Augmenting and Exploiting Auditory Perception for Complex Scene Analysis

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

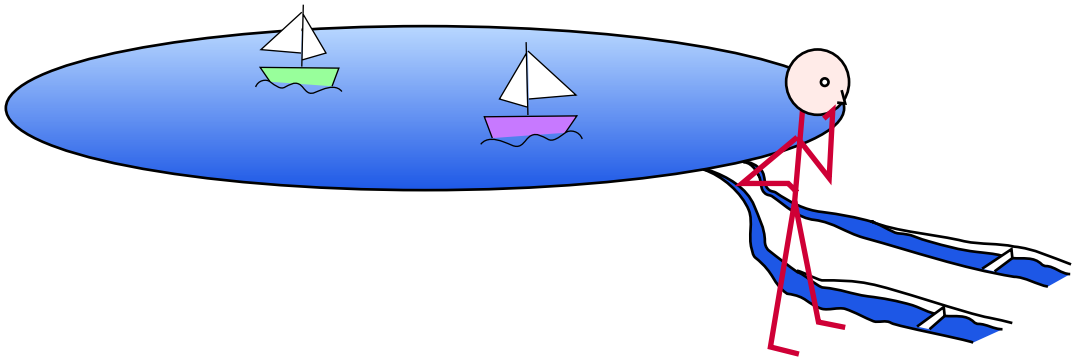
dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

1. Human Auditory Scene Analysis
2. Computational Acoustic Scene Analysis
3. Virtual and Augmented Audio
4. Future Audio Analysis & Display

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

Columbia University
In the City of New York

1. Human Auditory Scene Analysis



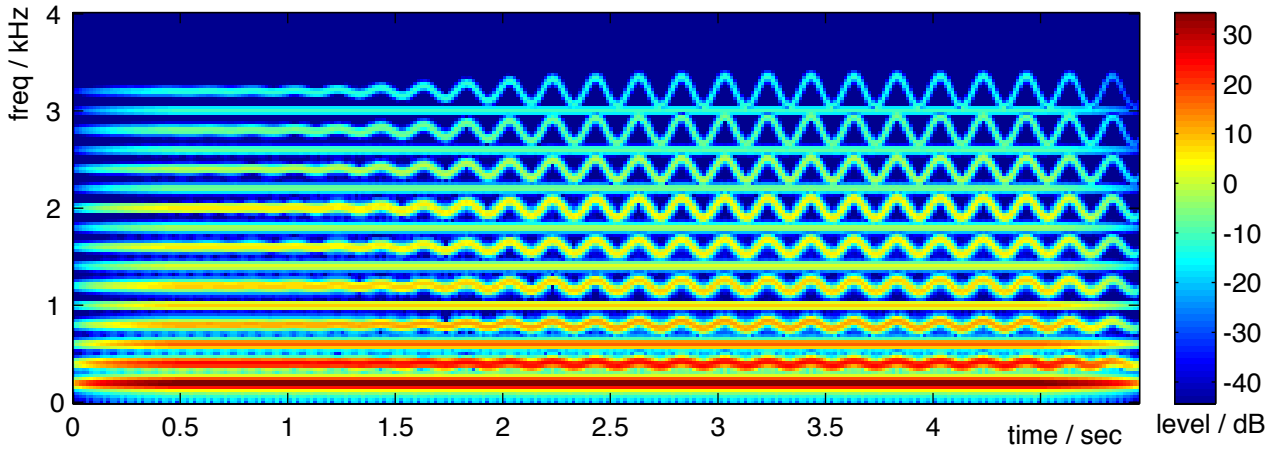
"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- Now:
  Hearing as the model for machine perception
- Future: Machines to enhance human perception

Auditory Scene Analysis

Bregman '90
Darwin & Carlyon '95

- Listeners organize sound mixtures into discrete perceived sources based on within-signal cues (audio + ...)
  - common onset + continuity
  - harmonicity
  - spatial, modulation, ...
  - learned "schema"

