presence of other sources", *Speech Communication*, to appear.
http://www.ee.columbia.edu/~dpwe/papers/BarkCE04-sfd.pdf

Dan Ellis (1996) "Prediction-driven Computational Auditory Scene Analysis", Ph.D. dissertation, MIT EECS.
http://www.ee.columbia.edu/~dpwe/pdcasa/