A History and Overview of Machine Listening

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

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

1. A Machine Listener
2. Key Tools in Machine Listening
3. Outstanding Problems

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

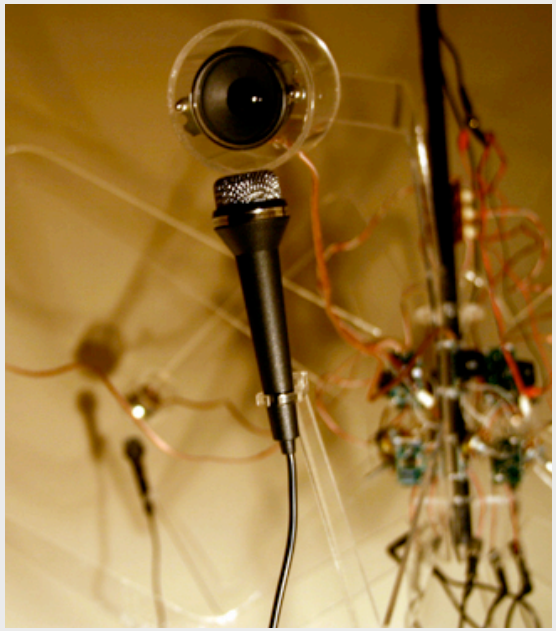
Columbia University
in the City of New York

1. Machine Listening

"Listening puts us in the world" (Handel, 1989)

- Listening
  = Getting useful information from sound
  - signal processing + abstraction
  - "useful" depends on what you're doing

- Machine Listening
  = devices that respond to particular sounds



Listening Machines

- 1922

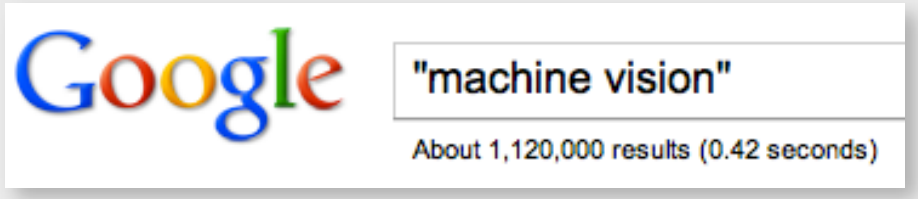
  - magnet releases dog in response to 500 Hz energy

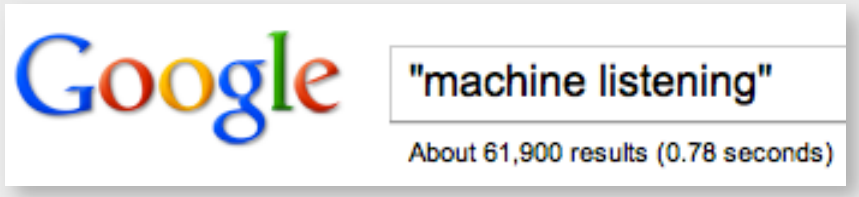
- 1984

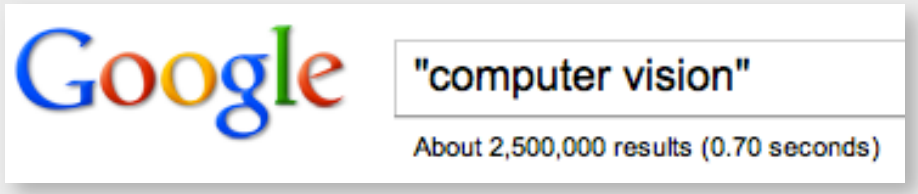
  - two claps toggle power outlet on/off

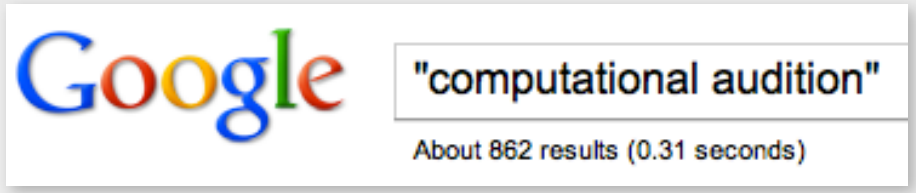
Why is Machine Listening obscure?

- A poor second to Machine Vision:



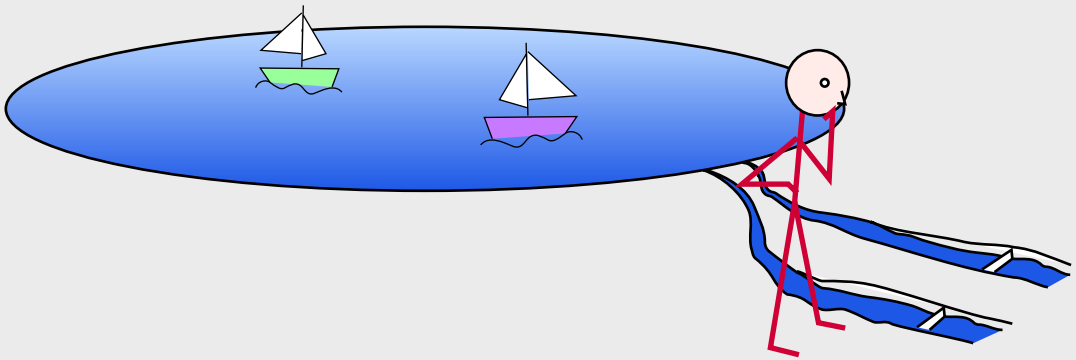






  - vision leads to more immediate practical applications (robotics)?
  - "machine listening" has been subsumed by speech recognition?
  - images are more tangible?

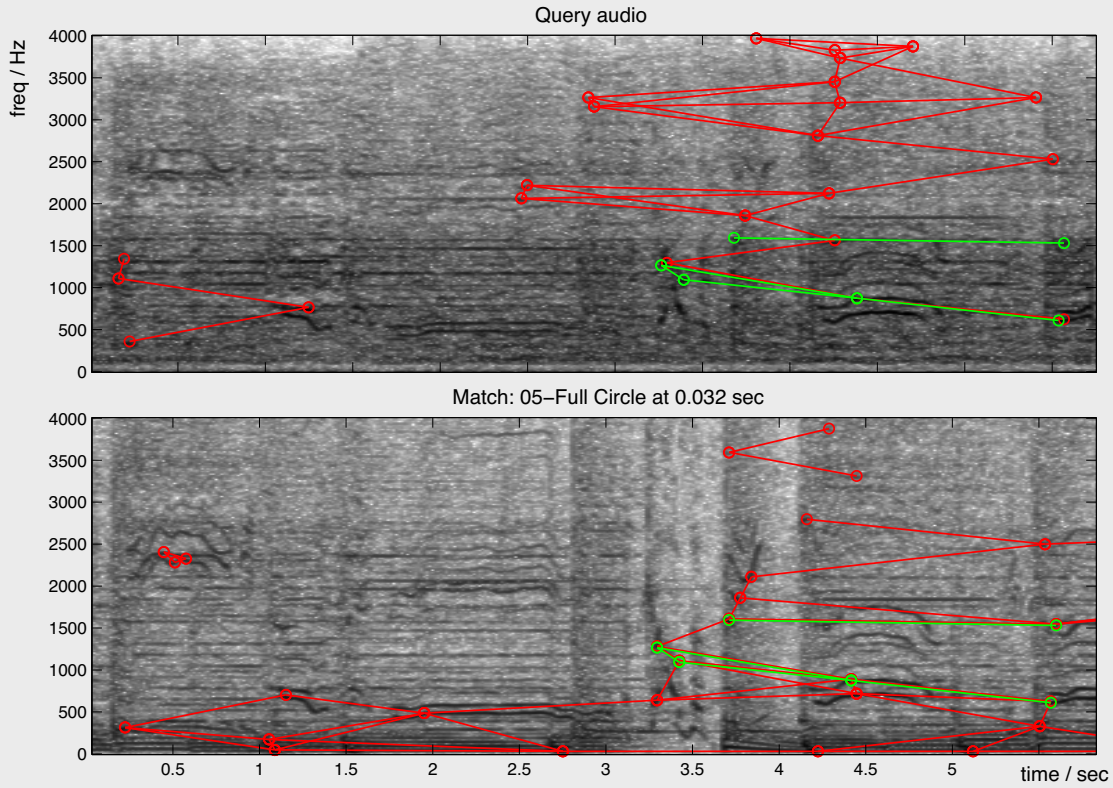
Listening to Mixtures



- The world is cluttered & sound is transparent
  - mixtures are a certainty
- Useful information is structured by 'sources'
  - specific definition of a 'source': intentional independence

