Computational Auditory Scene Analysis

Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Engineering, Columbia Univ., NY USA

dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

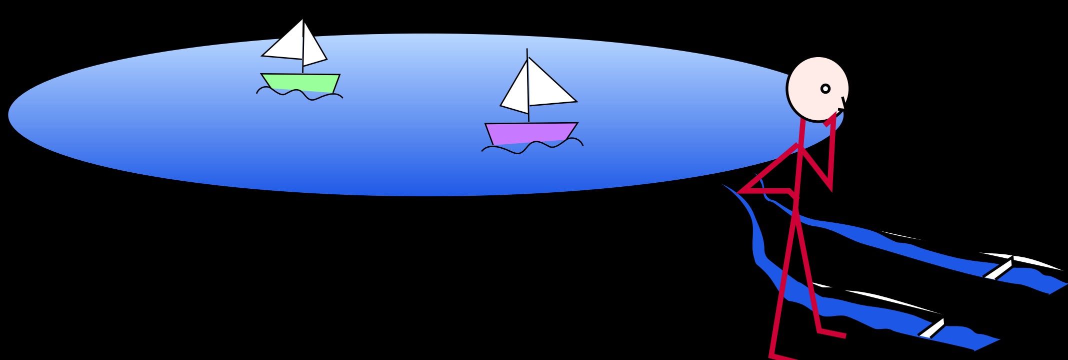
1. ASA and CASA
2. The Development of CASA
3. The Prospects for Computational Audition

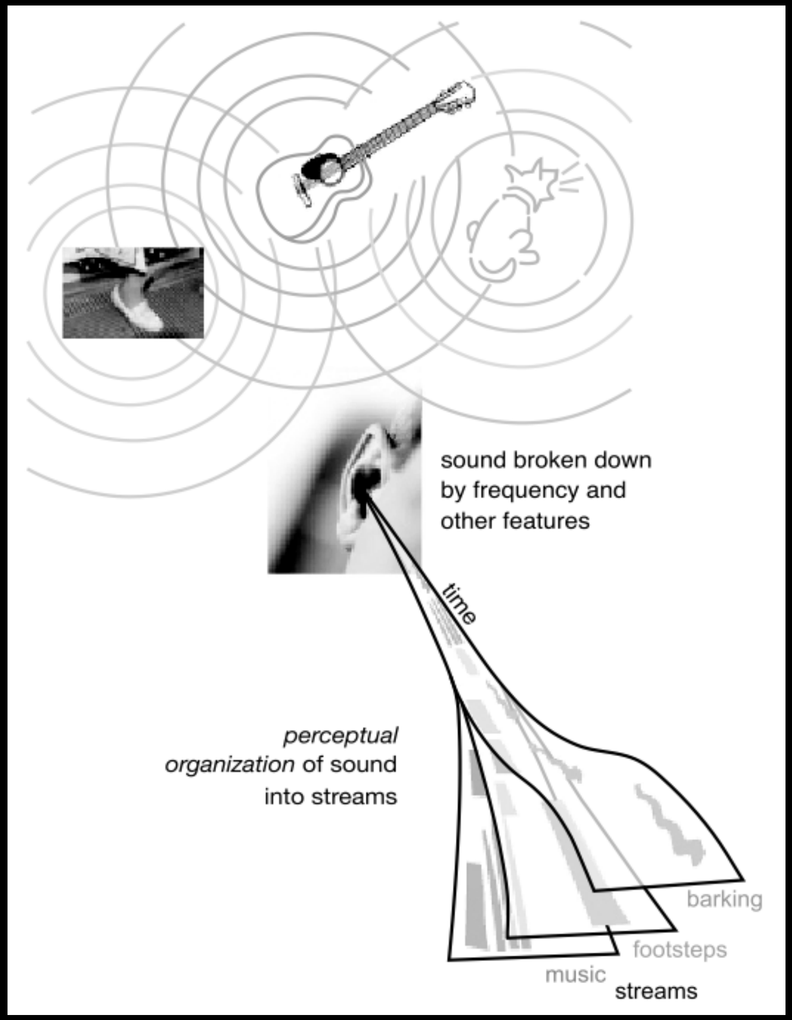
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

1. Auditory Scene Analysis (ASA)

"To recognize the component sounds that have been added together to form the mixture that reaches our ears, the auditory system must somehow create individual descriptions that are based only on those components of the sound that have arisen from the same environmental event."

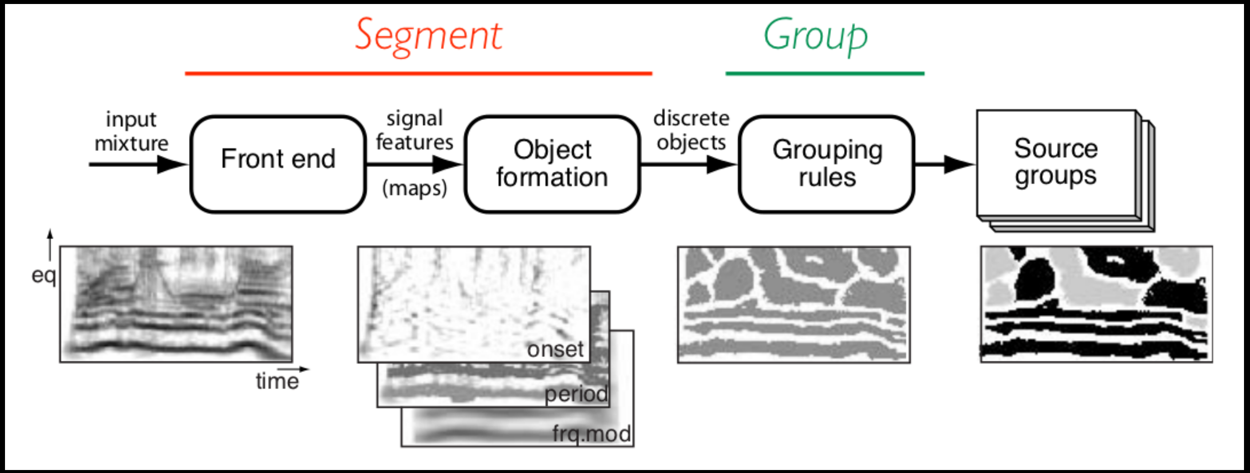




Cusack & Carlyon 2004

What is CASA?

