EE E6820: Speech & Audio Processing & Recognition

Lecture 10: Signal Separation

Dan Ellis <dpwe@ee.columbia.edu>
Michael Mandel <mim@ee.columbia.edu>

Columbia University Dept. of Electrical Engineering
http://www.ee.columbia.edu/∼dpwe/e6820

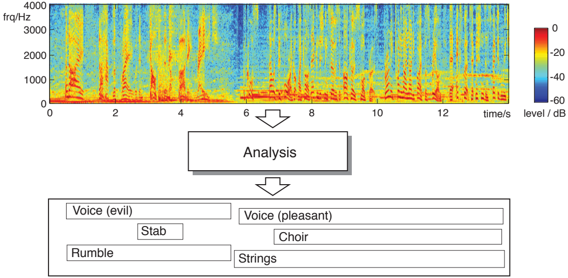
April 14, 2009

1. Sound mixture organization
2. Computational auditory scene analysis
3. Independent component analysis
4. Model-based separation

Outline

1. Sound mixture organization
2. Computational auditory scene analysis
3. Independent component analysis
4. Model-based separation

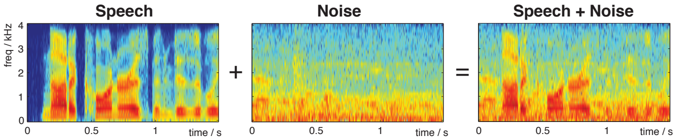
Sound Mixture Organization



- 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

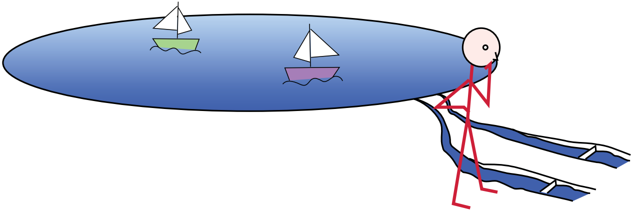
Sound, mixtures, and learning



- 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



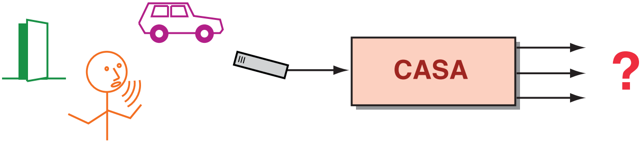
"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman, 1990)

- 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 sound mixture analysis?



- 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, and finer details)
  - need to identify and 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?

$$p(O, x, y) \rightarrow p(x, y \mid O)$$

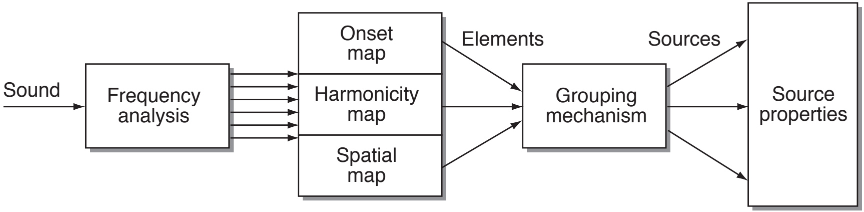
  - or: just skip those values & infer from higher-level context

Outline

1. Sound mixture organization
2. Computational auditory scene analysis
3. Independent component analysis
4. Model-based separation

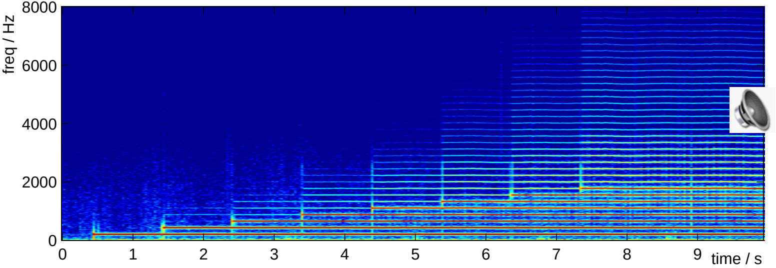
Auditory Scene Analysis (Bregman, 1990)

- 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 and Carlyon, 1995)
  - cues: common onset/offset/modulation, harmonicity, spatial location, ...