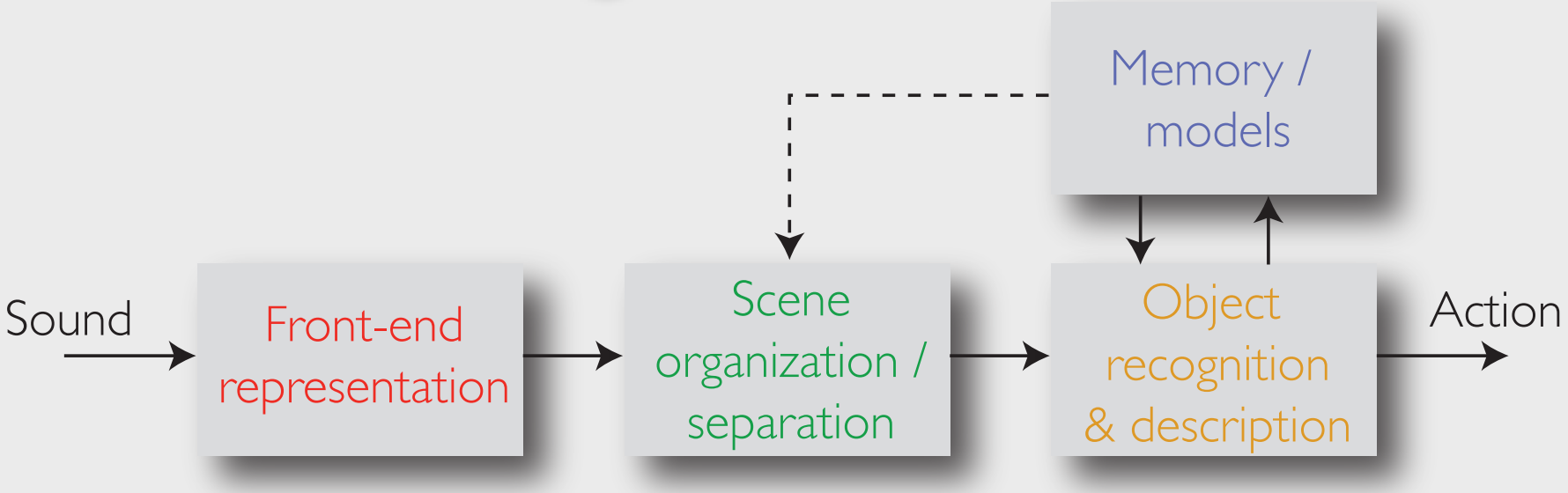
Listening vs. Separation

- Extracting information
  does not require reconstructing waveforms
  - e.g. Shazam fingerprinting [Wang '03]



- But... high-level description permits resynthesis

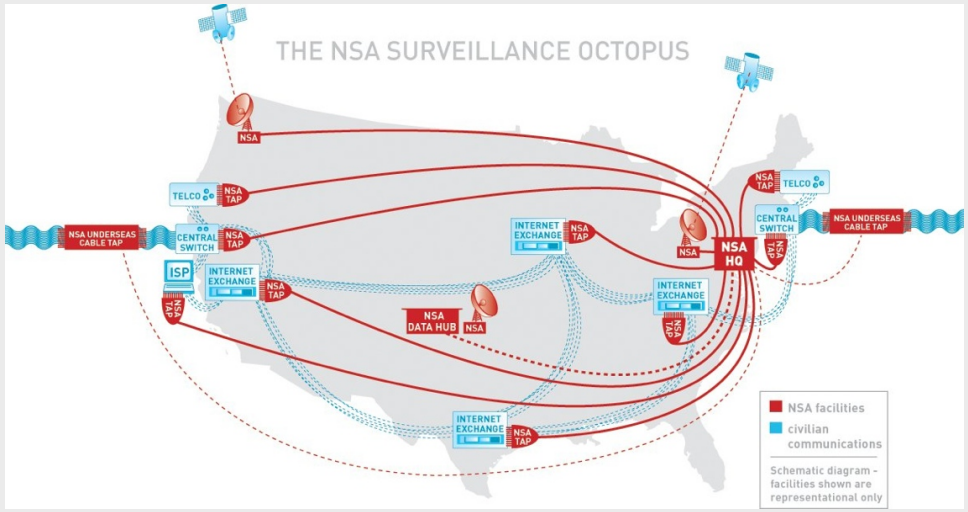
Listening Machine Parts



- Representation: what is perceived
- Organization: handling overlap and interference
- Recognition: object classification & description
- Models: stored information about objects

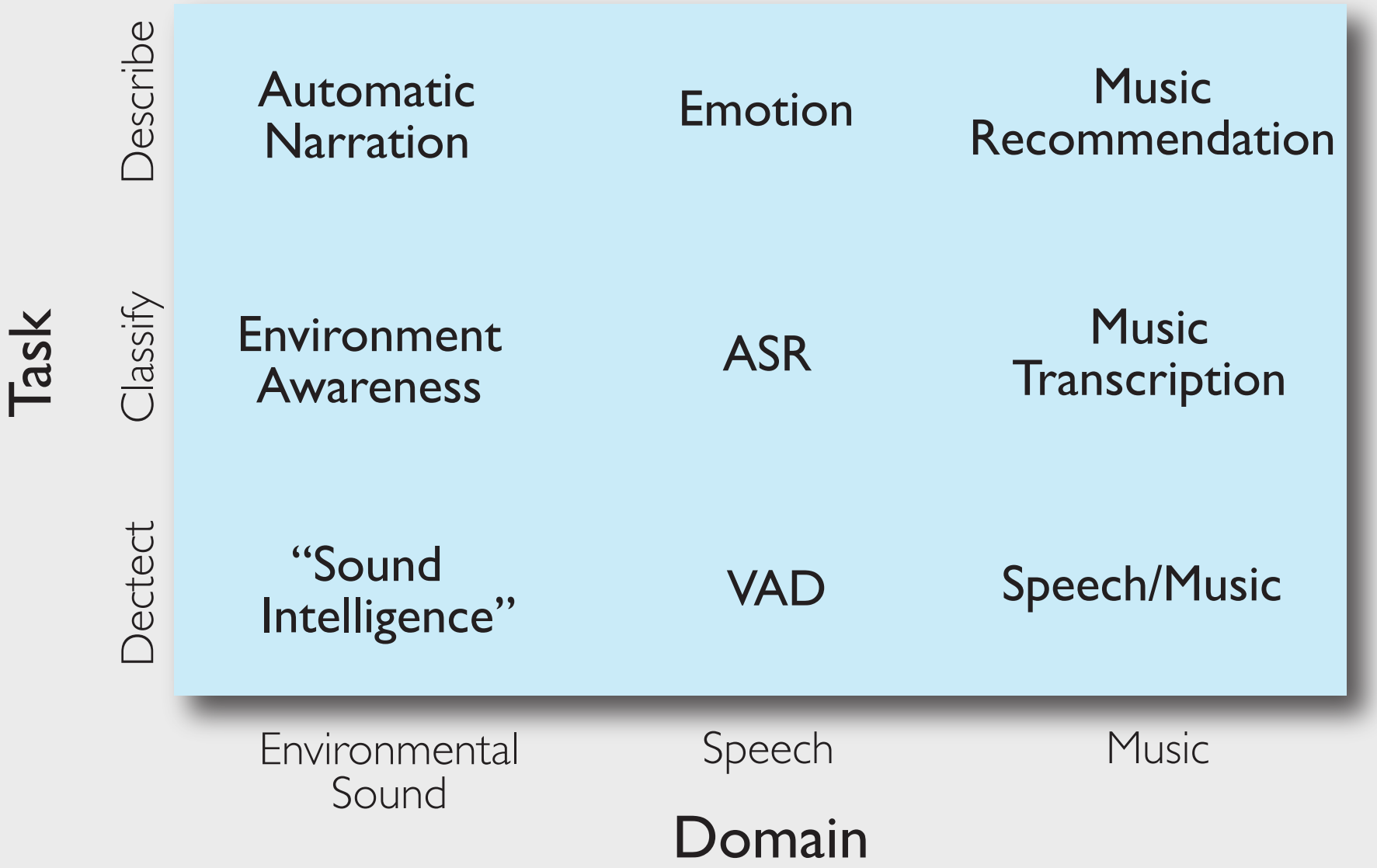
Listening Machines

- What would count as a machine listener, ... and why would we want one?
  - replace people
  - surpass people
  - interactive devices
  - robot musicians