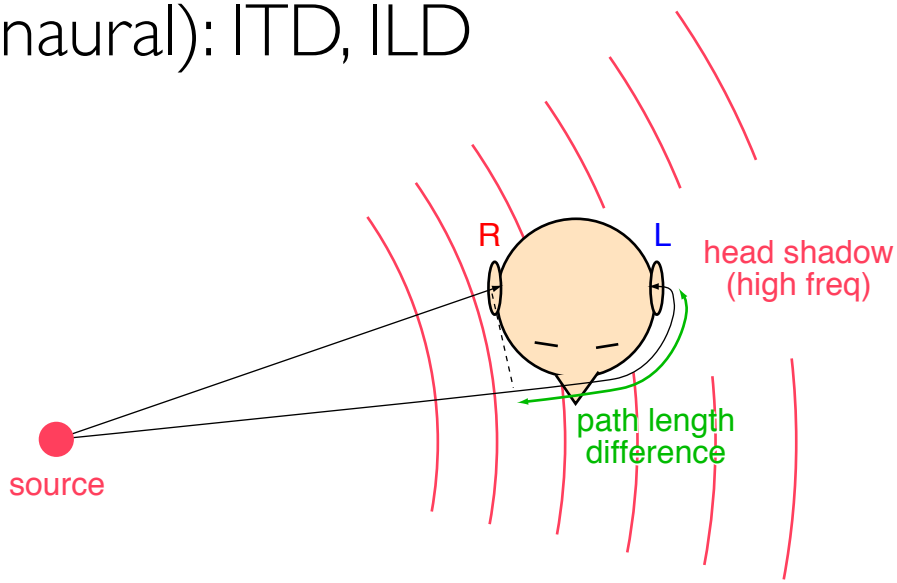


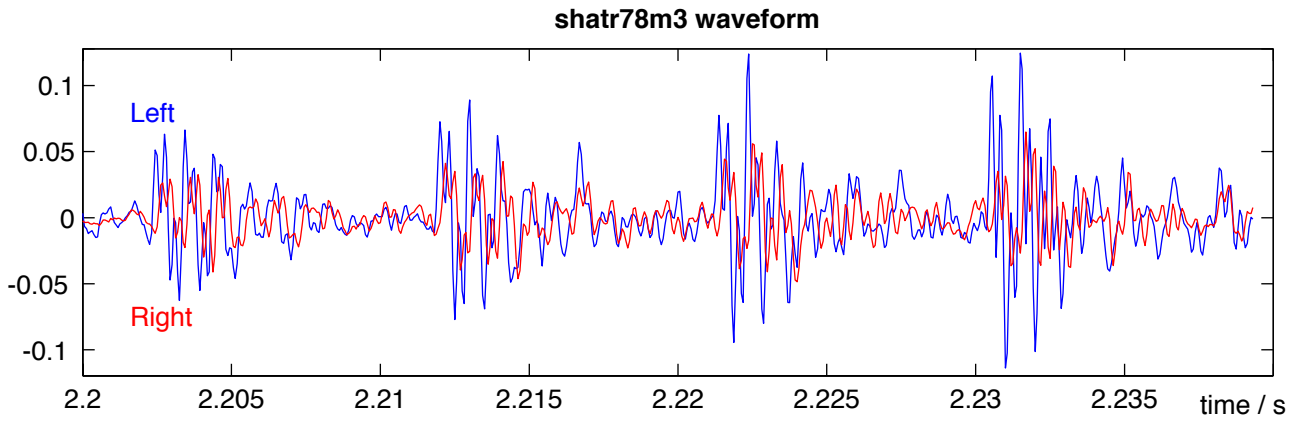
reynolds-mcadams-dpwe.wav

Spatial Hearing

Blauert '96

- People perceive sources based on cues
  - spatial (binaural): ITD, ILD





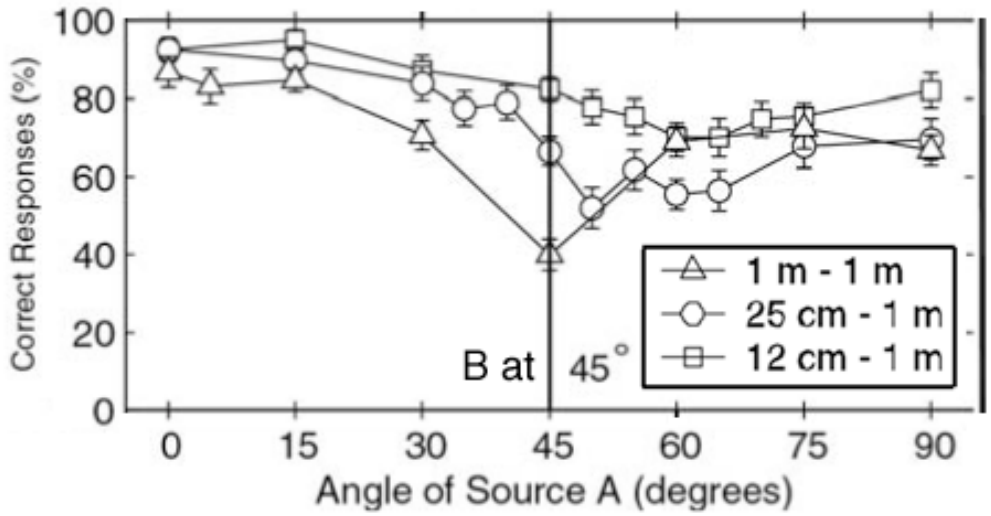
Human Performance: Spatial Info

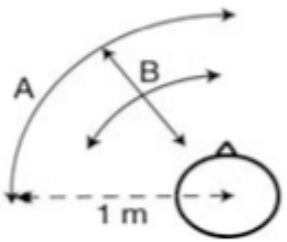
Brungart et al.'02

- Task: Coordinate Response Measure
  - "Ready Baron go to green eight now"
  - 256 variants, 16 speakers
  - correct = color and number for "Baron"

crm-11737+16515.wav

- Accuracy as a function of spatial separation:



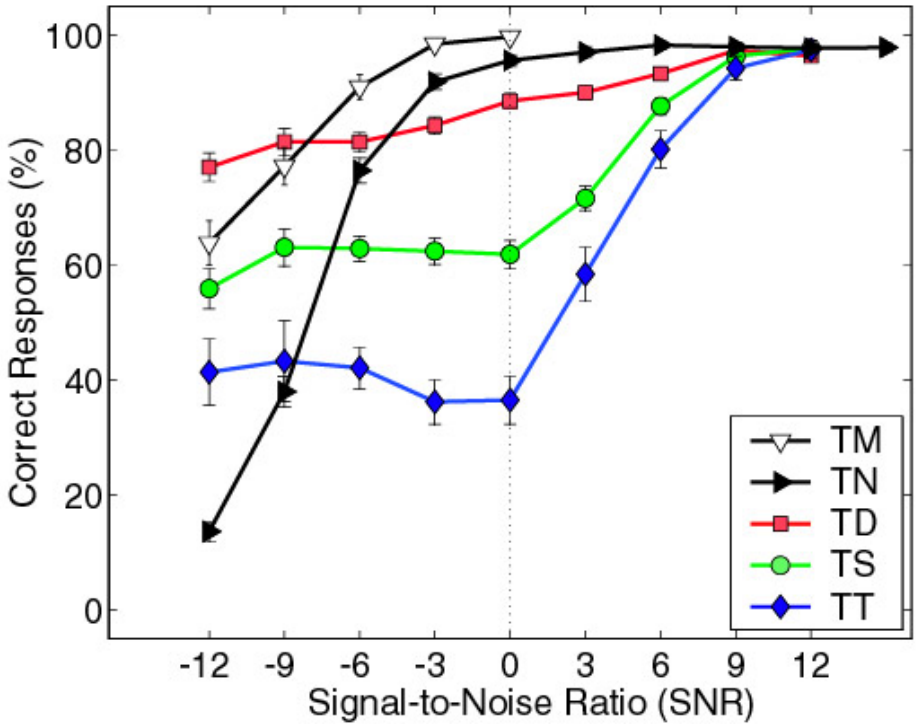


  - A, B same speaker

Human Performance: Source Info

Brungart et al.'01

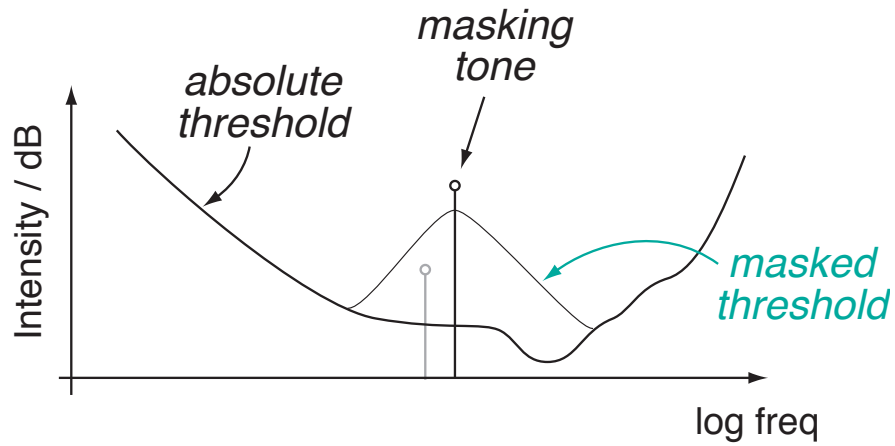
- **CRM varying the level and voice character**
  - (same spatial location)



  - energetic vs. informational masking

Human Hearing: Limitations

- Sensor number: just 2 ears
- Sensor location: short, horizontal baseline
- Sensor performance: local dynamic range