Cues to simultaneous grouping

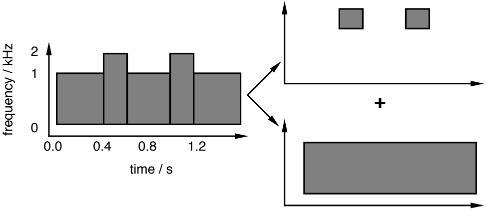
- Elements + attributes



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

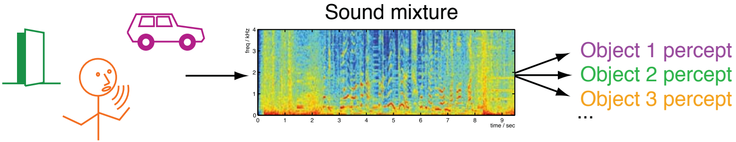
The effect of context

- 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

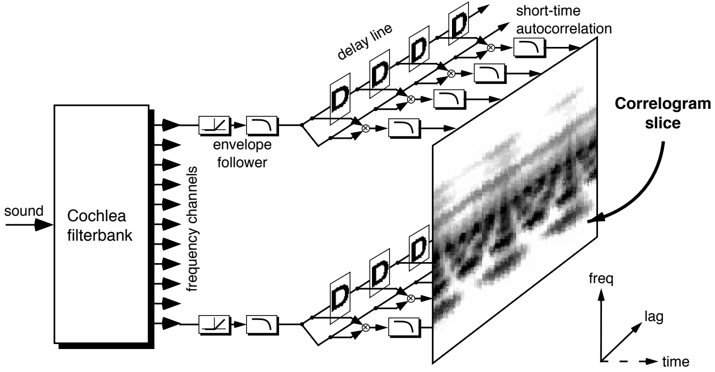
  - to improve speech recognition

- Approach
  - psychoacoustics describes grouping 'rules'

  ... just implement them?

CASA front-end processing

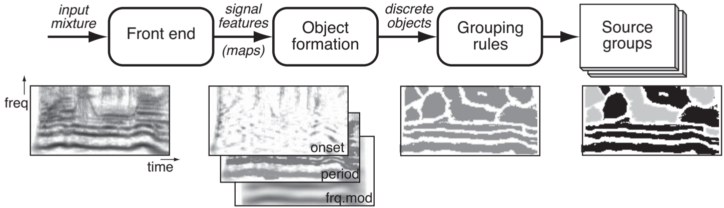
- Correlogram: Loosely based on known/possible physiology



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

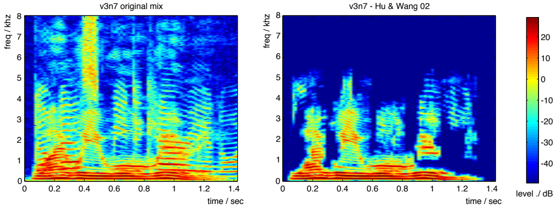
Bottom-Up Approach (Brown and Cooke, 1994)

- Implement psychoacoustic theory

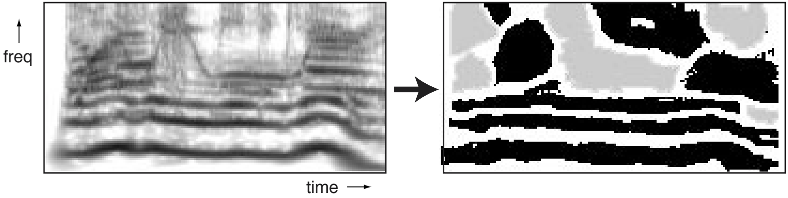


  - left-to-right processing
  - uses common onset & periodicity cues

- Able to extract voiced speech



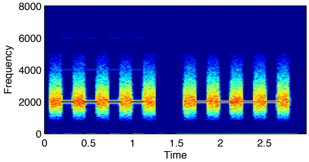
Problems with ‘bottom-up’ CASA

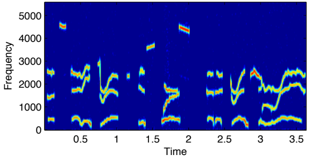


- Circumscribing time-frequency elements
  - need to have ‘regions’, but hard to find
- Periodicity is the primary cue
  - how to handle aperiodic energy?
- Resynthesis via masked filtering
  - cannot separate within a single t-f element
- Bottom-up leaves no ambiguity or context
  - how to model illusions?