- **Computer systems for separating sounds**
  - based on biological "inspiration" (ASA)
  - based on a source / stream formation paradigm
  - frequently using pitch information (less binaural)
  - frequently involving time-frequency masking

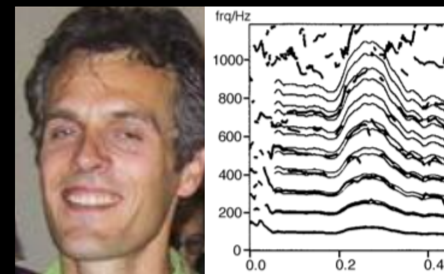


after Brown 1992

"If the study of human audition were able to lay bare the principles that govern the human skill, there is some hope that a computer could be designed to mimic it."

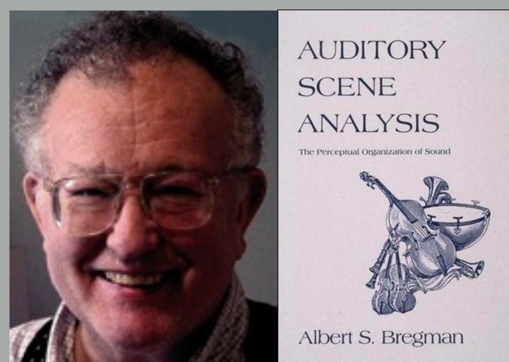
CASA History

Lyon 1984

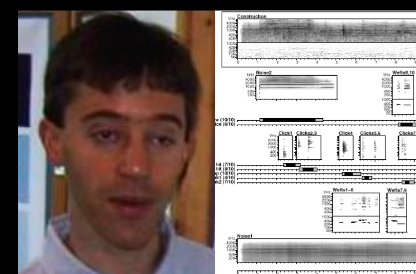


Cooke 1991

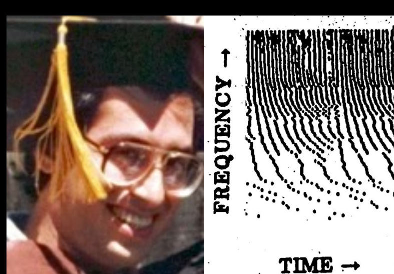
Wang & Brown 2006



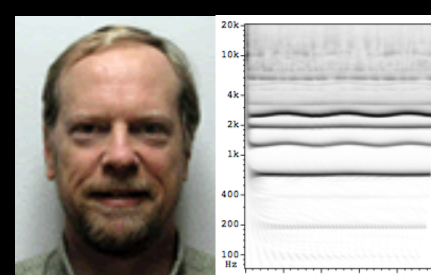
Bregman 1990



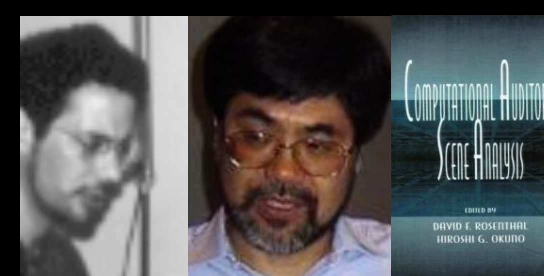
Ellis 1996



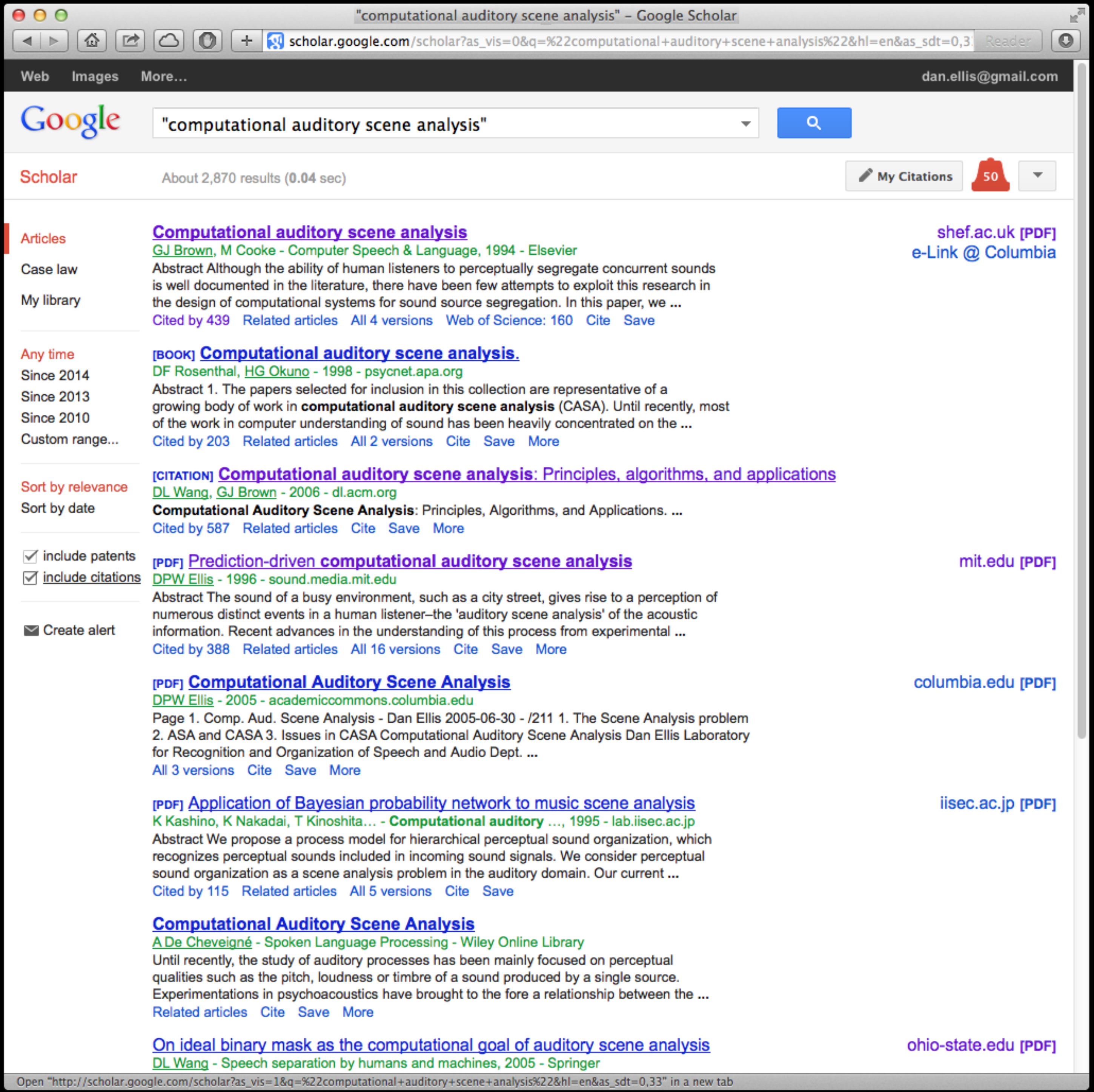
Weintraub 1985



Mellinger 1991

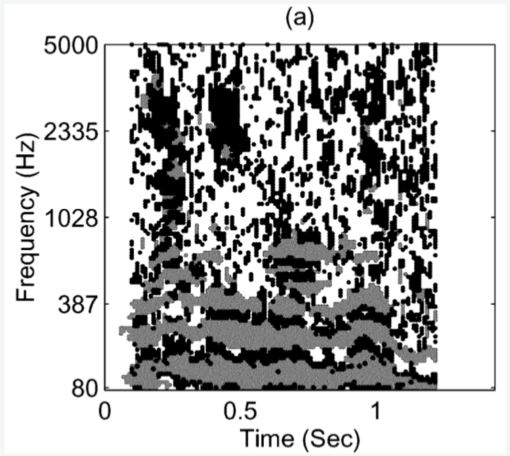


Rosenthal & Okuno 1998


"computational auditory scene analysis" – Google Scholar
scholar.google.com/scholar?as_vis=0&q=%22computational+auditory+scene+analysis%22&hl=en&as_sdt=0,3
Reader
Web Images More…
dan.ellis@gmail.com
Google
"computational auditory scene analysis"
Scholar
About 2,870 results (0.04 sec)
My Citations
50
Articles
Case law
My library
Any time
Since 2014
Since 2013
Since 2010
Custom range...
Sort by relevance
Sort by date
include patents
include citations
Create alert
Computational auditory scene analysis
GJ Brown, M Cooke - Computer Speech & Language, 1994 - Elsevier
Abstract Although the ability of human listeners to perceptually segregate concurrent sounds is well documented in the literature, there have been few attempts to exploit this research in the design of computational systems for sound source segregation. In this paper, we ...
Cited by 439 Related articles All 4 versions Web of Science: 160 Cite Save
shef.ac.uk [PDF]
e-Link @ Columbia
[BOOK] Computational auditory scene analysis.
DF Rosenthal, HG Okuno - 1998 - psycnet.apa.org
Abstract 1. The papers selected for inclusion in this collection are representative of a growing body of work in computational auditory scene analysis (CASA). Until recently, most of the work in computer understanding of sound has been heavily concentrated on the ...
Cited by 203 Related articles All 2 versions Cite Save More
[CITATION] Computational auditory scene analysis: Principles, algorithms, and applications