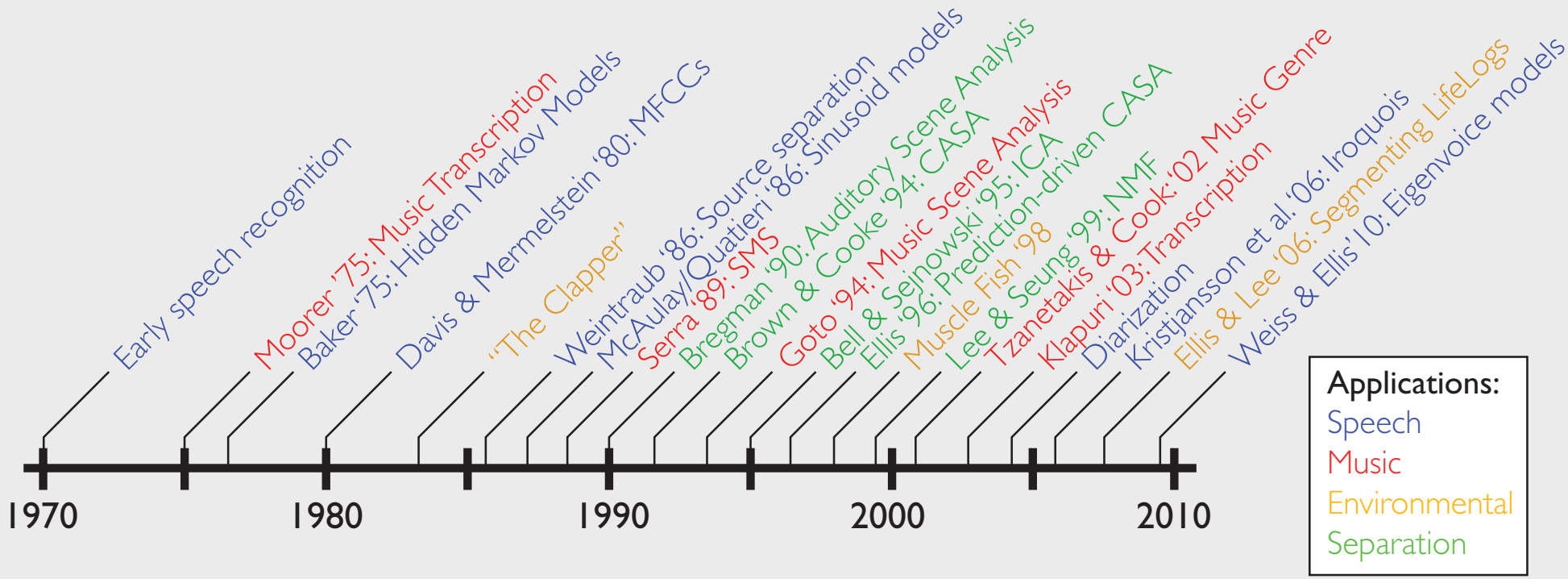
- ASR? yes
- Separation? maybe

Machine Listening Tasks



2. Key Tools in Machine Listening

- History: an eclectic timeline:

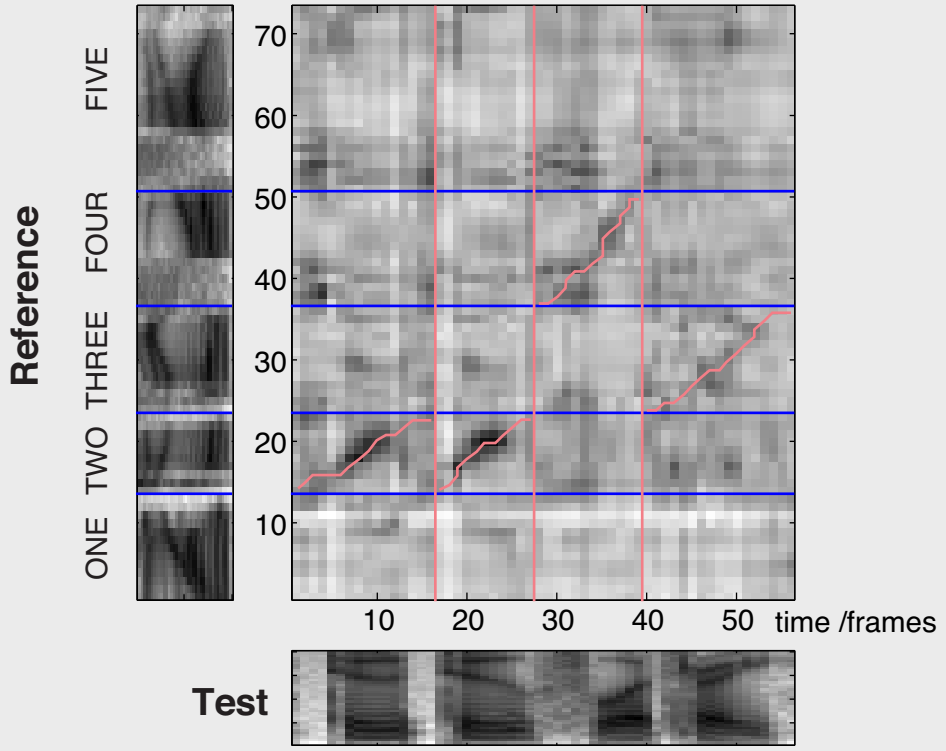


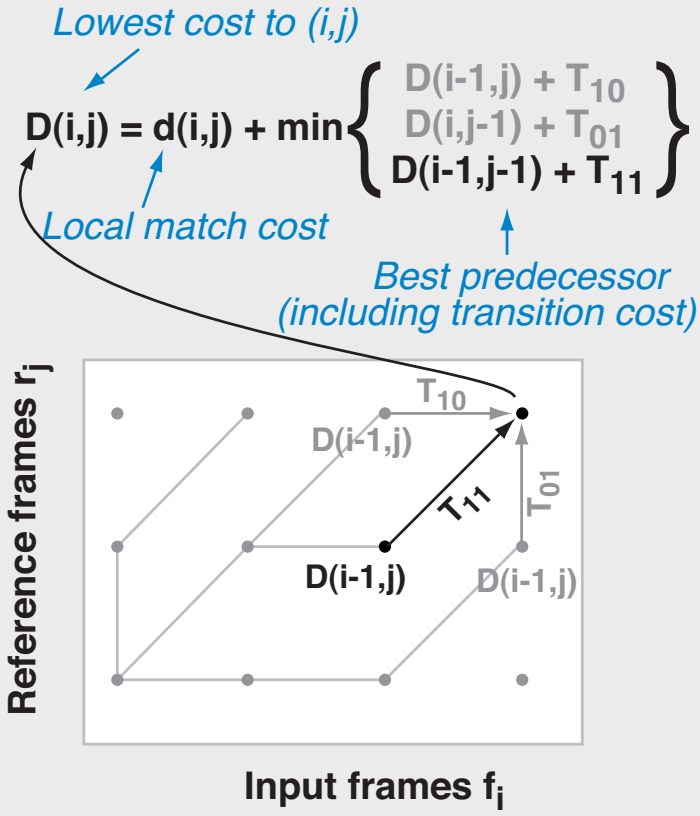
  - what worked? what is useful?

Early Speech Recognition

Vintsyuk '68
Ney '84

- DTW template matching



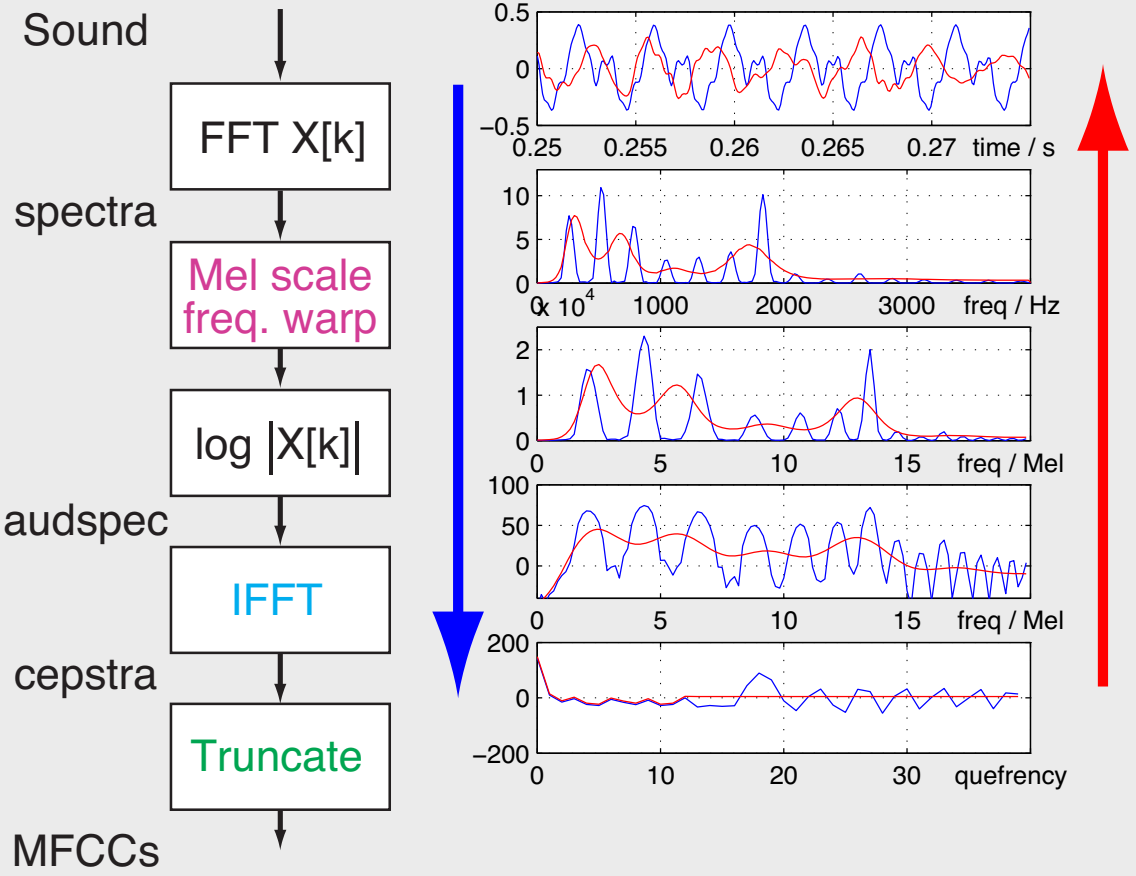


- Innovations:
  - template models
  - time warping

MFCCs

Davis & Mermelstein '80
Logan '00

- One feature to rule them all?