Restoration in sound perception

- Auditory 'illusions' = hearing what's not there
- The continuity illusion & Sinewave Speech (SWS)

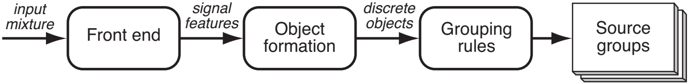




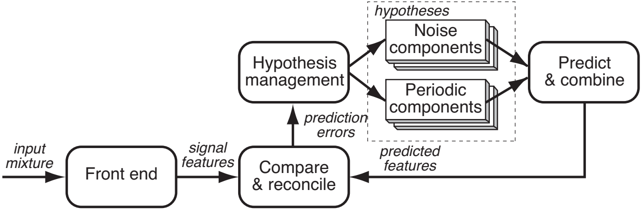
  - duplex perception
- What kind of model accounts for this?
  - is it an important part of hearing?

Adding top-down constraints: Prediction-Driven CASA

- Perception is not direct
  but a search for plausible hypotheses -
- Data-driven (bottom-up)...



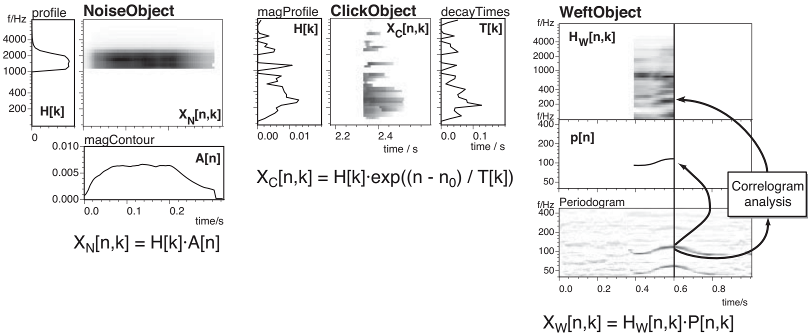
  - objects irresistibly appear
- vs. Prediction-driven (top-down)



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

Generic sound elements for PDCASA (Ellis, 1996)

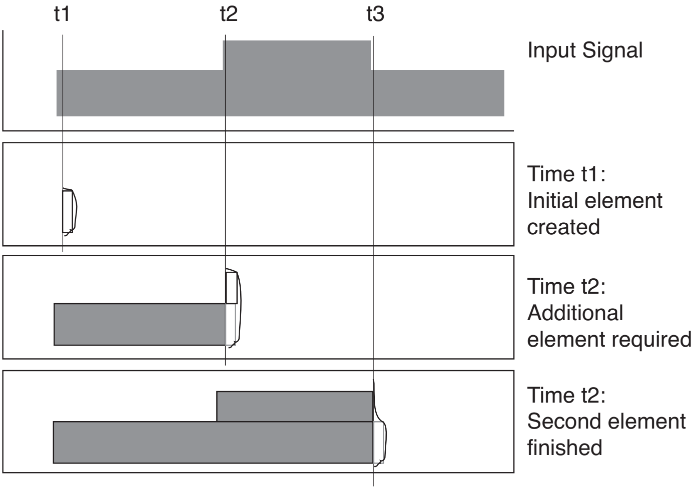
- Goal is a representational space that
  - covers real-world perceptual sounds
  - minimal parameterization (sparseness)
  - separate attributes in separate parameters



- Object hierarchies built on top. . .

PDCASA for old-plus-new

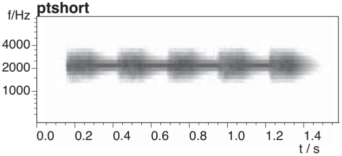
- Incremental analysis



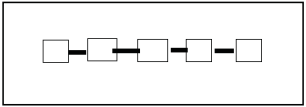
PDCASA for the continuity illusion

- Subjects hear the tone as continuous

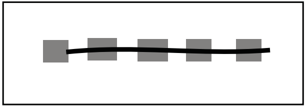
  ... if the noise is a plausible masker



- Data-driven analysis gives just visible portions:

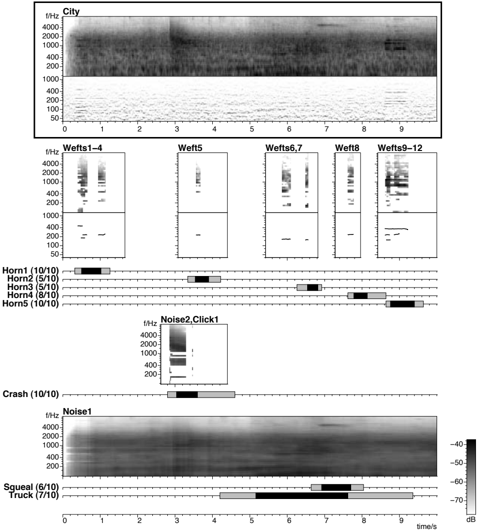


- Prediction-driven can infer masking:



Prediction-Driven CASA

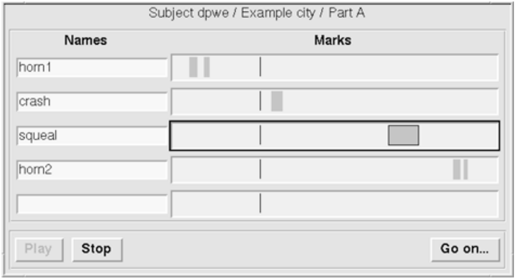
- Explain a complex sound with basic elements



Aside: Ground Truth

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

→ Listening tests to collect 'perceived events':



Aside: Evaluation

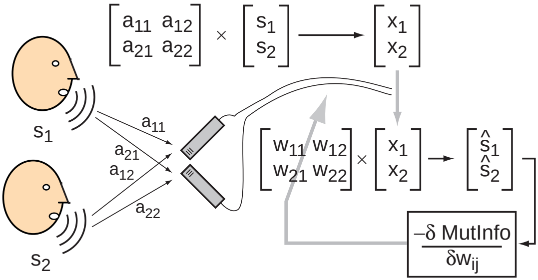
- Evaluation is a big problem for CASA
  - what is the goal, really?
  - what is a good test domain?
  - how do you measure performance?
- SNR improvement
  - tricky to derive from before/after signals: correspondence problem
  - can do with fixed filtering mask
  - differentiate removing signal from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

Outline

1. Sound mixture organization
2. Computational auditory scene analysis
3. Independent component analysis
4. Model-based separation

Independent Component Analysis (ICA) (Bell and 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

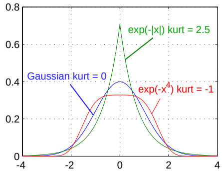
- A signal can be characterized by its PDF $p(x)$

  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 (central limit theorem)

- Measures of deviations from Gaussianity: 4th moment is Kurtosis ("bulging")

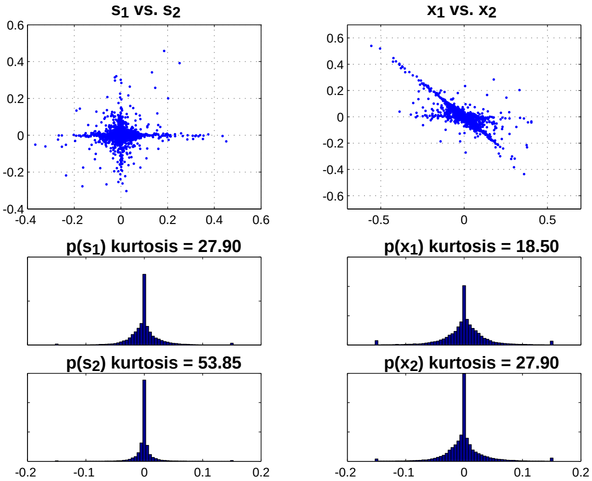
$$\mathrm{kurt}(y) = \mathrm{E}\left[\left(\frac{y-\mu}{\sigma}\right)^4\right] - 3$$



  - kurtosis of Gaussian is zero (this def.)
  - 'heavy tails' $\rightarrow$ kurt > 0
  - closer to uniform dist. $\rightarrow$ kurt < 0
  - Directly related to KL divergence from Gaussian PDF