DL Wang, GJ Brown - 2006 - dl.acm.org
Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. ...
Cited by 587 Related articles Cite Save More
[PDF] Prediction-driven computational auditory scene analysis
DPW Ellis - 1996 - sound.media.mit.edu
Abstract The sound of a busy environment, such as a city street, gives rise to a perception of numerous distinct events in a human listener–the 'auditory scene analysis' of the acoustic information. Recent advances in the understanding of this process from experimental ...
Cited by 388 Related articles All 16 versions Cite Save More
mit.edu [PDF]
[PDF] Computational Auditory Scene Analysis
DPW Ellis - 2005 - academiccommons.columbia.edu
Page 1. Comp. Aud. Scene Analysis - Dan Ellis 2005-06-30 - /211 1. The Scene Analysis problem 2. ASA and CASA 3. Issues in CASA Computational Auditory Scene Analysis Dan Ellis Laboratory for Recognition and Organization of Speech and Audio Dept. ...
All 3 versions Cite Save More
columbia.edu [PDF]
[PDF] Application of Bayesian probability network to music scene analysis
K Kashino, K Nakadai, T Kinoshita... - Computational auditory ..., 1995 - lab.iisec.ac.jp
Abstract We propose a process model for hierarchical perceptual sound organization, which recognizes perceptual sounds included in incoming sound signals. We consider perceptual sound organization as a scene analysis problem in the auditory domain. Our current ...
Cited by 115 Related articles All 5 versions Cite Save
iisec.ac.jp [PDF]
Computational Auditory Scene Analysis
A De Cheveigné - Spoken Language Processing - Wiley Online Library
Until recently, the study of auditory processes has been mainly focused on perceptual qualities such as the pitch, loudness or timbre of a sound produced by a single source. Experimentations in psychoacoustics have brought to the fore a relationship between the ...
Related articles Cite Save More
On ideal binary mask as the computational goal of auditory scene analysis
DL Wang - Speech separation by humans and machines, 2005 - Springer
ohio-state.edu [PDF]
Open "http://scholar.google.com/scholar?as_vis=1&q=%22computational+auditory+scene+analysis%22&hl=en&as_sdt=0,33" in a new tab

2. CASA Systems

- Literal implementations of the process described in Bregman 1990:
  - compute "regularity" cues:
    - common onset
    - gradual change
    - harmonic patterns
    - common fate