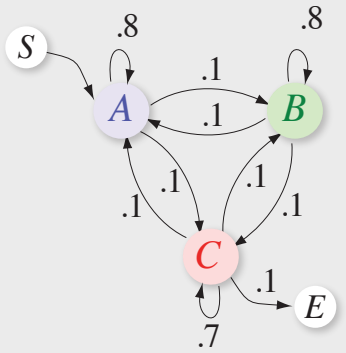
MFCC resynthesis:

- just the right amount of blurring

Hidden Markov Models

Baker '75
Jelinek '76

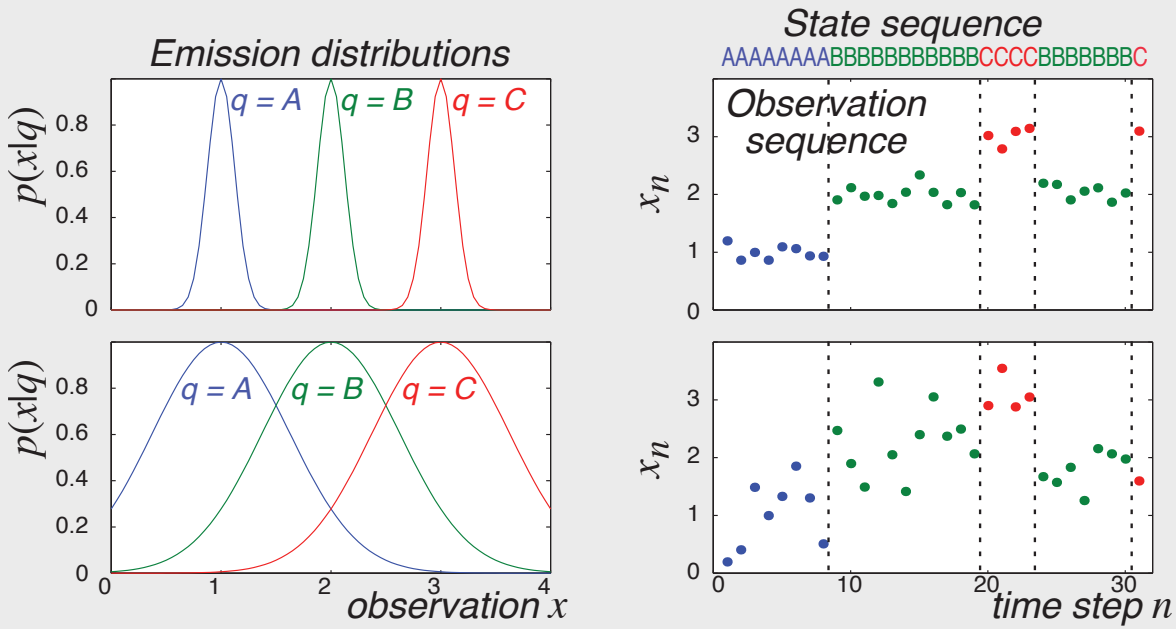
- Recognition as inferring the parameters of a generative model



$p(q_{n+1}\|q_n)$	q_{n+1} S	A	B	C	E
S	0	1	0	0	0
A	0	.8	.1	.1	0
q_n B	0	.1	.8	.1	0
C	0	.1	.1	.7	.1
E	0	0	0	0	1

S A A A A A A A A B B B B B B B B B C C C C B B B B B B B C E

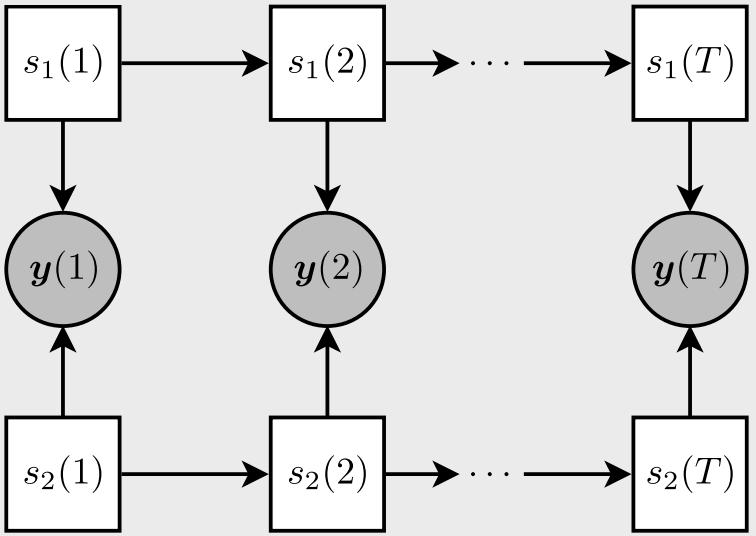
- Time warping + probability + training

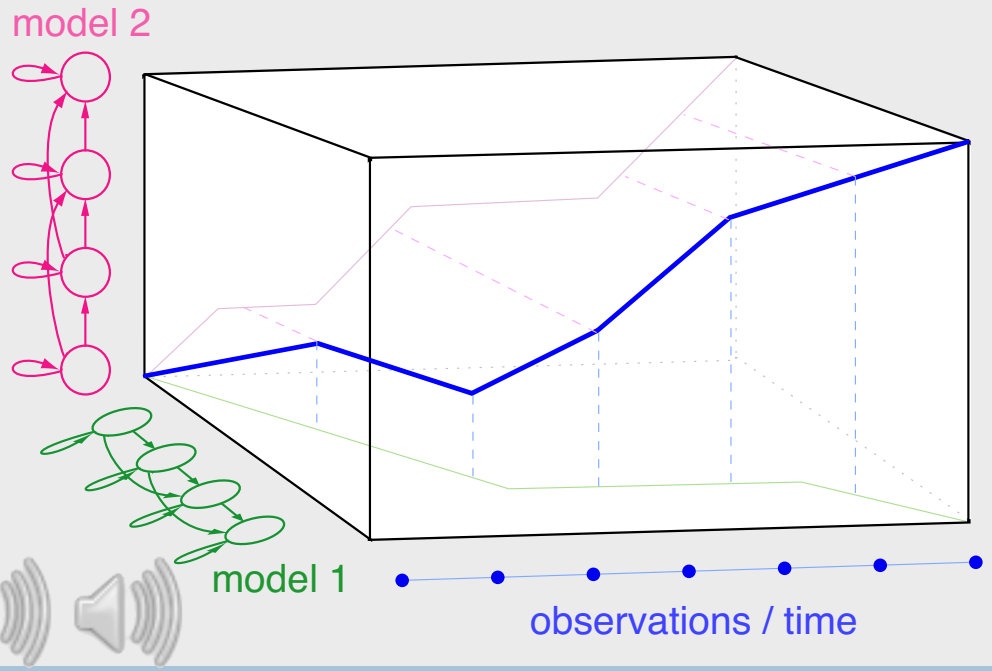


Model Decomposition

- HMMs applied to multiple sources
  - infer generative states for many models
  - combination of observations...
  - exponential complexity...

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

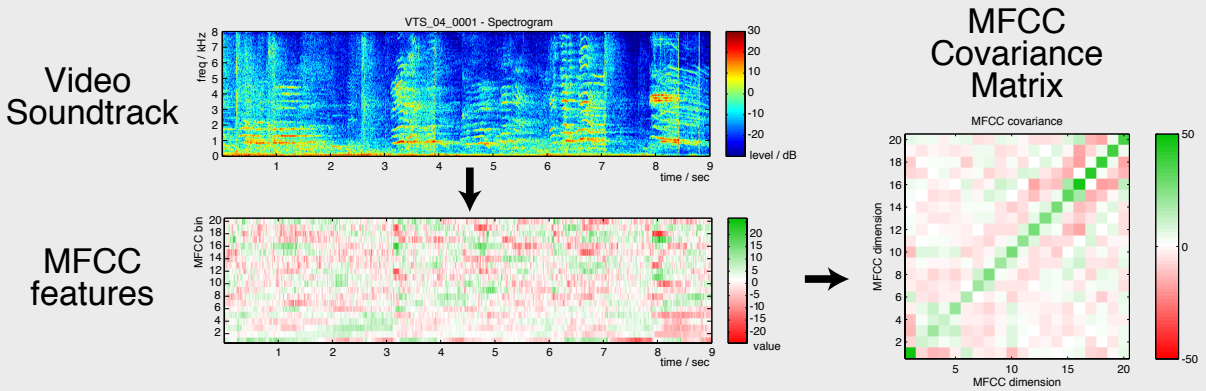




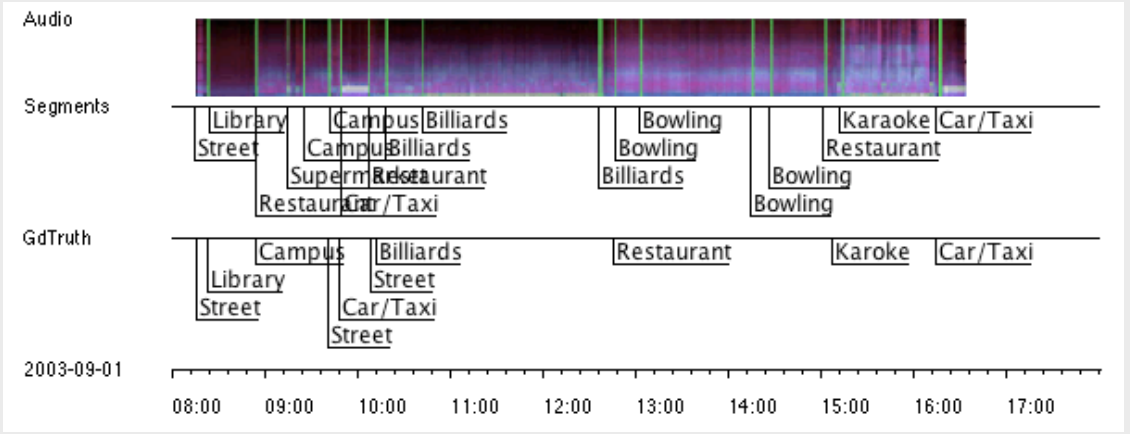
Segmentation & Classification

Chen & Gopalakrishnan '98
Tzanetakis & Cook '02
Lee & Ellis '06

- Label audio at scale of ~ seconds
  - segment when signal changes
  - describe by statistics
  - classify with existing models (HMMs?)