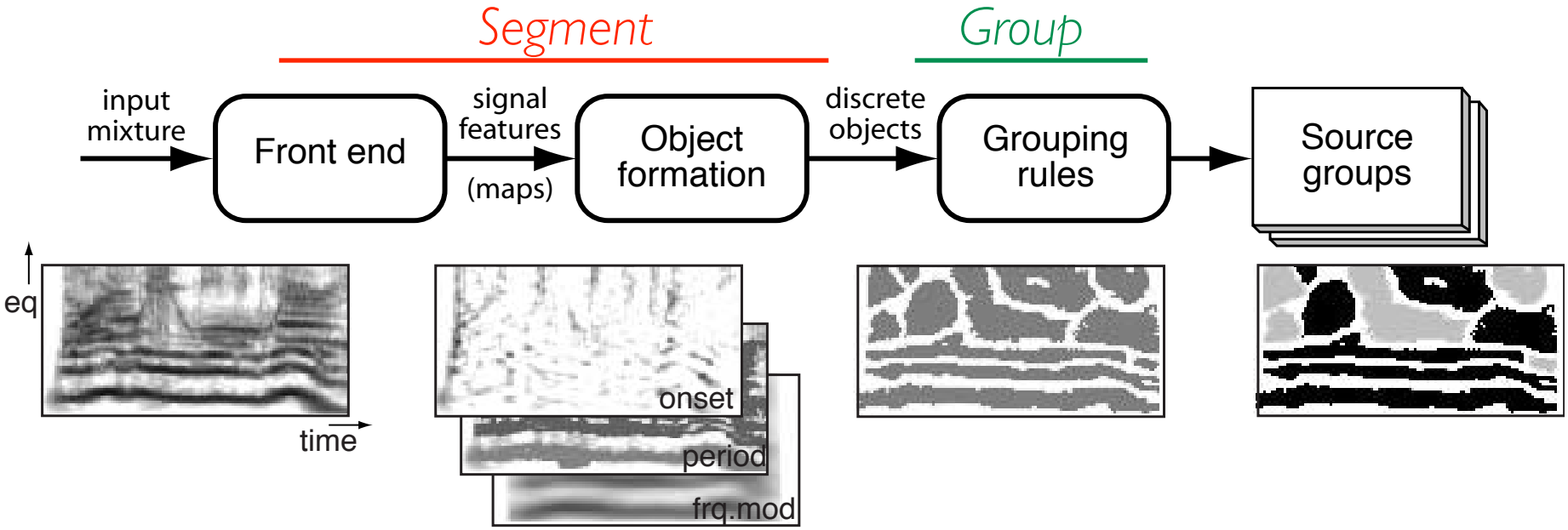


- Processing: Attention & Memory limits
  - integration time

2. Computational Scene Analysis

Brown & Cooke'94
Okuno et al.'99
Hu & Wang'04 ...

- Central idea:
  Segment time-frequency into sources
  based on perceptual grouping cues

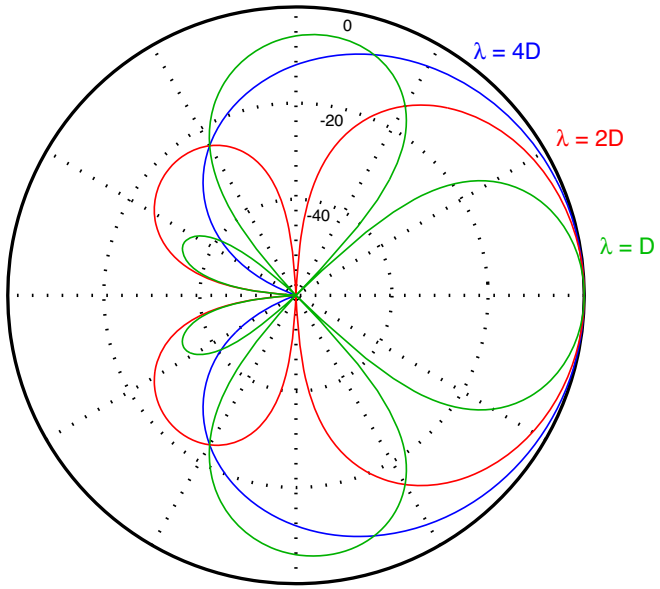


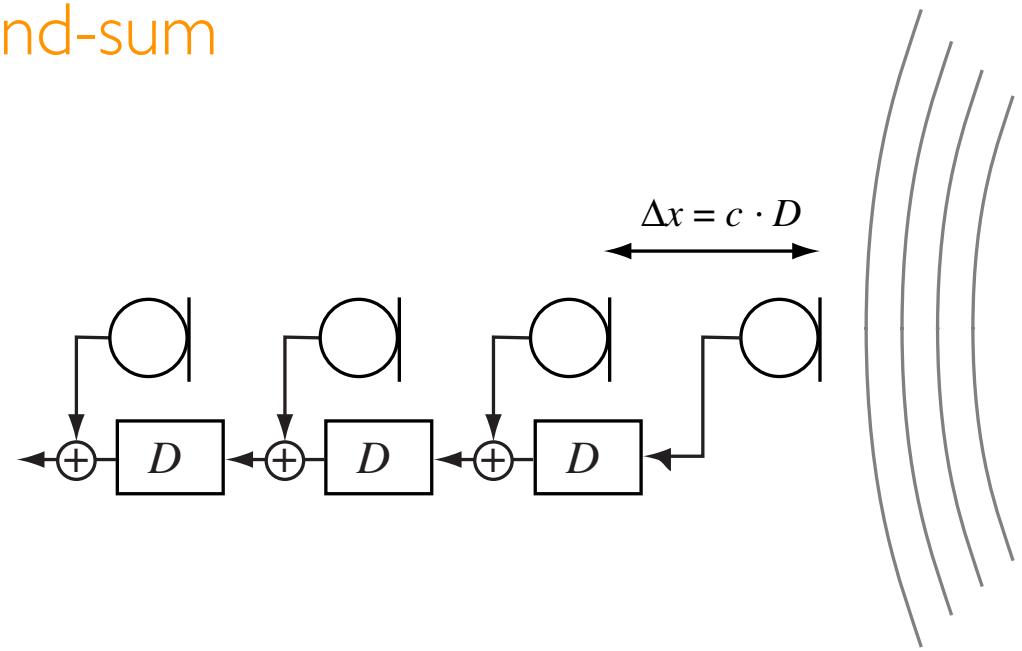
- ... principal cue is harmonicity

Spatial Info: Microphone Arrays

Benesty, Chen, Huang '08

- If interference is diffuse, can simply boost energy from target direction
  - e.g. shotgun mic - delay-and-sum



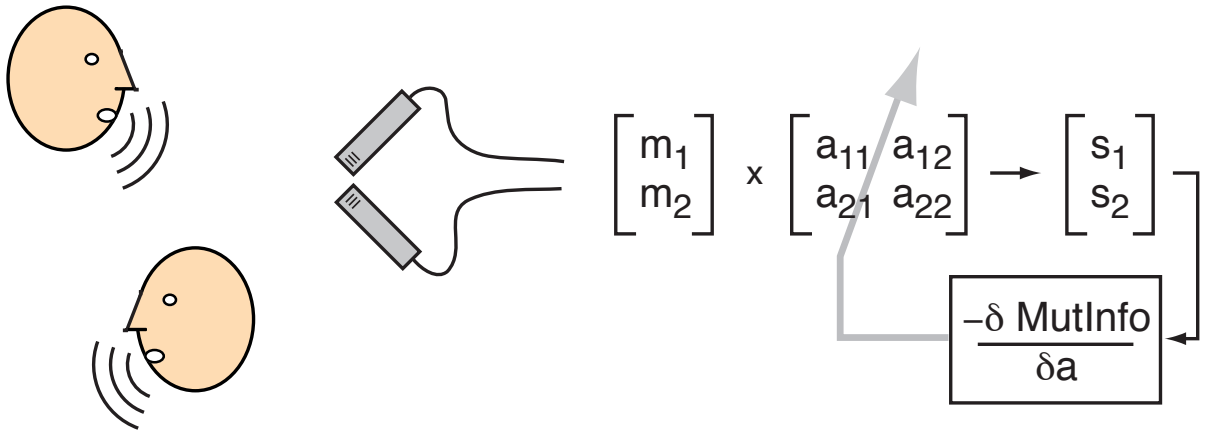


  - off-axis spectral coloration
  - many variants - filter & sum, sidelobe cancelation ...