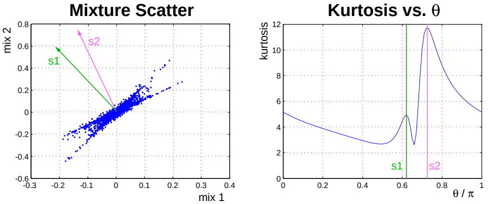
Independence in Mixtures

- Scatter plots & Kurtosis values



Finding Independent Components

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



- Solve by Gradient descent or Newton-Raphson:

$$w^+ = \mathrm{E}[xg(w^T x)] - \mathrm{E}[g'(w^T x)]w$$

$$w = \frac{w^+}{\|w^+\|}$$

- "Fast ICA", (Hyvärinen and Oja, 2000)
  `http://www.cis.hut.fi/projects/ica/fastica/`

Limitations of ICA

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

$$x_1(t) = a_{11}(t) * s_1(t) + a_{12}(t) * s_2(t)$$
$$\Rightarrow X_1(\omega) = A_{11}(\omega)S_1(\omega) + A_{12}(\omega)S_2(\omega)$$

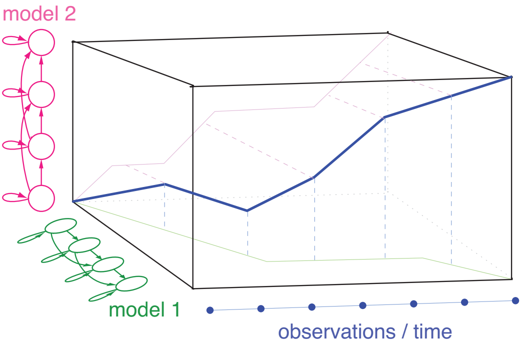
  - Solve ω subbands separately, match up answers
- Searching for best possible inverse matrix
  - cannot find more than N outputs from N inputs
  - but: "projection pursuit" ideas + time-frequency masking...
- Cancellation inherently fragile
  - $\hat{s}_1 = w_{11}x_1 + w_{12}x_2$ to cancel out s_2
  - 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

Model-Based Separation: HMM decomposition

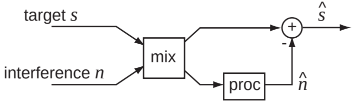
- (*e.g.* Varga and Moore, 1990; Gales and Young, 1993)
- Independent state sequences for 2+ component source models



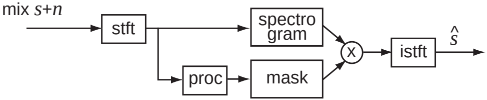
- New combined state space $q' = q1 \times q2$
  - need pdfs for combinations $p(X \mid q_1, q_2)$

One-channel Separation: Masked Filtering

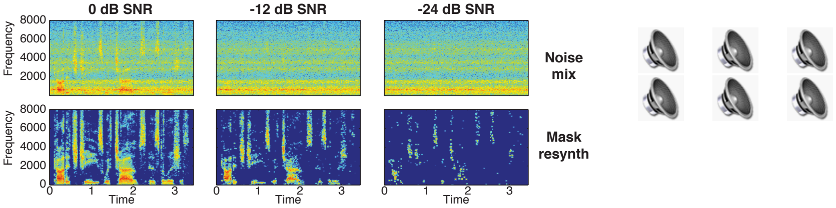
- Multichannel → ICA: Inverse filter and cancel



- One channel: find a time-frequency mask

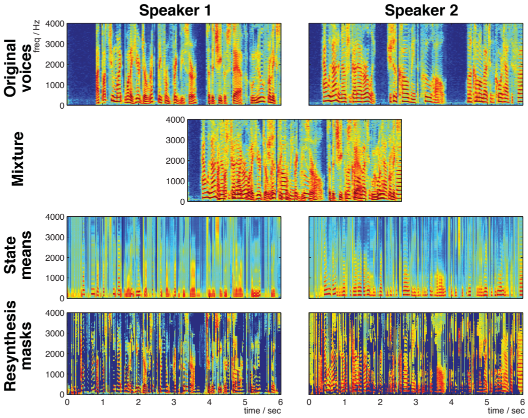


- Cannot remove overlapping noise in t-f cells, but surprisingly effective (psych masking?):