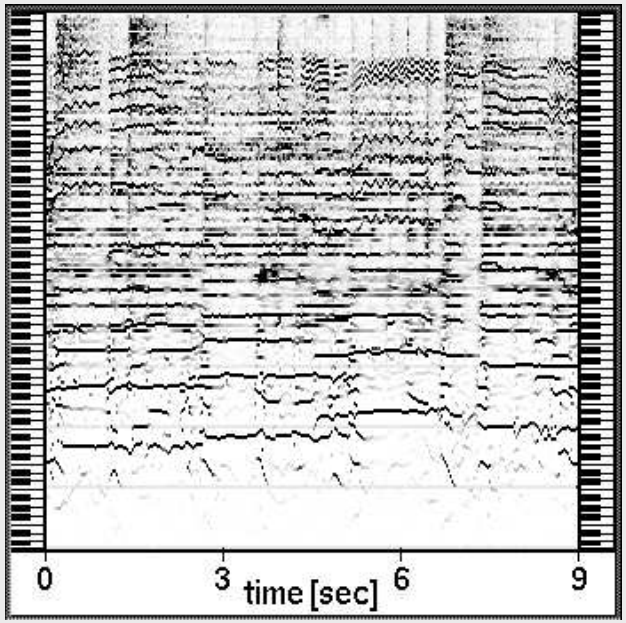
- Many supervised applications
  - speaker ID
  - music genre
  - environment ID

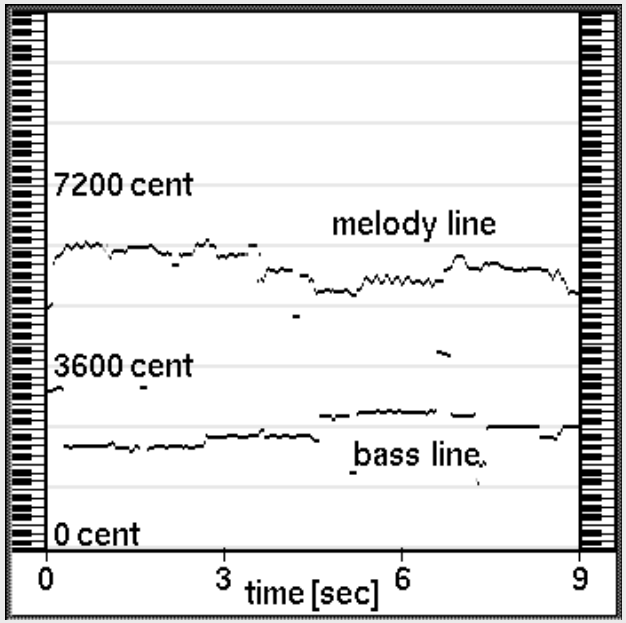


Music Transcription

- Music audio has a very specific structure
  - ... and an explicit abstract content

Moorer '75
Goto '94,'04
Klapuri '02,'06





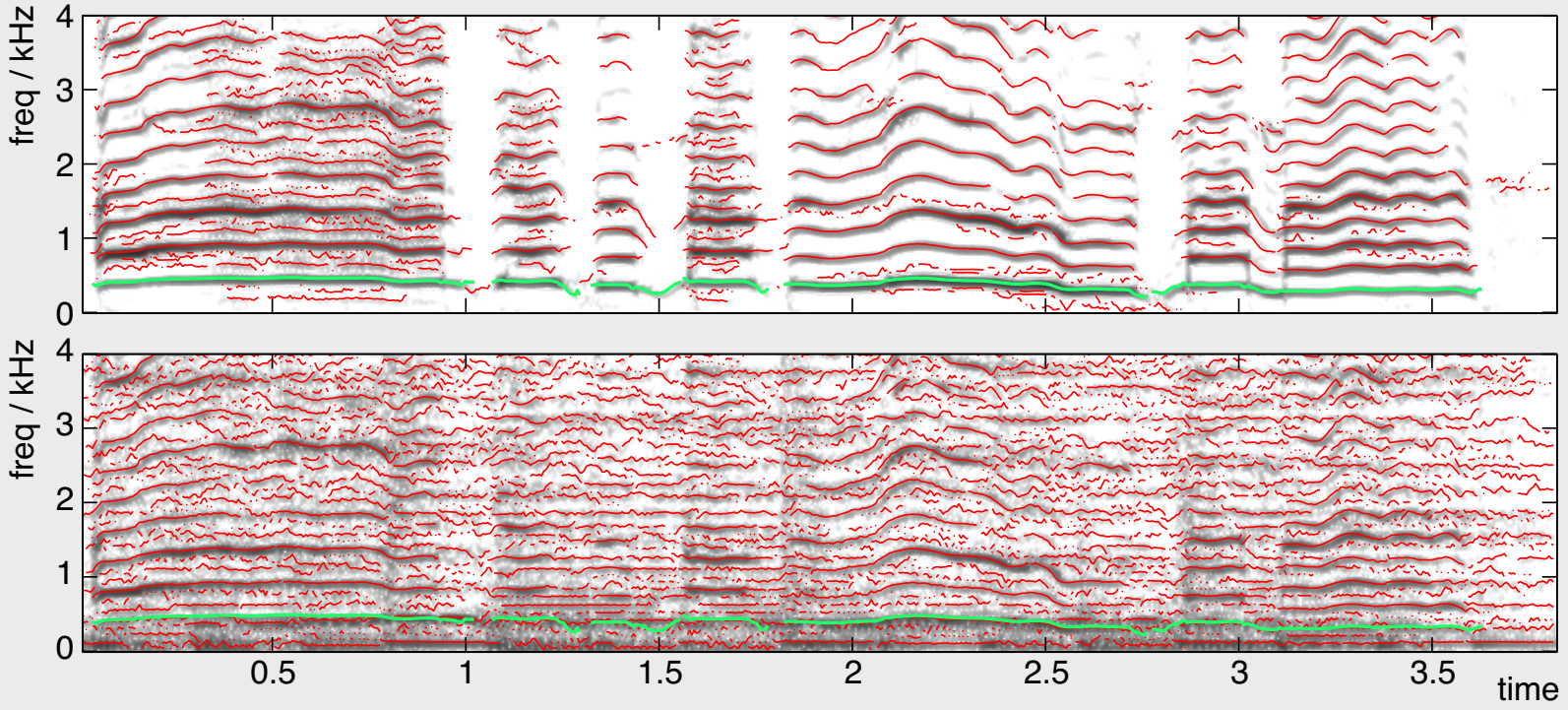
Goto '04

  - pitch + harmonics
  - rhythm
  - 'harmonized' voices

Sinusoidal Models

McAulay & Quatieri '84
Serra '89
Maher & Beauchamp '94

- Stylize a spectrogram into discrete components



from Wang '95

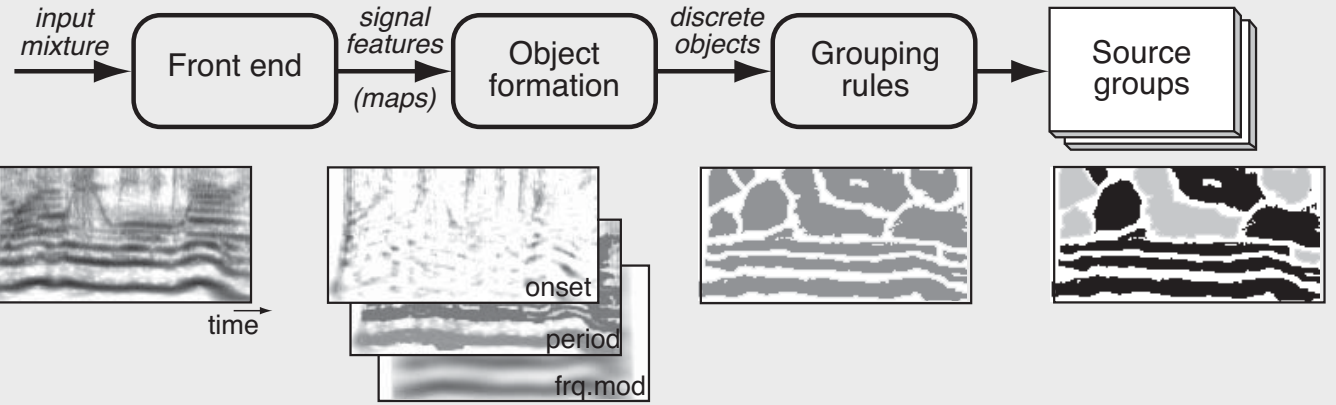
- discrete pieces → objects
- good for modification & resynthesis