Independent Component Analysis

Bell & Sejnowski '95
Smaragdis '98

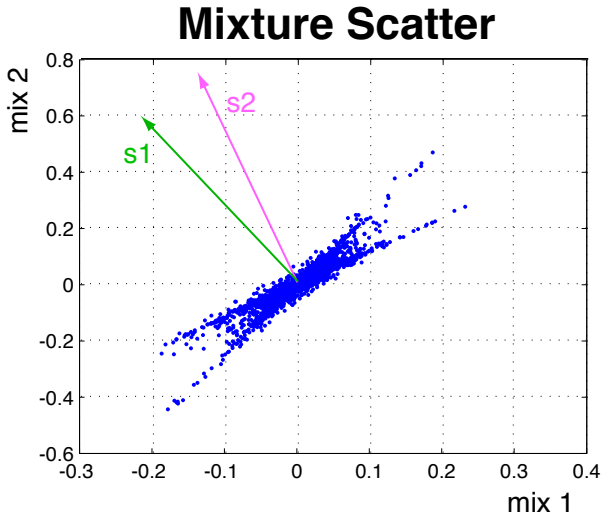
- Separate "blind" combinations by maximizing independence of outputs

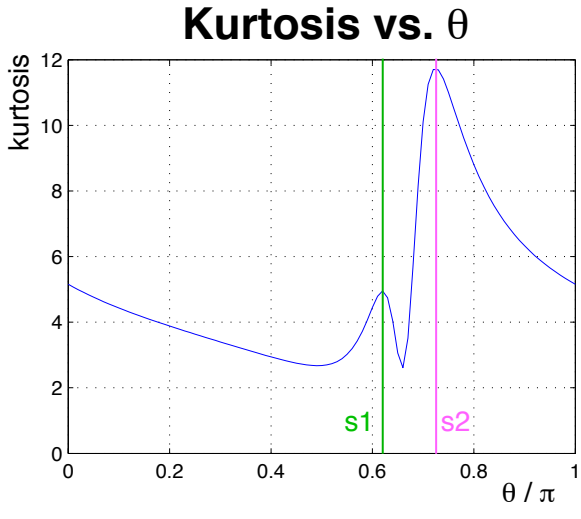


- kurtosis

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

as a measure of independence?

