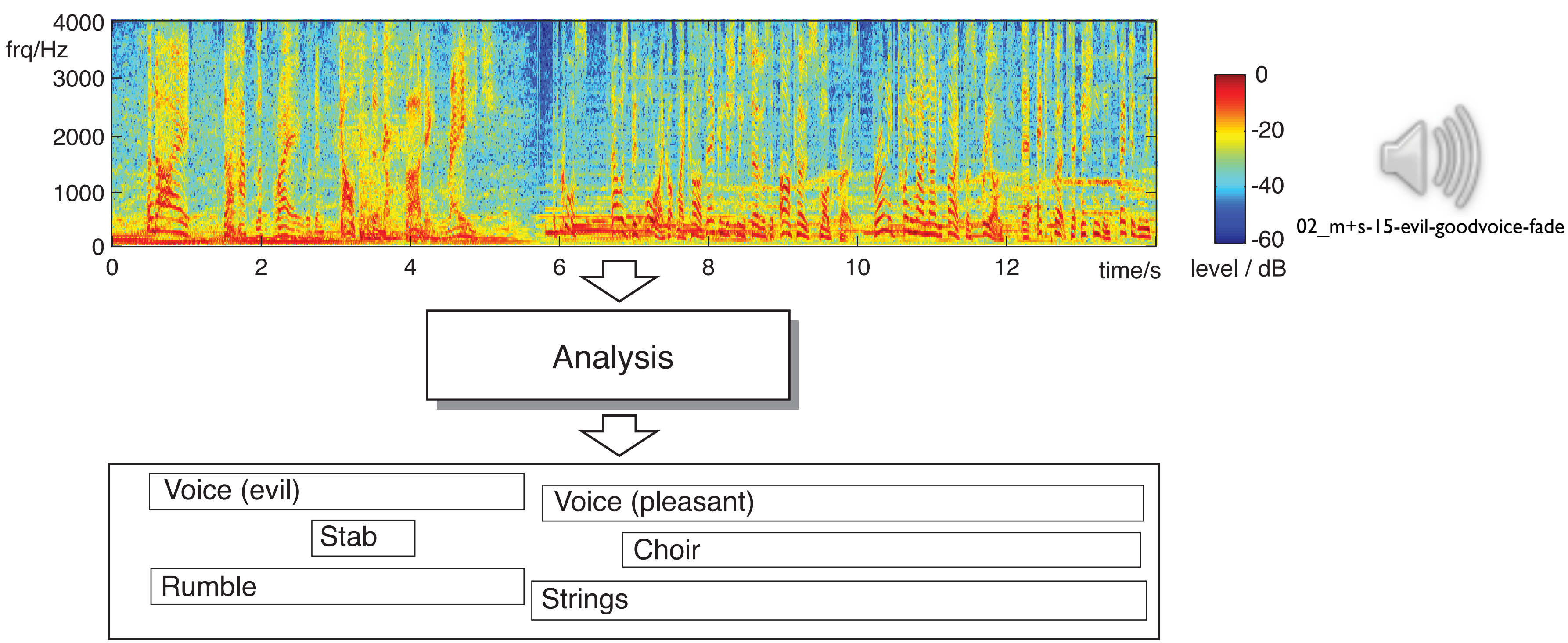
Lecture 14: Source Separation

1. Sources, Mixtures, & Perception
2. Spatial Filtering
3. Time-Frequency Masking
4. Model-Based Separation

Dan Ellis
Dept. Electrical Engineering, Columbia University
dpwe@ee.columbia.edu   http://www.ee.columbia.edu/~dpwe/e4896/

1. Sources, Mixtures, & Perception

- Sound is a linear process (superposition)
  - no "opacity" (unlike vision)
  - sources → "auditory scenes" (polyphony)