Comp. Aud. Scene Analysis

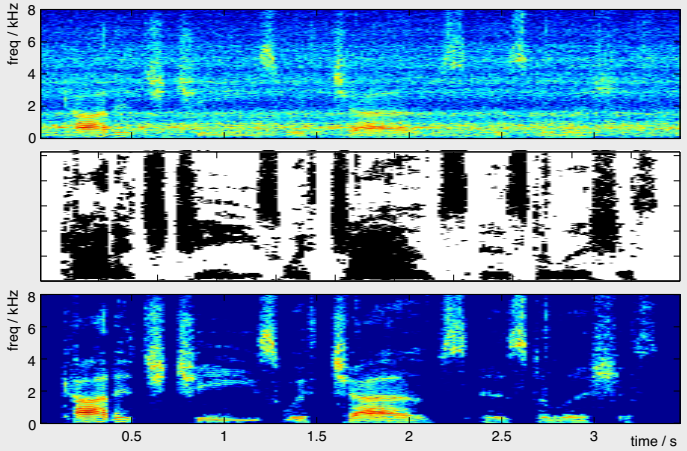
- Computer implementations of principles from [Bregman 1990] etc.

Weintraub '85
Brown & Cooke '94
Ellis '96
Roweis '03
Hu & Wang '04

- harmonicity, onset cues



- time-frequency masking for resynthesis

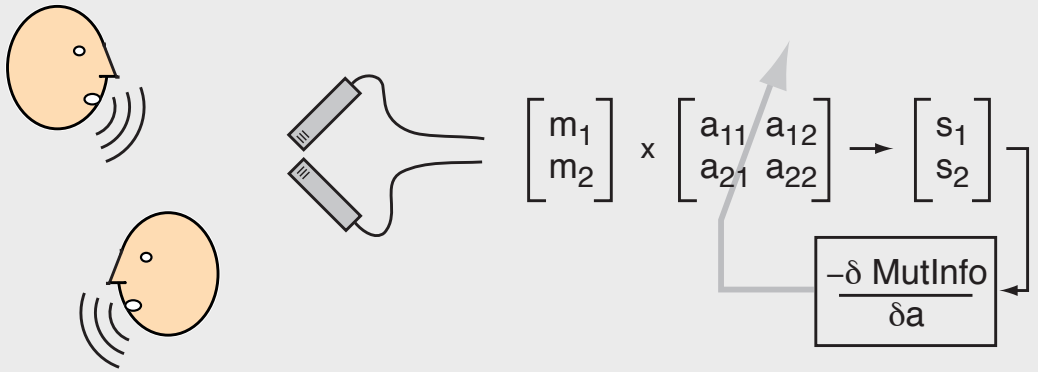


Oracle mask:

Independent Component Analysis

- Can separate "blind" combinations by maximizing independence of outputs

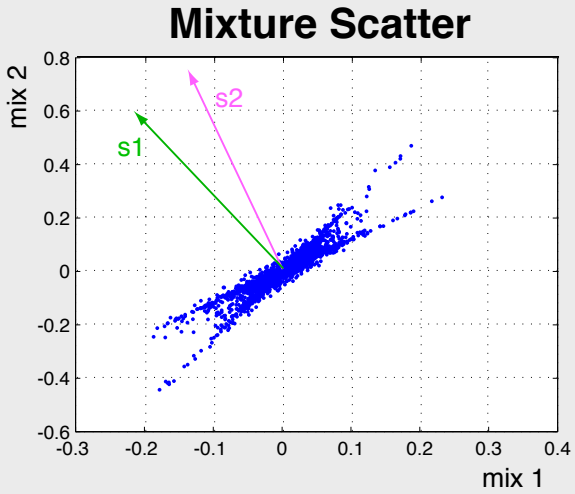
Bell & Sejnowski '95
Smaragdis '98

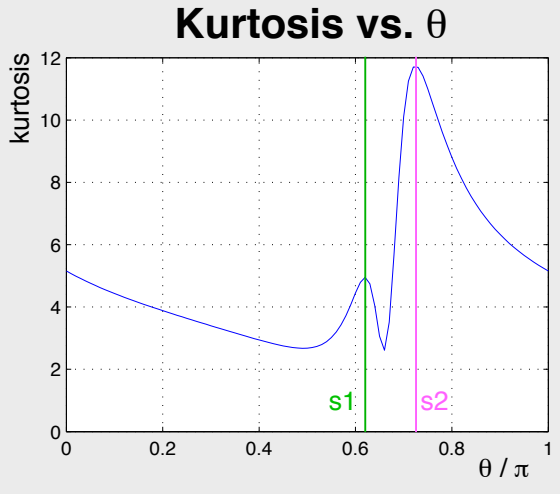


- kurtosis

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

as a measure of independence?





Nonnegative Matrix Factorization

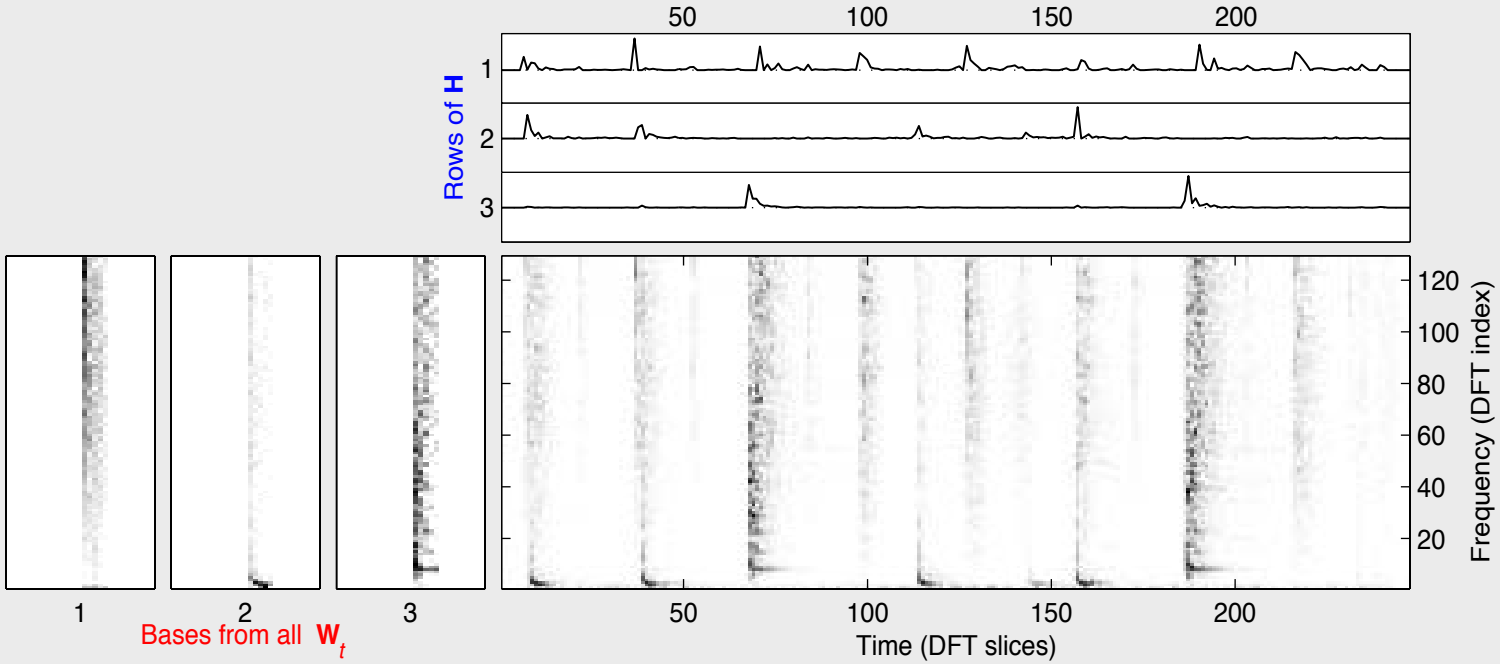
Lee & Seung '99
Abdallah & Plumbley '04
Smaragdis & Brown '03
Virtanen '07