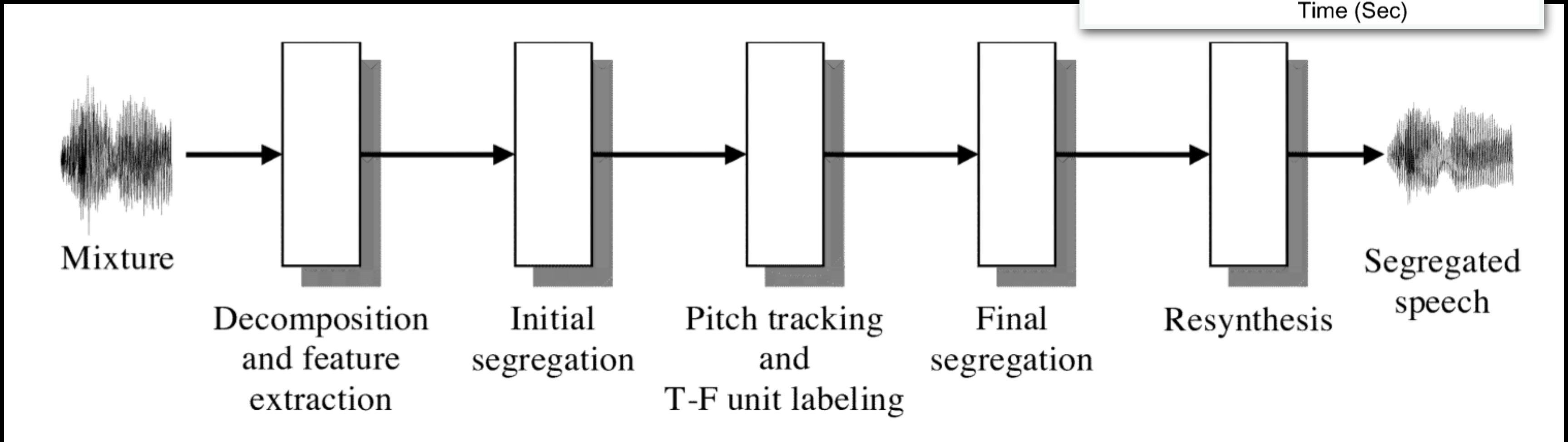


Original v3n7

Brown 1992

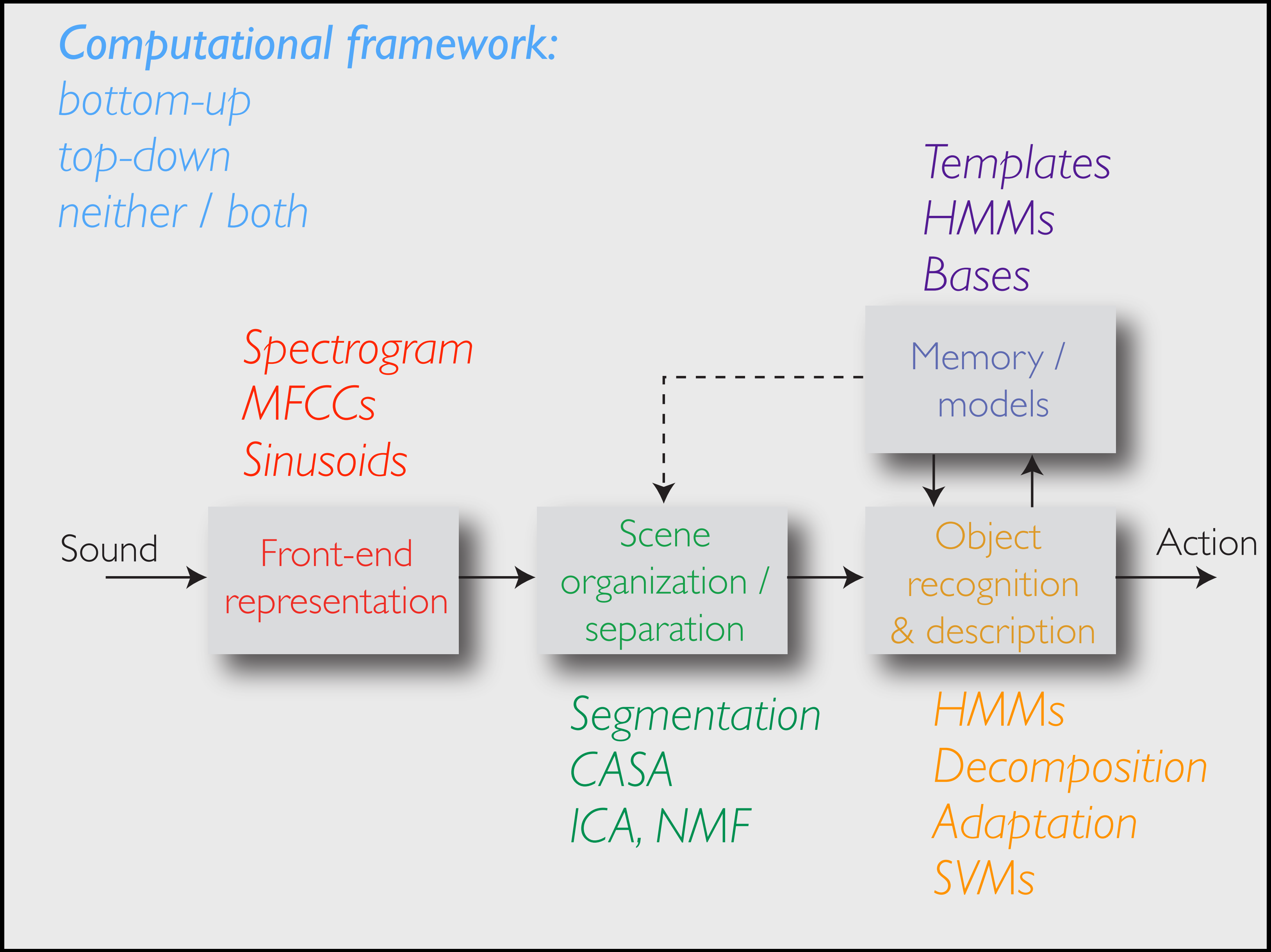
Ellis 1996

Hu & Wang 2004



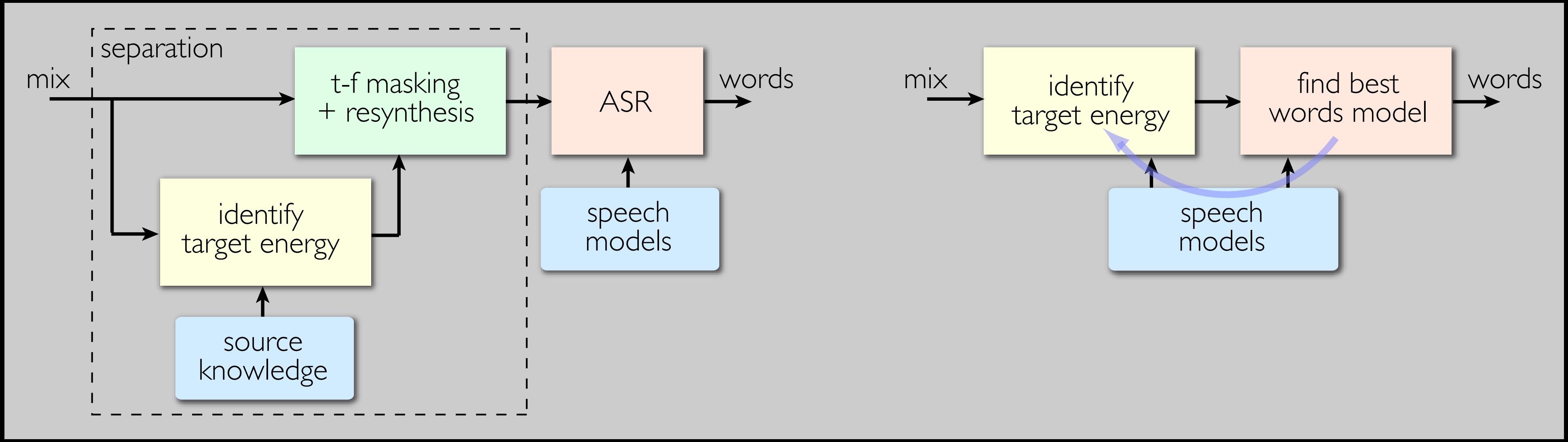
Hu & Wang 2004

Key CASA Components



How Important is Separation?

- **Separation systems often evaluated by SNR**
  - i.e., comparison to pre-mix components
  - is this relevant?
- **Best systems use resynthesis**
  - e.g. IBM's Superhuman Speech Recognizer
    - "separate then recognize"



- **Separating original signals is not necessary**

Machine Listening Tasks

- What is the goal? How to evaluate?

Task	Environmental Sound	Speech	Music
Describe	Automatic Narration	Emotion	Music Recommendation
Classify	Environment Awareness	ASR	Music Transcription