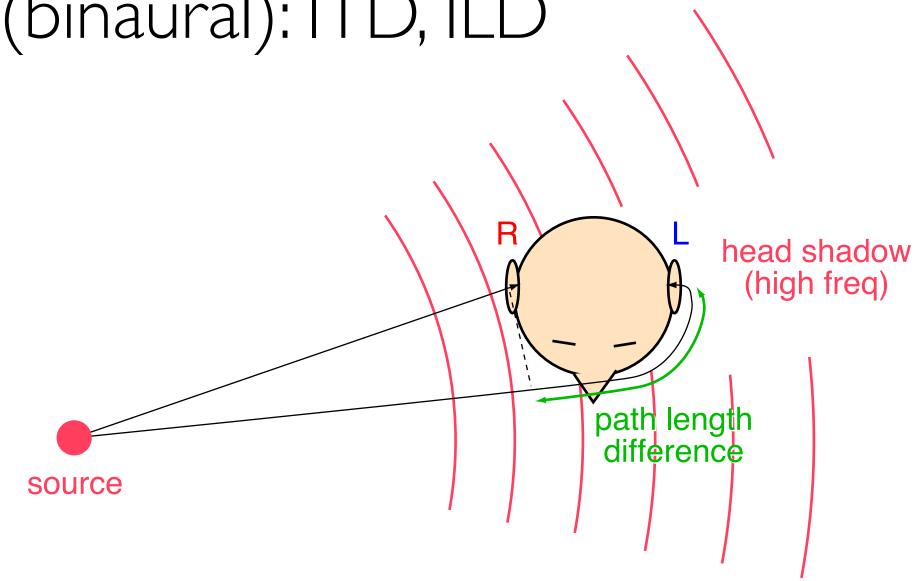


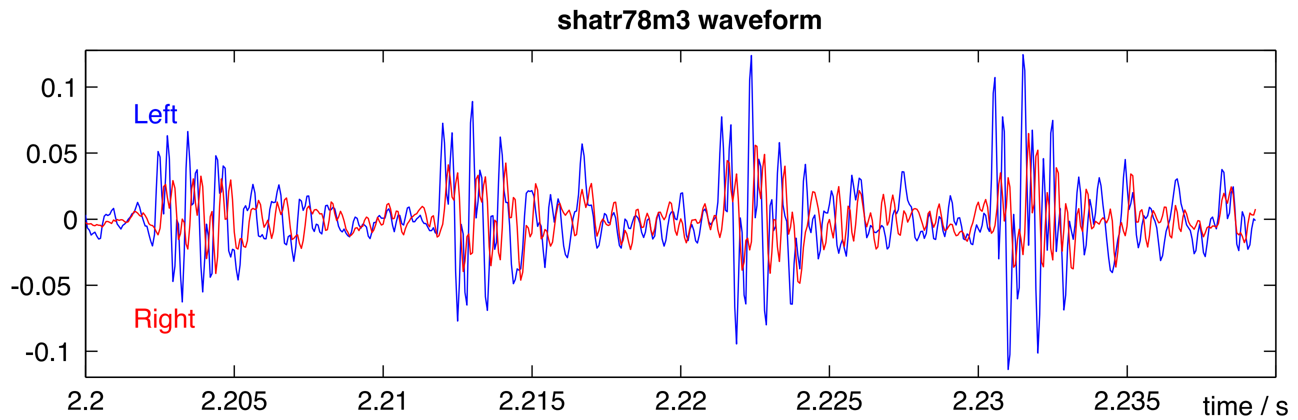
- Humans perceive discrete sources
  - .. a subjective construct