- Decomposition of spectrograms into templates + activation

$$\mathbf{X} = \mathbf{W} \cdot \mathbf{H}$$

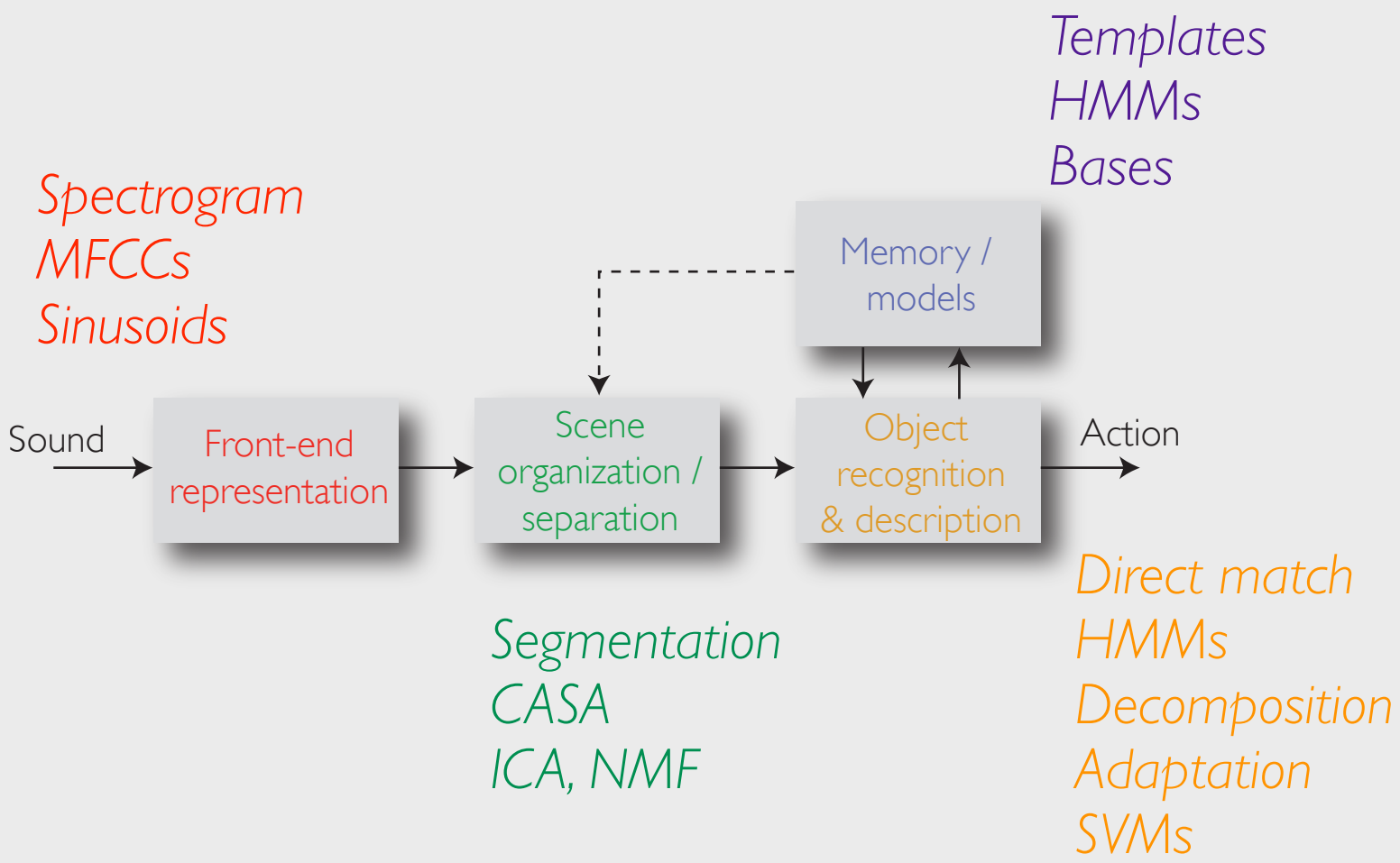
  - fast & forgiving gradient descent algorithm
  - fits neatly with time-frequency masking

Virtanen '03 sounds



Smaragdis '04

Summary of Key Ideas

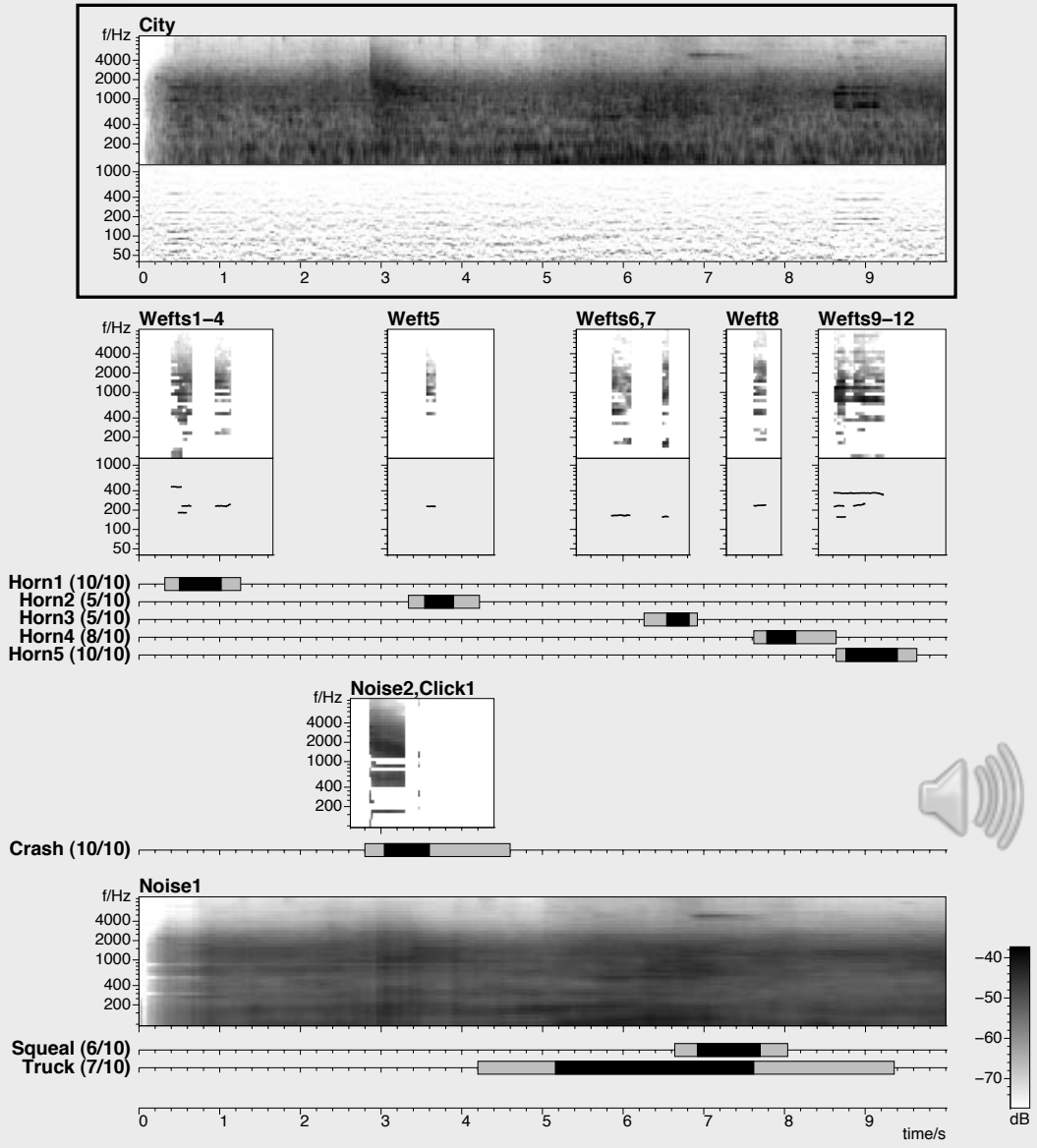


3. Open Issues

- Where to focus to advance machine listening?
  - task & evaluation
  - separation
  - models & learning
  - ASA
  - computational theory
  - attention & search
  - spatial vs. source

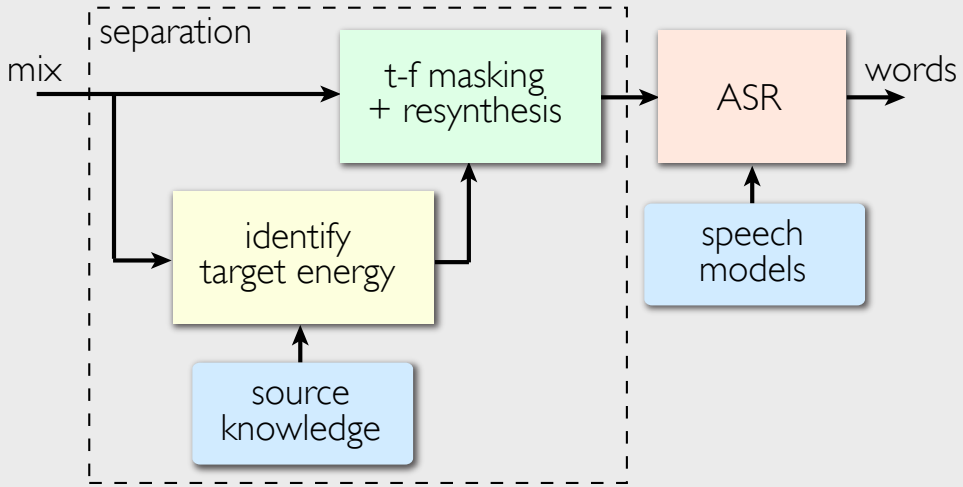
Scene Analysis Tasks

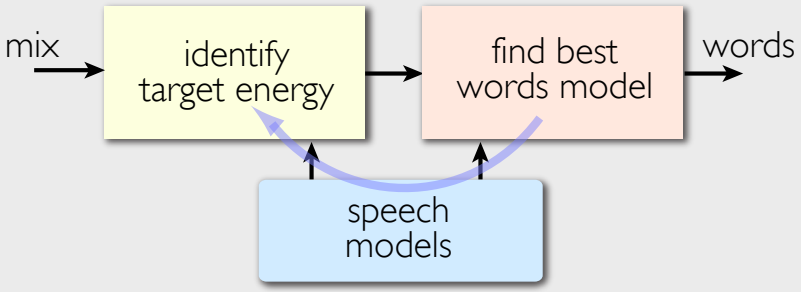
- Real scenes are complex!
  - background noise
  - many objects
  - prominence
- Is this just scaling, or is it different?
- Evaluation?