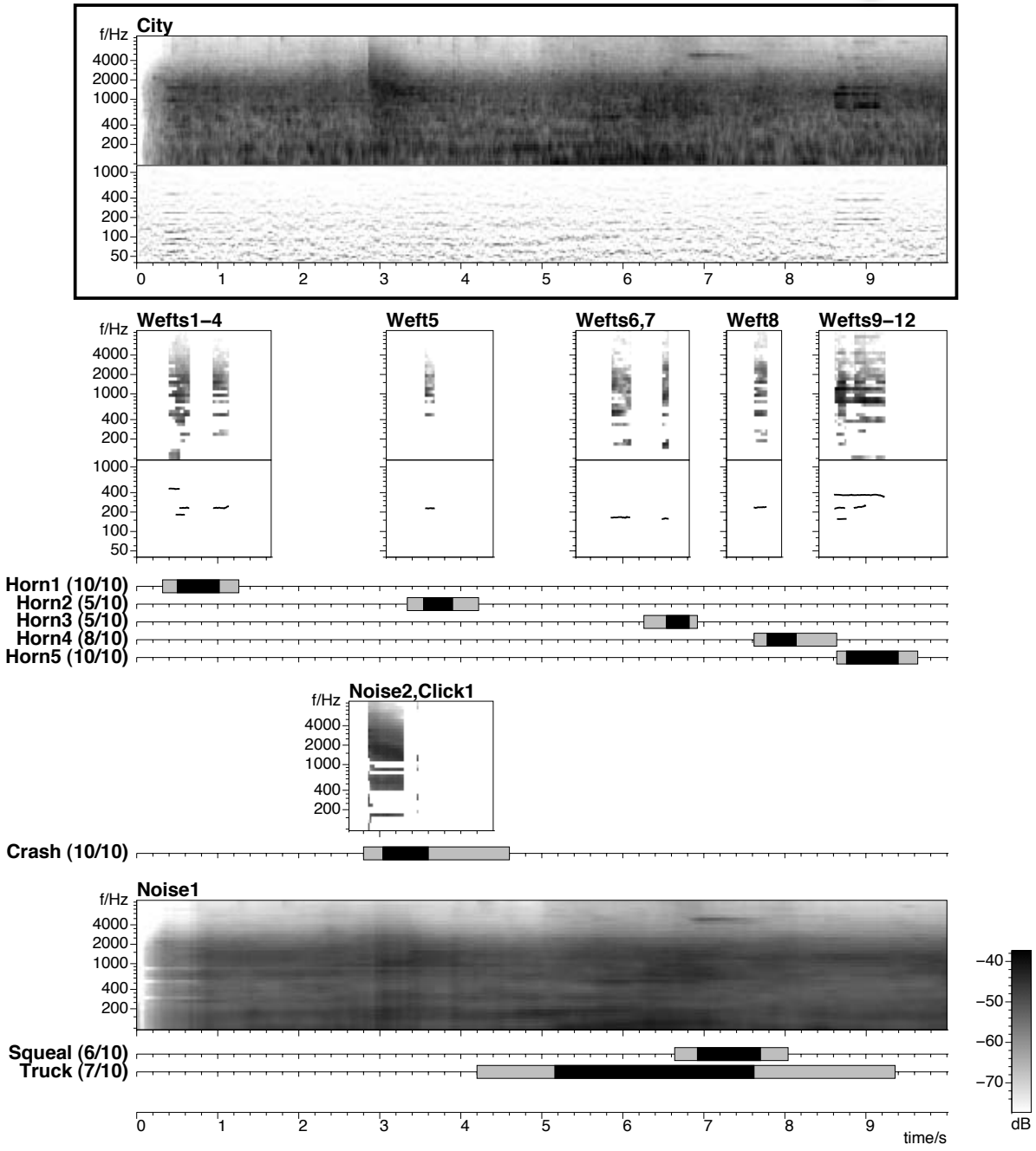




Environmental Scene Analysis

- Find the pieces a listener would report

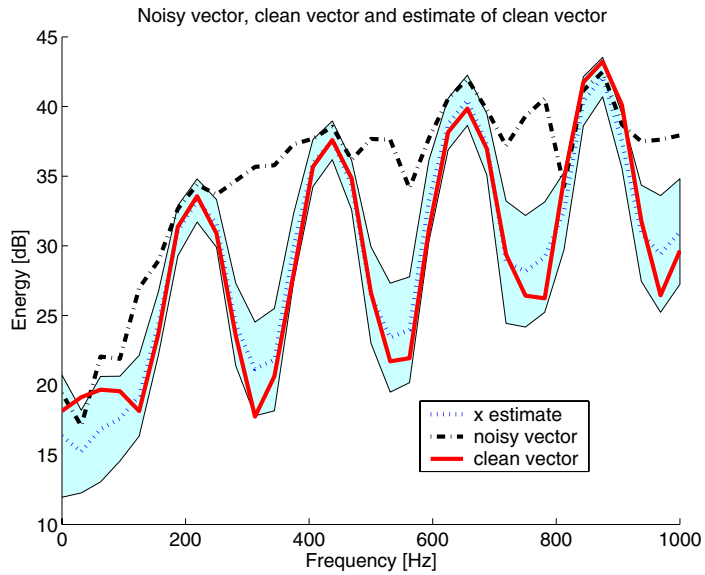
Ellis '96

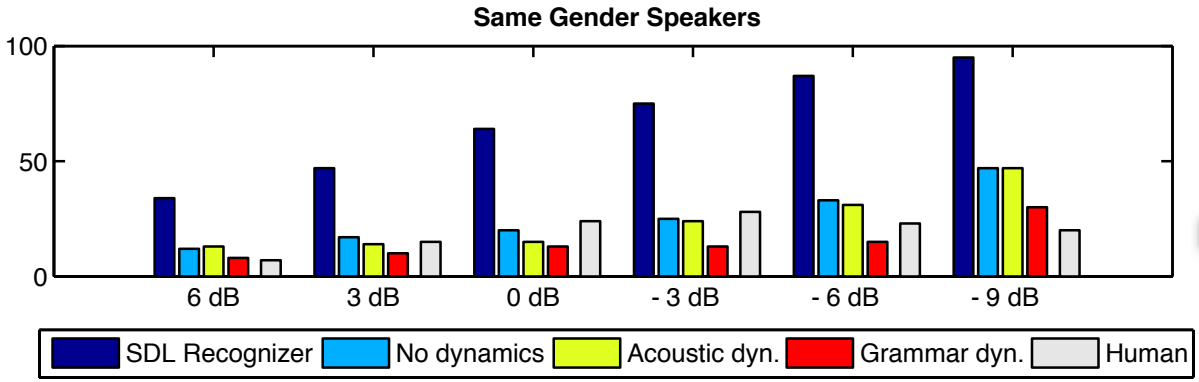


"Superhuman" Speech Analysis

Kristjansson, Hershey et al. '06

- IBM's 2006 Iroquois speech separation system
  Key features:
  - detailed state combinations
  - large speech recognizer
  - exploits grammar constraints
  - 34 per-speaker models
- "Superhuman" performance
  - ... in some conditions