"One microphone source separation"

- (Roweis, 2001)
- State sequences → t-f estimates → mask



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

Speech Fragment Recognition

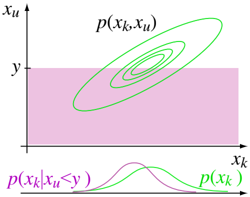
- (Barker et al., 2005)
- Signal separation is too hard! Instead:
  - segregate features into partially-observed sources
  - then classify
- Made possible by missing data recognition
  - integrate over uncertainty in observations for true posterior distribution
- Goal: Relate clean speech models $P(X \mid M)$ to speech-plus-noise mixture observations

  . . . and make it tractable

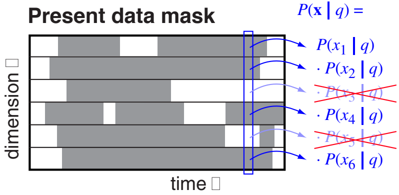
Missing Data Recognition

- Speech models $p(x \mid m)$ are multidimensional...
  i.e. means, variances for every freq. channel
  - need values for all dimensions to get $p(\cdot)$
- But: can evaluate over a subset of dimensions x_k

$$p(x_k \mid m) = \int p(x_k, x_u \mid m)\, dx_u$$



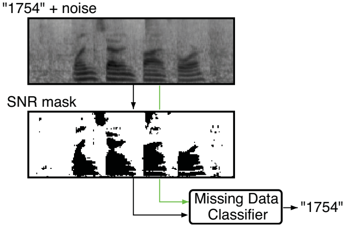
- Hence, missing data recognition:

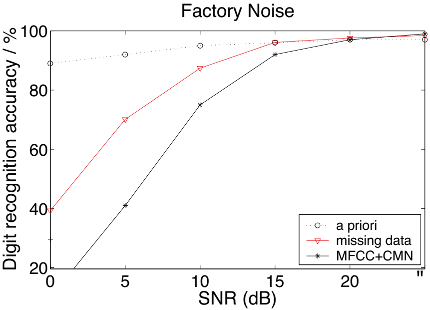


  - hard part is finding the mask (segregation)

Missing Data Results

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





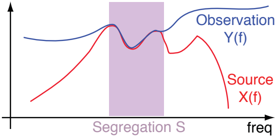
  - must use spectral features!
- But: nonstationary noise → spurious mask bits
  - can we try removing parts of mask?

Comparing different segregations

- Standard classification chooses between models M to match source features X

$$M^* = \underset{M}{\operatorname{argmax}}\, p(M \mid X) = \underset{M}{\operatorname{argmax}}\, p(X \mid M) p(M)$$

- Mixtures: observed features Y, segregation S, all related by $p(X \mid Y, S)$



- Joint classification of model and segregation:

$$p(M, S \mid Y) = p(M) \int p(X \mid M) \frac{p(X \mid Y, S)}{p(X)}\, dX\, p(S \mid Y)$$

  - $P(X)$ no longer constant

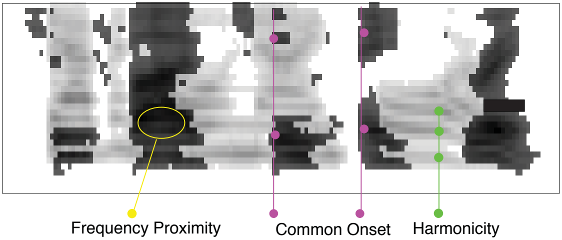
Calculating fragment matches

$$p(M, S \mid Y) = p(M) \int p(X \mid M) \frac{p(X \mid Y, S)}{p(X)} \, dX \, p(S \mid Y)$$

- $p(X \mid M)$ – the clean-signal feature model
- $\frac{p(X \mid Y,S)}{p(X)}$ – 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 Ss. . .
- Result:
  - probabilistically-correct relation between
  - clean-source models $p(X \mid M)$ and
  - inferred, recognized source + segregation $p(M, S \mid Y)$