How Important is Separation?

- Separation systems often evaluated by SNR
  - based on pre-mix components - is this relevant?
- Best machine listening systems have resynthesis
  - e.g. Iroquois speech recognition - "separate then recognize"

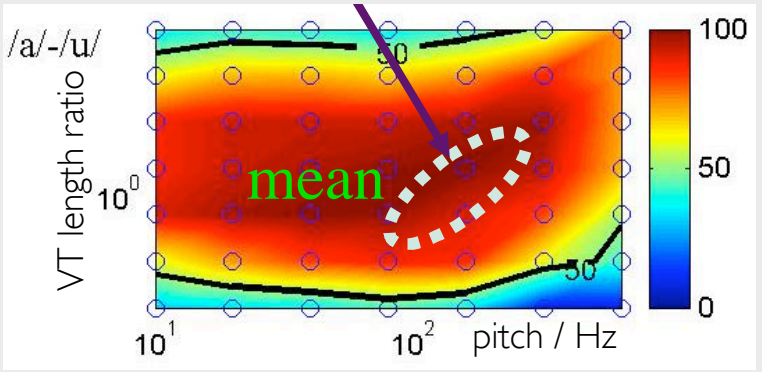




- Separated signals don't have to match originals to be useful

How Many Models?

- More specific models → better analysis
  - need dictionaries for "everything"??
- Model adaptation and hierarchy
  - speaker adapted models : base + parameters
  - extrapolation beyond normal

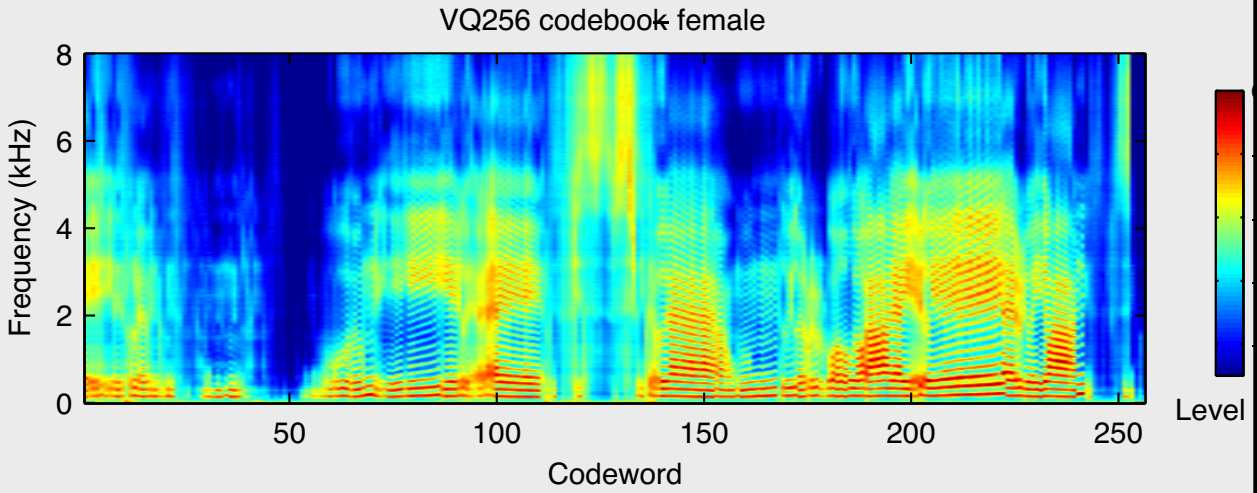


Smith, Patterson et al. '05

- Time scales of model acquisition
  - innate/evolutionary (hair-cell tuning)
  - developmental (mother tongue phones)
  - dynamic - the "Bolero" effect

Auditory Scene Analysis?

- Codebook models learn harmonicity, onset
  - ... to subsume rules/ representations of CASA



- Can also capture sequential structure
  - e.g. consonants follow vowels
  - use overlapping patches?
- But: computational factors

Computational Theory

- Marr's (1982) perspective on perception

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

- What is the computational theory of machine listening?
  - independence? sources?

Summary

- Machine Listening:
  Getting useful information from sound

- Techniques for:
  - representation
  - separation / organization
  - recognition / description
  - memory / models

- Where to go?
  - separation?
  - computational theory?

References 1/2

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

J. Baker (1975) "The DRAGON system – an overview," *IEEE Trans. Acoust. Speech, Sig. Proc.* 23, 24–29, 1975.

A. Bell & T. Sejnowski (1995) "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, 7, 1129-1159, 1995.

A. Bregman (1990) *Auditory Scene Analysis*, MIT Press.

G. Brown & M. Cooke (1994) "Computational auditory scene analysis," *Comp. Speech & Lang.* 8(4), 297–336, 1994.

S. Chen & P. Gopalakrishnan (1998) "Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion," *Proc. DARPA Broadcast News Transcription and Understanding Workshop*.

S. Davis & P. Mermelstein (1980) "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Sig. Proc.* 28, 357–366, 1980.

D. Ellis (1996) *Prediction-driven computational auditory scene analysis*, Ph.D thesis, EECS dept., MIT.

D. Ellis & K. Lee (2006) "Accessing minimal impact personal audio archives," *IEEE Multimedia* 13(4), 30-38, Oct-Dec 2006.

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

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

M. Goto & Y. Muraoka (1994) "A beat tracking system for acoustic signals of music," *Proc. ACM Intl. Conf. on Multimedia*, 365–372, 1994.

M. Goto (2004) "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Comm.*, 43(4), 311–329, 2004.

S. Handel (1989) *Listening*, MIT Press.

J. Hershey, S. Rennie, P. Olsen, T. Kristjansson (2006) "Super-human multi-talker speech recognition: A graphical modeling approach," *Comp. Speech & Lang.*, 24, 45–66.

G. Hu and D.L. Wang (2004) "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Tr. Neural Networks*, 15(5), Sep. 2004.

F. Jelinek (1976) "Continuous speech recognition by statistical methods," *Proc. IEEE* 64(4), 532–556.

A. Klapuri (2003) "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," IEEE Trans. Speech & Audio Proc., 11(6).

A. Klapuri (2006) "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes," *Proc. Int. Symp. Music Info. Retr.*

T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, and R. Gopinath (2006) "Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system," *Proc. Interspeech*, 775-1778, 2006.

D. Lee & S. Seung (1999) "Learning the Parts of Objects by Non-negative Matrix Factorization", *Nature* 401, 788.

B. Logan (2000) "Mel frequency cepstral coefficients for music modeling," *Proc. Int. Symp. Music Inf. Retrieval*, Plymouth, September 2000.

R. Maher & J. Beachamp (1994) "Fundamental frequency estimation of musical signals using a two-way mismatch procedure," *J. Acoust. Soc. Am.*, 95(4), 2254–2263, 1994.

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

References 2/2

R. McAulay & T. Quatieri (1986) "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Sig. Proc.* 34, 744–754, 1986.

A. Moorer (1975) *On the segmentation and analysis of continuous musical sound by computer*, Ph.D. thesis, CS dept., Stanford Univ.

H. Ney (1984) "The use of a one stage dynamic programming algorithm for connected word recognition," *IEEE Trans. Acoust. Speech Sig. Proc.* 32, 263–271, 1984.

S. Roweis (2003) "Factorial Models and Refiltering for Speech Separation and Denoising", *Proc. Eurospeech*, 2003.

X. Serra (1989) *A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition,* Ph.D. thesis, Music dept., Stanford Univ.

P. Smaragdis (1998) "Blind separation of convolved mixtures in the frequency domain," *Intl. Wkshp. on Indep. & Artif. Neural Networks*, Tenerife, Feb. 1998.

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

P. Smaragdis (2004) "Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs," *Proc. ICA*, LNCS 3195, 494–499.

D. Smith, R. Patterson, R. Turner, H. Kawahara & T. Irino (2005) "The processing and perception of size information in speech sounds," J Acoust Soc Am. 117(1), 305–318.

G. Tzanetakis & P. Cook (2002) "Musical genre classification of audio signals," *IEEE Tr. Speech & Audio Proc.* 10(5).

A. Varga & R. Moore (1990) "Hidden Markov Model decomposition of speech and noise," *IEEE ICASSP*, 845–848, 1990.

T. Vintsyuk (1971) "Element-wise recognition of continuous speech composed of words from a specified dictionary," *Kibernetika* 7, 133–143, 1971.

T. Virtanen (2007) "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.

A. Wang (2006) "The Shazam music recognition service," *Comm. ACM* 49 (8), 44-48, 2006.

M. Weintraub (1986) *A theory and computational model of auditory monaural sound separation*, Ph.D. thesis, EE dept., Stanford Univ.

R. Weiss & D. Ellis (2010) "Speech separation using speaker-adapted Eigenvoice speech models," *Comp. Speech & Lang.*, 24(1), 16–29, Jan 2010.