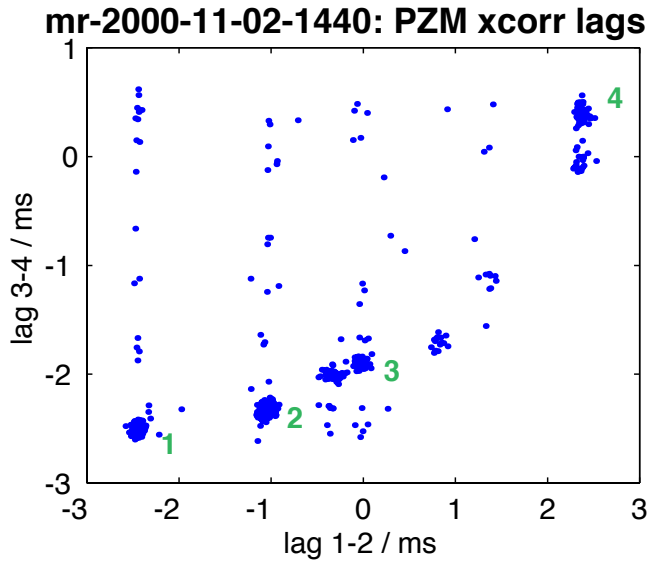


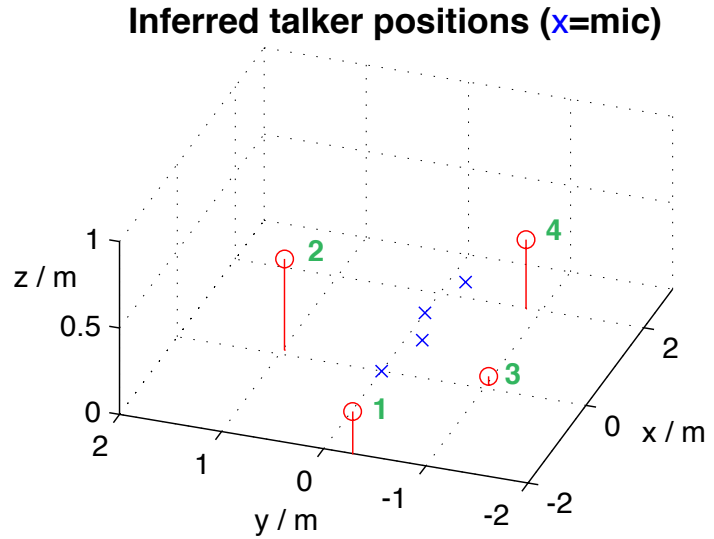
Meeting Recorders

Janin et al. '03
Ellis & Liu '04

- Distributed mics in meeting room

- Between-mic correlations locate sources

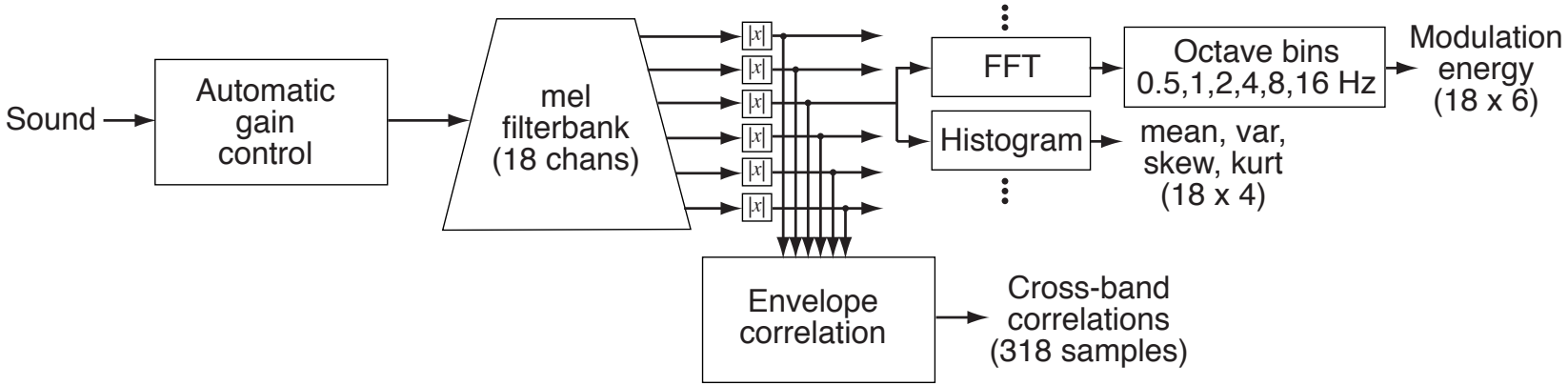




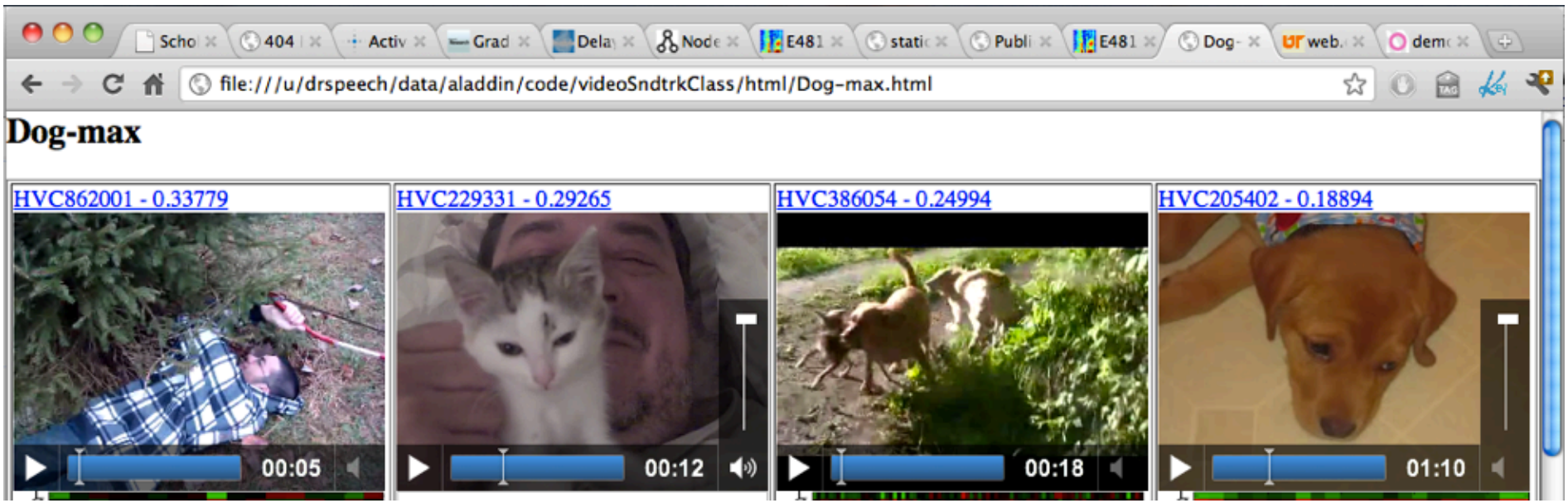
Environmental Sound Classification

Ellis, Zheng, McDermott '11

- Trained models using e.g. "texture" features



- Paradoxical results

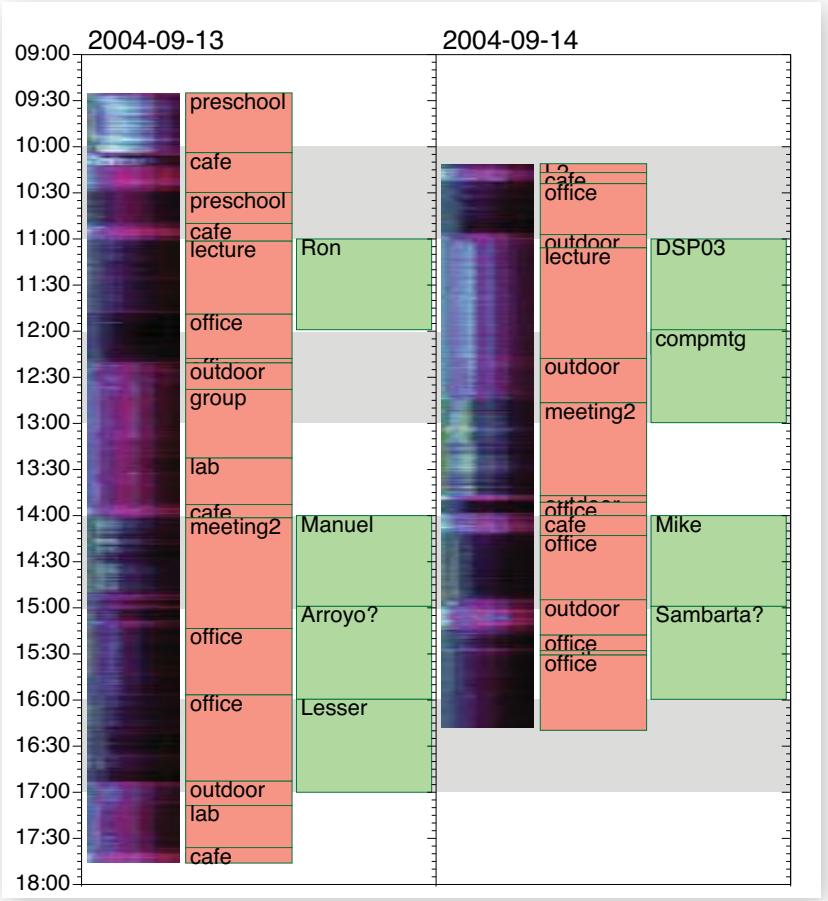


Audio Lifelogs

Lee & Ellis '04

- Body-worn continuous recording

- Long time windows for episode-scale segmentation, clustering, and classification



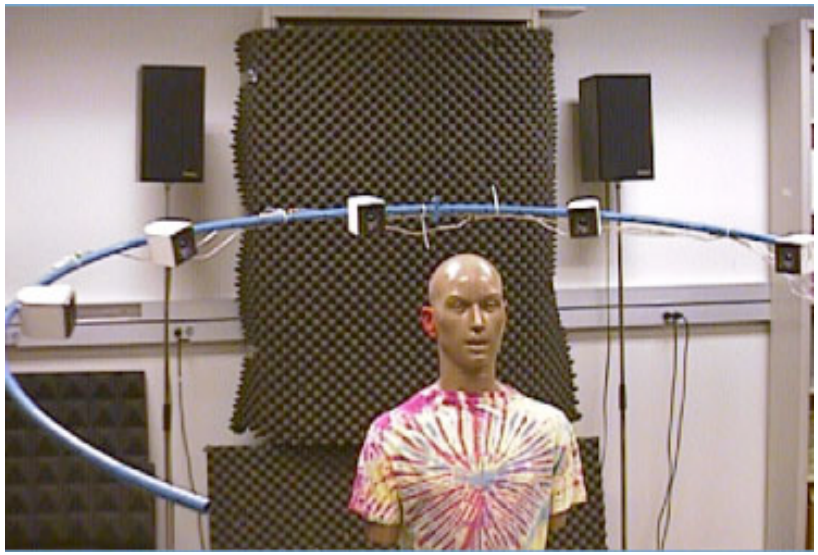
Machines: Current Limitations

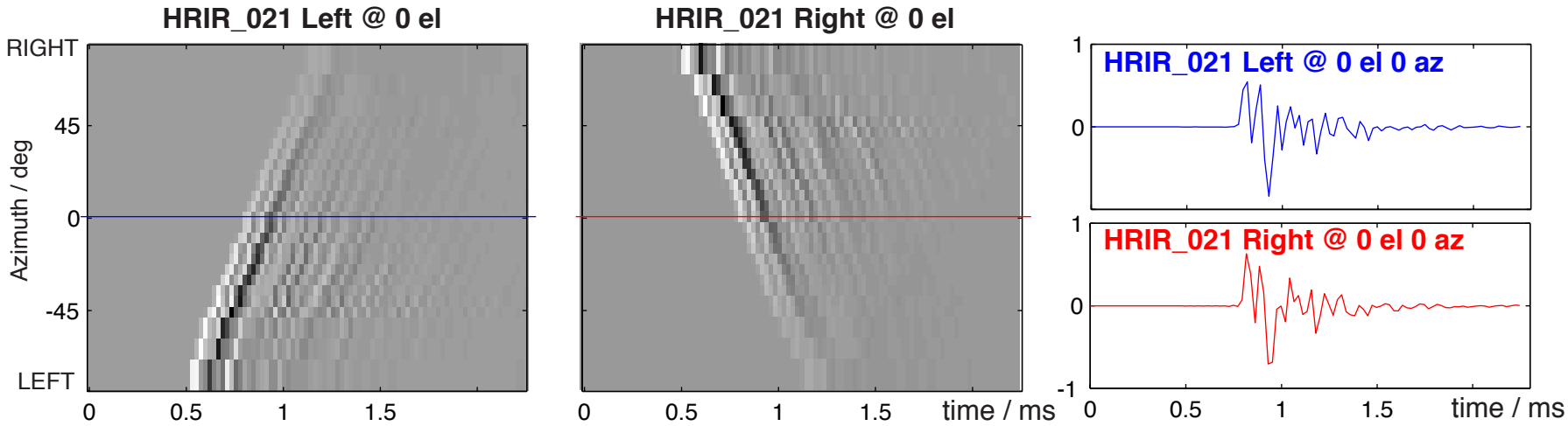
- Separating overlapping sources
  - blind source separation
- Separating individual events
  - segmentation
- Learning & classifying source categories
  - recognition of individual sounds and classes