Dectect	"Sound Intelligence"	VAD	Speech/Music

Domain

Sound Separation Techniques

- Marr (1982): Levels of a perceptual problem:

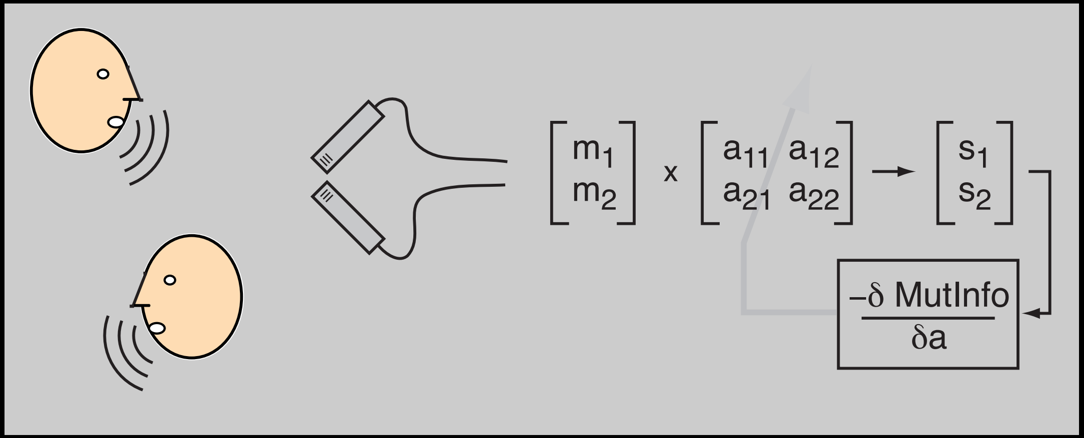
Computational Theory	Properties of the world that make the problem solvable
Algorithm	Specific calculations & operations
Implementation	Details of how it's done

- What is ASA's "computational theory"?
  - Environmental regularities → CASA
  - Independence → ICA
  - Efficient / sparse description → NMF
  - Underlying explanation → MaxVQ, Factorial HMM

Independent Component Analysis

Bell & Sejnowski '95
Smaragdis '98

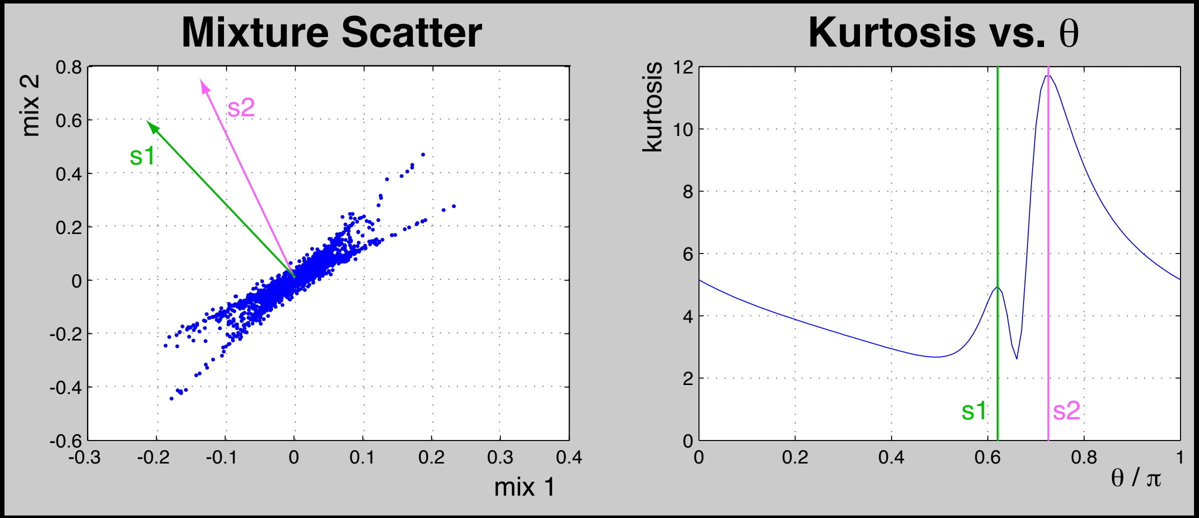
- Separate "blind" combinations by maximizing independence of outputs:



- e.g. Kurtosis

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

as a measure of independence



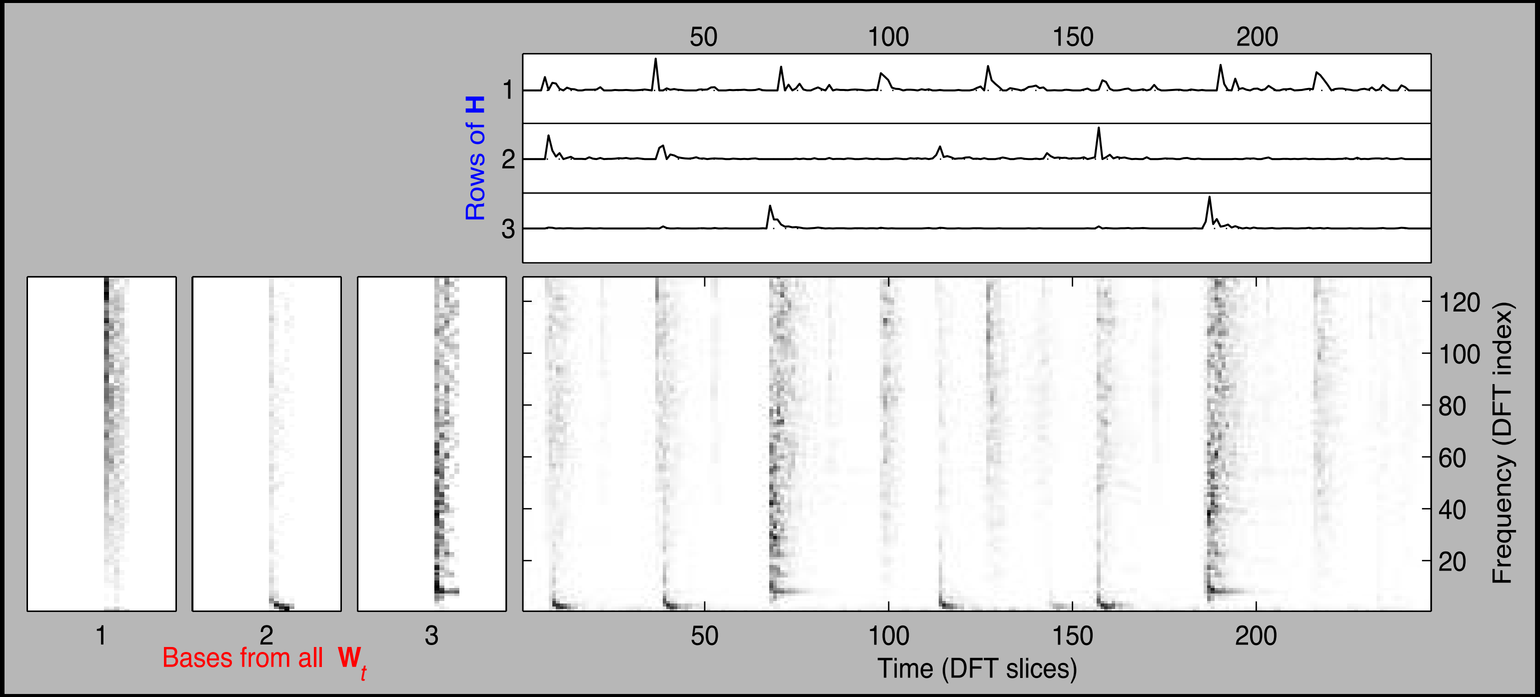
Nonnegative Matrix Factorization

Lee & Seung '99
Smaragdis & Brown '03
Abdallah & Plumbley '04
Virtanen '07

- Decomposition of spectrograms into templates + activation

$$X = W \cdot H$$

  - fits neatly with time-frequency masking
  - useful for repeated events e.g. music

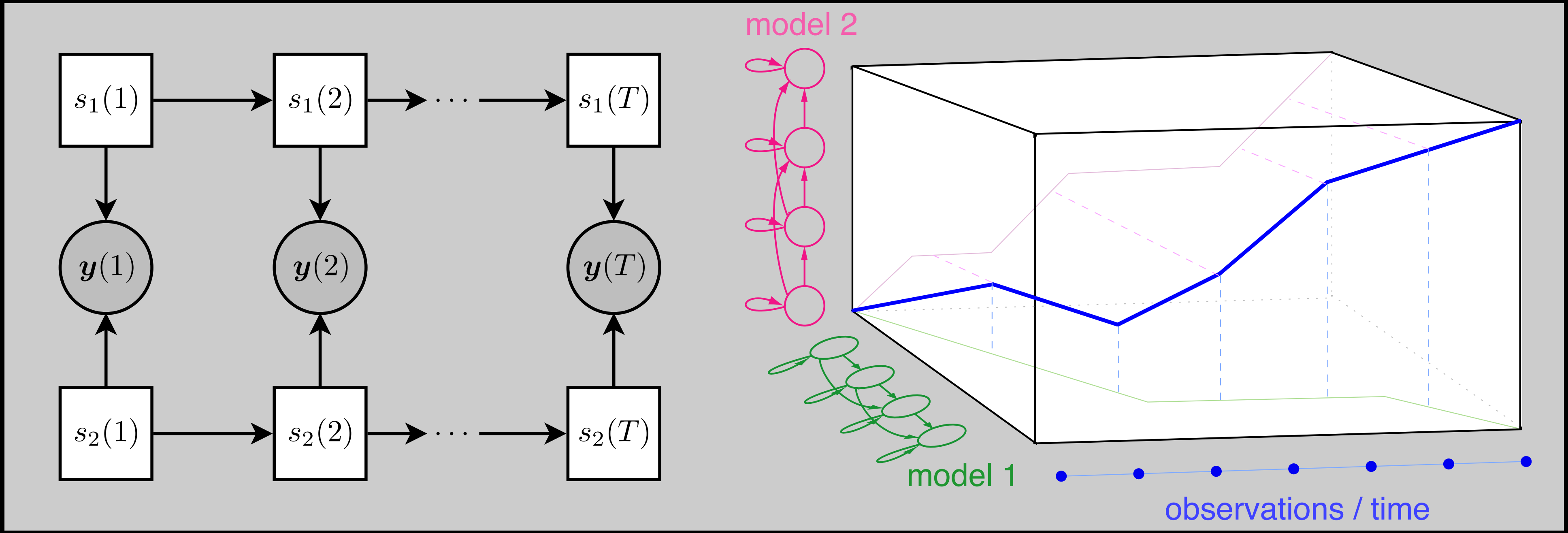


Model-Based Explanation

- Probabilistic approach:
  Find most likely parameters
  of some model

Varga & Moore '90
Gales & Young '95
Ghahramani & Jordan '97
Roweis '01
Kristjansson et al '06
Hershey et al '10