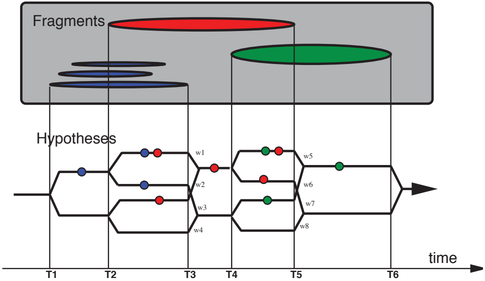
Using CASA features

- $p(S \mid Y)$ 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 $p(S \mid Y)p(X \mid M)$

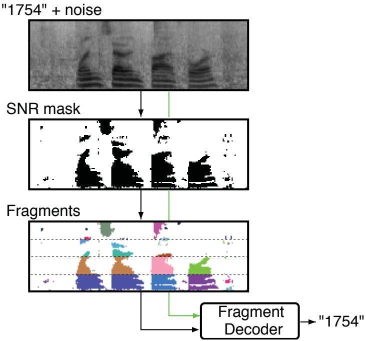
  i.e. best combination of segregation and match to speech models

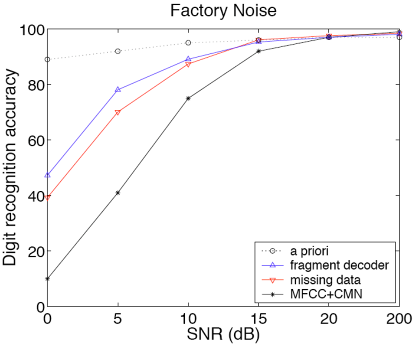
- Merging hypotheses limits space demands

  ... but erases specific history

Speech fragment decoder results

- Simple $p(S \mid Y)$ model forces contiguous regions to stay together
  - big efficiency gain when searching S space

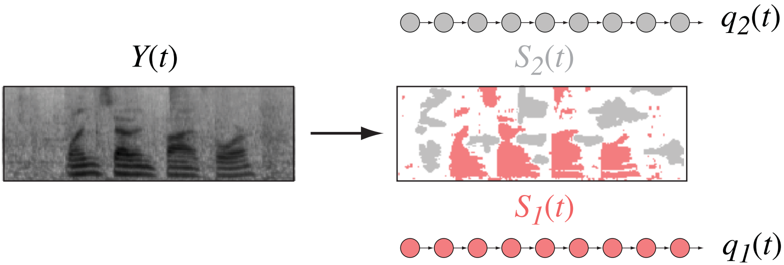




- Clean-models-based recognition rivals trained-in-noise recognition

Multi-source decoding

- Search for more than one source



- Mutually-dependent data masks
  - disjoint subsets of cells for each source
  - each model match $p(M_x \mid S_x, Y)$ is independent
  - masks are mutually dependent: $p(S_1, S_2 \mid Y)$
- 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

Parting thought

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

References

Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 1990. ISBN 0262521954.

C.J. Darwin and R.P. Carlyon. Auditory grouping. In B.C.J. Moore, editor, *The Handbook of Perception and Cognition, Vol 6, Hearing*, pages 387–424. Academic Press, 1995.

G. J. Brown and M. P. Cooke. Computational auditory scene analysis. *Computer speech and language*, 8:297–336, 1994.

D. P. W. Ellis. *Prediction–driven computational auditory scene analysis*. PhD thesis, Department of Electrtical Engineering and Computer Science, M.I.T., 1996.

Anthony J. Bell and Terrence J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995. `ftp://ftp.cnl.salk.edu/pub/tony/bell.blind.ps.Z`.

A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4–5):411–430, 2000. `http://www.cis.hut.fi/aapo/papers/IJCNN99_tutorialweb/`.

A. Varga and R. Moore. Hidden markov model decomposition of speech and noise. In *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 845–848, 1990.

M.J.F. Gales and S.J. Young. Hmm recognition in noise using parallel model combination. In *Proc. Eurospeech-93*, volume 2, pages 837–840, 1993.

S. Roweis. One-microphone source separation. In *NIPS*, volume 11, pages 609–616. MIT Press, Cambridge MA, 2001.

Jon Barker, Martin Cooke, and Daniel P. W. Ellis. Decoding speech in the presence of other sources. *Speech Communication*, 45(1):5–25, 2005. URL `http://www.ee.columbia.edu/~dpwe/pubs/BarkCE05-sfd.pdf`.