3. Virtual and Augmented Audio

Brown & Duda '98

- Audio signals can be effectively spatialized by convolving with Head-Related Impulse Responses (HRIRs)



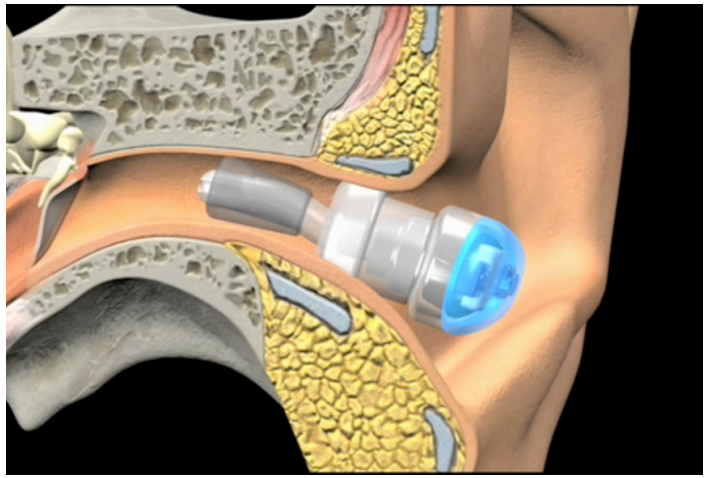


- Auditory localization also uses head-motion cues

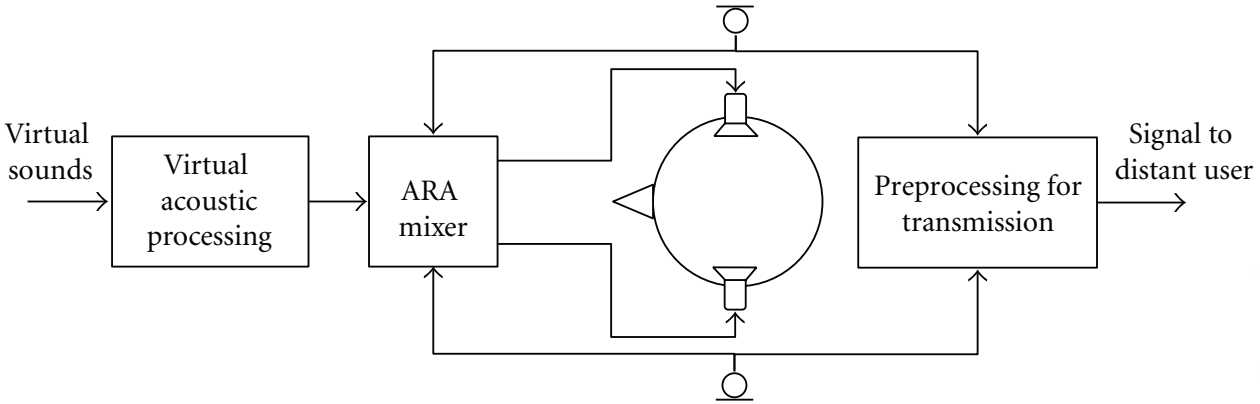
Augmented Audio Reality

- Pass-through and/or mix-in

Härmä et al. '04
Hearium '12



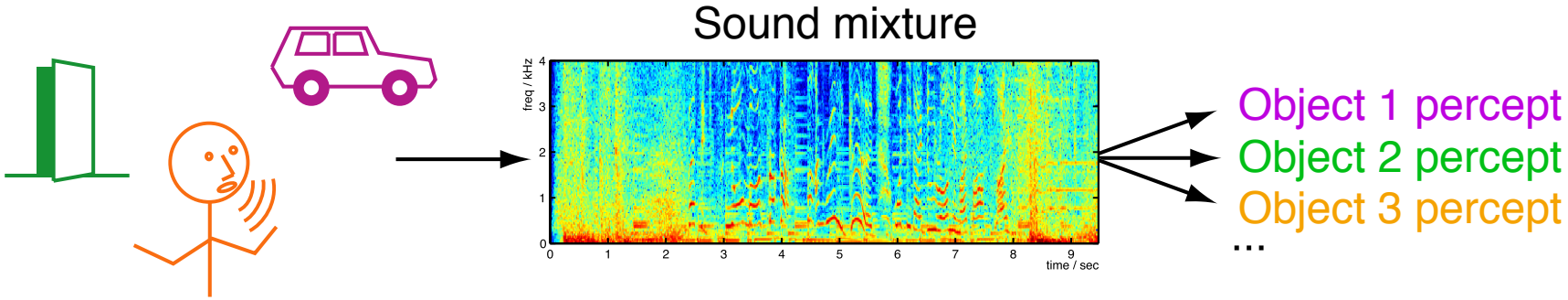
Hearium



QuietPro+

4. Future Audio Analysis & Display

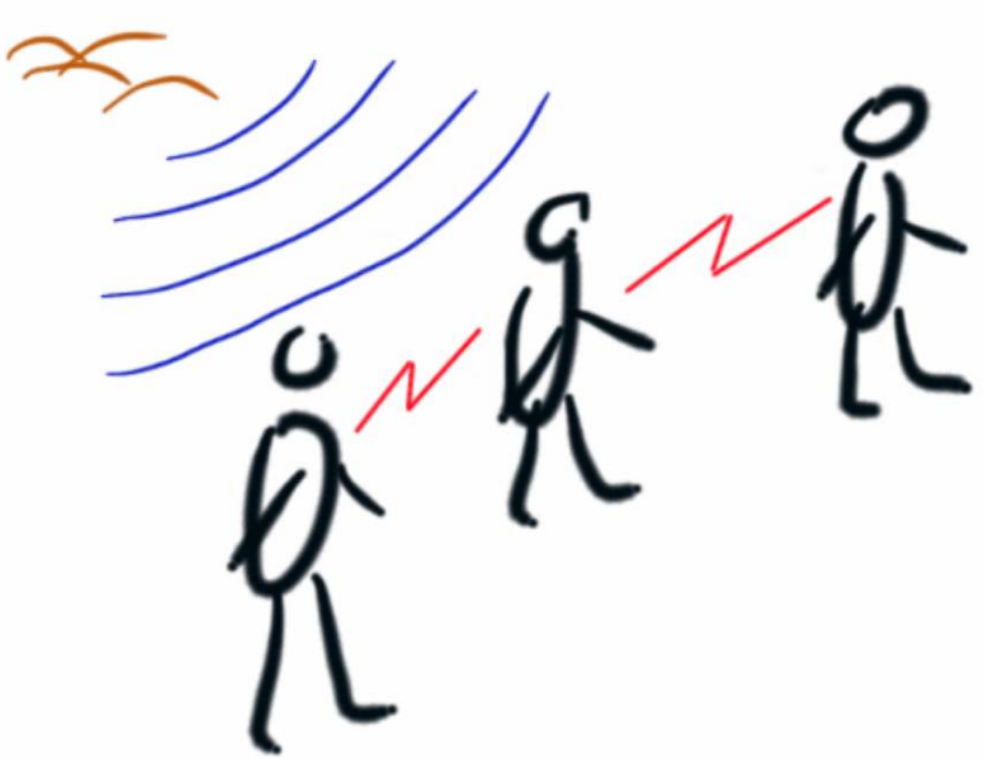
- Better Scene Analysis
  - overcoming the limitations of human hearing: sensors, geometry



- Challenges
  - fine source discrimination
  - modeling & classification (language ID?)
  - Integrating through time: single location, sparse sounds