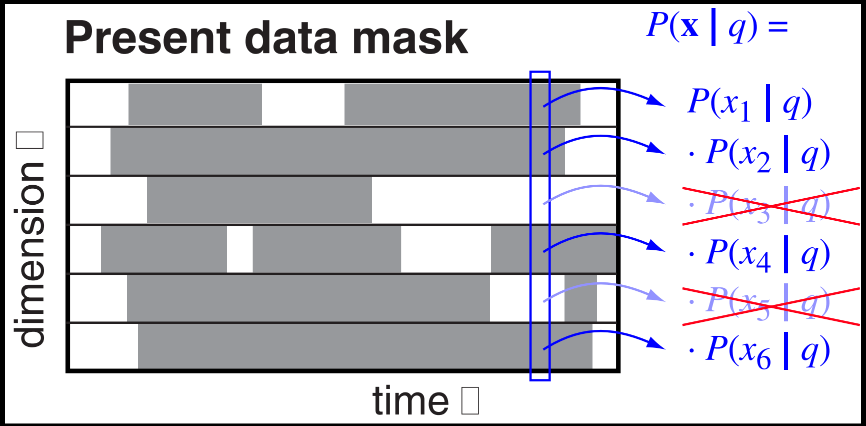
$$\{s_i(t)\}^* = \arg\max_{\{s_i(t)\}} \Pr(y(t)|\{s_i(t)\}) \cdot \Pr(\{s_i(t)\})$$



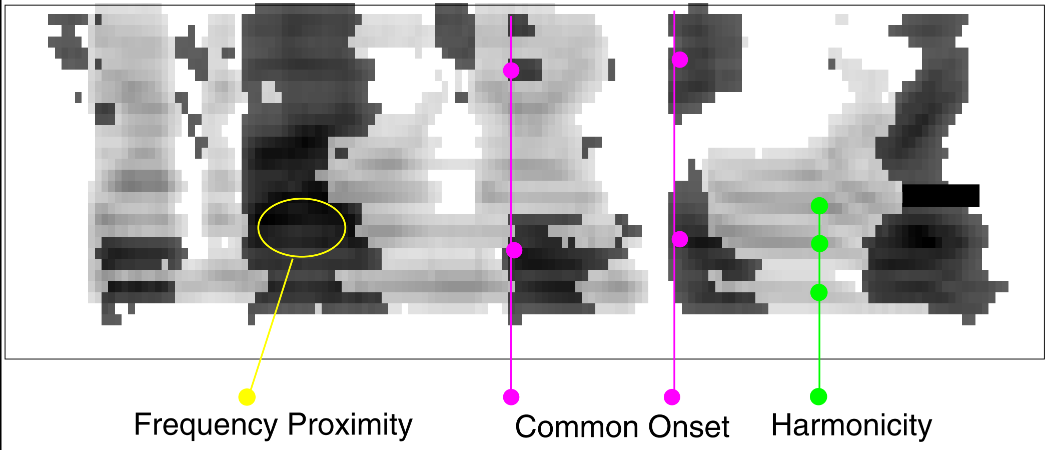
Missing Data Recognition

Cooke et al. '01
Barker et al. '05

- Integrate out missing information needed to recognizing a source…
  - no need to estimate missing/masked values



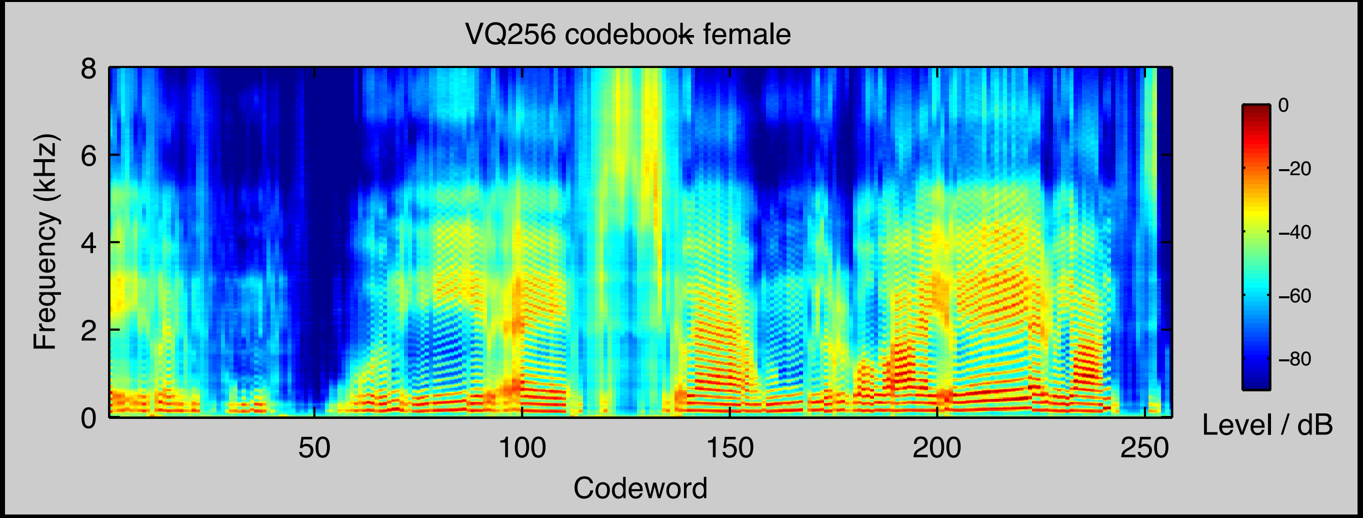
- Joint search for model M and segregation S
  - use CASA as prior on segregations



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

3. Whither ASA?

- Dictionary models can learn harmonicity, onset, etc.

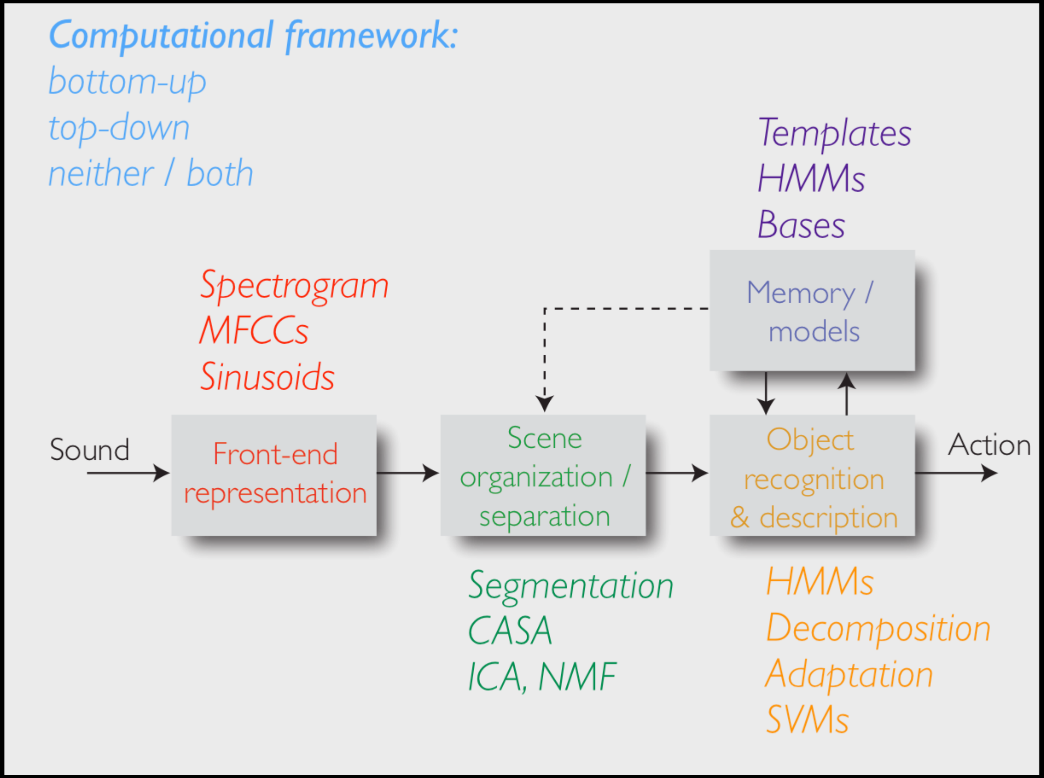


- secondary effects (harmony)
- subsume the ideas of CASA?

- Can also capture sequential structure
  - e.g., consonants follow vowels ("schema")
  - use overlapping patches?
- Computational theory or implementation?

Future CASA Systems

- **Representation**
  - still missing the key basis of fusion?
- **Models, Separation**
  - learn from examples
- **Object description**
  - what is salient to listeners? what is attention?
- **Computational framework**
  - pragmatic search for solution (illusions)



Summary

- Auditory Scene Analysis
  - the functional problem of hearing
- Computational Auditory Scene Analysis
  - computer implementations
- Automatic Sound Source Separation
  - different problems, different solutions

"We have reached the point where we have a good appreciation of many of the kinds of evidence that the human brain uses for partitioning sound, and it seems appropriate to begin to explore the formal patterns of computation by which the process could be accomplished." Bregman, 1990

References

[Abdallah & Plumbley 2004] S. Abdallah & M. Plumbley, ”Polyphonic transcription by non-negative sparse coding of power spectra”, *Proc. Int. Symp. Music Info. Retrieval* 2004

[Barker et al. 2005] J. Barker, M. Cooke, D. Ellis, “Decoding speech in the presence of other sources,” Speech Comm. 45, 5-25, 2005.

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

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

[Brown & Wang 2006] G. Brown & D. Wang (eds), *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications,* Wiley, 2006

[Brown 1992] G. Brown, *Computational auditory scene analysis: A representational approach,* Ph.D. Thesis, Univ. Sheffield, 1992.

[Brown & Cooke 1994] G. Brown & M. Cooke, “Computational auditory scene analysis,” Comp. Speech & Lang. 8(4), 1994.

[Cooke 1991] M. Cooke, *Modelling auditory processing and organisation*, Ph.D. thesis, Univ. Sheffied, 1991.

[Cooke et al. 2001] M. Cooke, P. Green, L. Josifovski, A. Vizinho, “Robust automatic speech recognition with missing and uncertain acoustic data,” Speech Communication 34, 267-285, 2001.

[Ellis 1996] D. Ellis, “Prediction-Driven Computational Auditory Scene Analysis,” Ph.D. thesis, MIT EECS, 1996.

[Gales & Young 1995] M. Gales & S. Young, “Robust speech recognition in additive and convolutional noise using parallel model combination,” *Comput. Speech Lang*. 9, 289–307, 1995.

[Ghahramani & Jordan 1997] Z. Ghahramani & M. Jordan, “Factorial hidden Markov models,” *Machine Learning*, 29(2-3,) 245–273, 1997.

[Hershey et al 2010] J. Hershey, S. Rennie, P. Olsen, T. Kristjansson, “Super-human multi-talker speech recognition: A graphical modeling approach,” *Comp. Speech & Lang*., 24, 45–66.

[Hu & Wang 2004] G. Hu and D.L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” IEEE Tr. Neural Networks, 15(5), Sep. 2004.

[Kristjansson et al 2006] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath, “Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system,” *Proc. Interspeech*, 775-1778, 2006.

[Lee & Seung 1999] D. Lee & S. Seung (1999) ”Learning the Parts of Objects by Non-negative Matrix Factorization”, *Nature* 401.

[Lyon 1984] R. Lyon, “Computational models of neural auditory processing”, *IEEE ICASSP*, 36.1.(1– 4), 1984.

[Marr 1982] D. Marr (1982) *Vision*, MIT Press.

[Mellinger 1991] D. Mellinger, *Event formation and separation in musical sound*, Ph.D. thesis, Stanford CCRMA, 1991.

[Rosenthal & Okuno 1998] D. Rosenthal & H. Okuno (eds), *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, 1998

[Roweis 2001] S. Roweis, “One-microphone source separation”, Proc. NIPS 2001.

[Smaragdis & Brown 2003] P. Smaragdis & J. Brown (2003) ”Non-negative Matrix Factorization for Polyphonic Music Transcription”, *Proc. IEEE WASPAA*, 177-180, October 2003

[Smaragdis 1998] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” Intl. Wkshp. on Indep. & Artif.l Neural Networks, Tenerife, Feb. 1998.

[Varga & Moore 1990] A. Varga & R. Moore, “Hidden Markov Model decomposition of speech and noise,” ICASSP, 1990.

[Virtanen 2007] T. Virtanen “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Tr. Audio, Speech, & Lang. Proc.* 15(3), 1066–1074, Mar. 2007.

[Weintraub 1985] M. Weintraub, *A theory and computational model of auditory monaural sound separation*, Ph.D. thesis, EE dept., Stanford Univ.