Ad-Hoc Mic Array

- **Multiple sensors, real-time sharing**
  - long-baseline beamforming

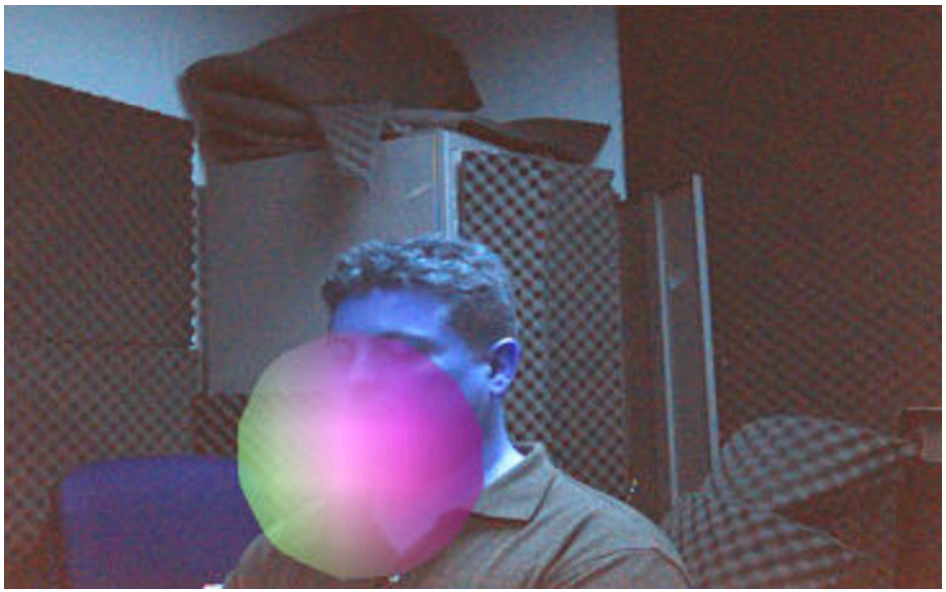


- Challenges
  - precise relative (dynamic) localization
  - precise absolute registration

Sound Visualization

O'Donovan et al. '07

- Making acoustic information visible
  - "synesthesia"

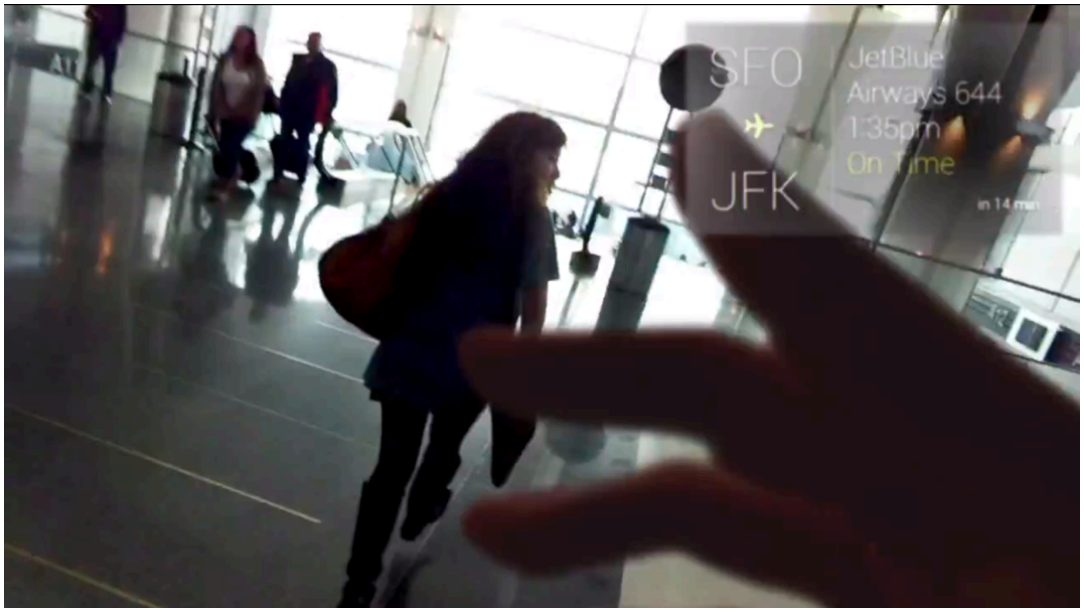


www.ultra-gunfirelocator.com

- Challenges
  - source formation & classification
  - registration: sensors, display

Auditory Display

- Acoustic channel **complements** vision
  - acoustic alarms
  - verbal information
  - "zoomed" ambience
  - instant replay



http://www.youtube.com/watch?v=v1uyQZNg2vE

- Challenges
  - Information management & prioritization
  - maximally exploit perceptual organization

Summary

- Human Scene Analysis
  Spatial & Source information

- Computational Scene Analysis
  Spatial & Source information
  World knowledge

- Augmented Audition
  Selective pass-through + insertion

References

- A. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.
- C. Darwin & R. Carlyon, "Auditory grouping" *Hbk of Percep. & Cogn. 6: Hearing*, 387–424, Academic Press, 1995.
- J. Blauert, *Spatial Hearing* (revised ed.), MIT Press, 1996.
- D. Brungart & B. Simpson, "The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal", JASA 112(2), Aug. 2002.
- D. Brungart, B. Simpson, M. Ericson, K. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," JASA 110(5), Nov. 2001.
- G. Brown & M. Cooke, "Computational auditory scene analysis," Comp. Speech & Lang. 8(4), 297–336, 1994.
- H. Okuno, T. Nakatani, T. Kawabata, "Listening to two simultaneous speeches," Speech Communication 27, 299–310, 1999.
- G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Tr. Neural Networks, 15(5), Sep. 2004.
- J. Benesty, J. Chen, Y. Huang, *Microphone Array Signal Processing*, Springer, 2008.
- A. Bell, T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," Neural Computation, vol. 7 no. 6, pp. 1129-1159, 1995.
- P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Intl. Wkshp. on Indep. & Artif.l Neural Networks, Tenerife, Feb. 1998.
- D. Ellis, "Prediction-Driven Computational Auditory Scene Analysis," Ph.D. thesis, MIT EECS, 1996.
- T Kristjansson, J Hershey, P Olsen, S Rennie, R Gopinath, Super-human multi-talker speech recognition: The IBM 2006 speech separation challenge system, Proce. of Interspeech, 97-100, 2006.
- A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, The ICSI Meeting Corpus, Proc. ICASSP-03, pp. I-364--367, 2003.
- D. Ellis and J. Liu, Speaker turn segmentation based on between-channel differences, NIST Meeting Recognition Workshop @ ICASSP, pp. 112-117, Montreal, May 2004.
- D. Ellis, X. Zheng, and J. McDermott, Classifying soundtracks with audio texture features, Proc. IEEE ICASSP, pp. 5880-5883, Prague, May 2011.
- D. Ellis and K.S. Lee, Minimal-Impact Audio-Based Personal Archives, First ACM workshop on Continuous Archiving and Recording of Personal Experiences CARPE-04, New York, pp. 39-47, Oct 2004.
- C. P. Brown, R. O. Duda, A structural model for binaural sound synthesis, IEEE Tr. Speech & Audio 6(5):476-488, 1998.
- A. Härma, J. Julia, M. Tikander, M. Karjalainen, T. Lokki, J, Hiipakka, G. Lorho, Augmented reality audio for mobile and wearable appliances, J. Aud Eng. Soc. 52(6): 618-639, 2004.
- A. O'Donovan, R. Duraiswami, J. Neumann, Microphone arrays as generalized cameras for integrated audio visual processing, IEEE CVPR, 1-8, 2007.