Spatial Hearing

Blauert '96

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



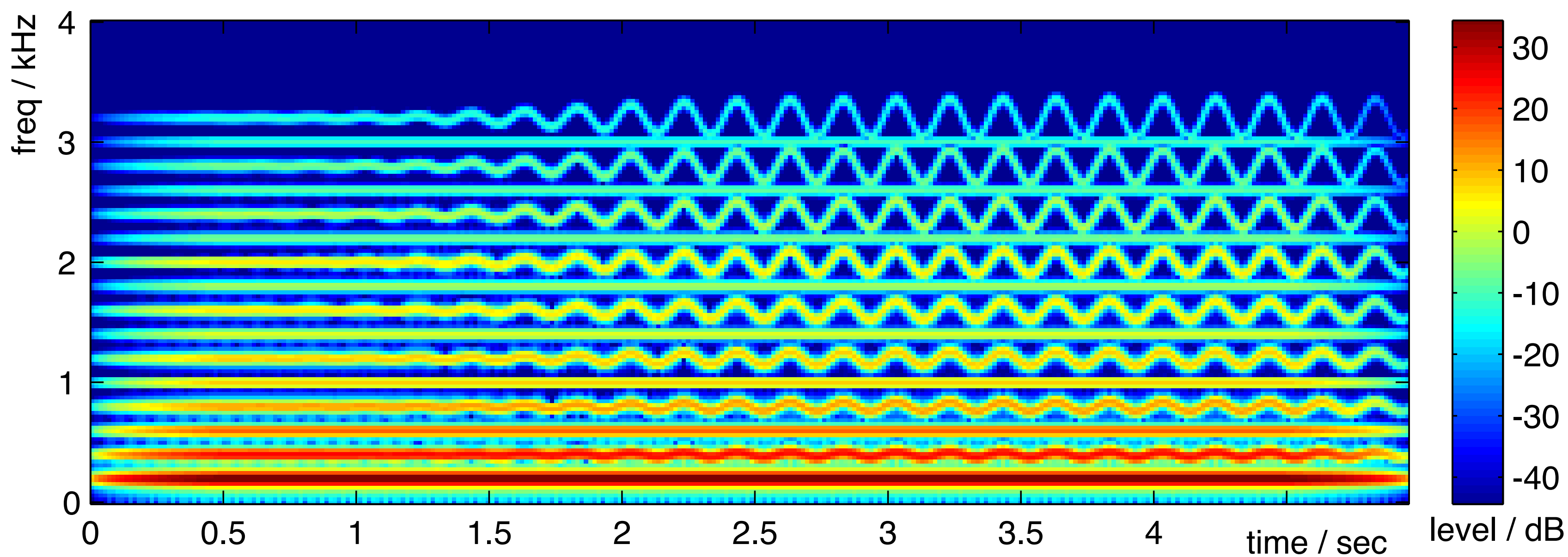
- 

Auditory Scene Analysis

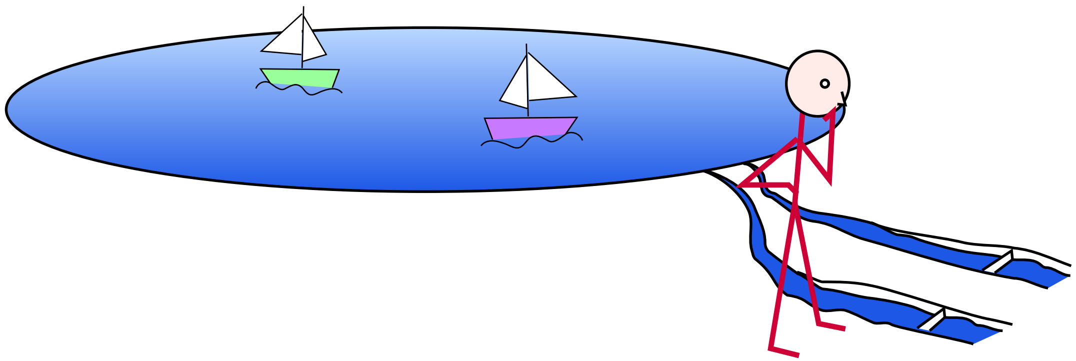
Bregman '90

- Spatial cues may not be enough/available
  - single channel signal
- Brain uses signal-intrinsic cues to form sources
  - onset, harmonicity

Reynolds-McAdams Oboe



Auditory Scene Analysis

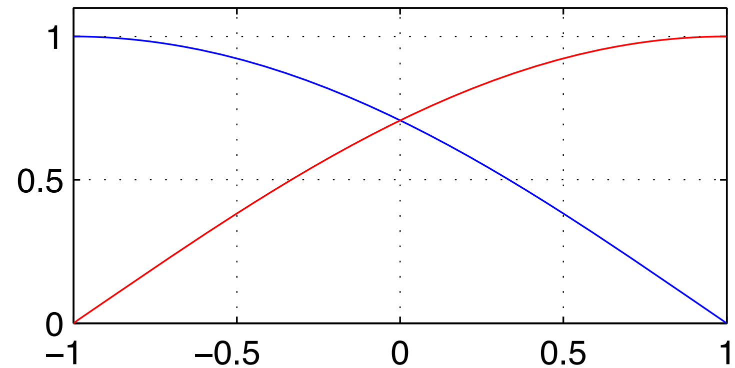


"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

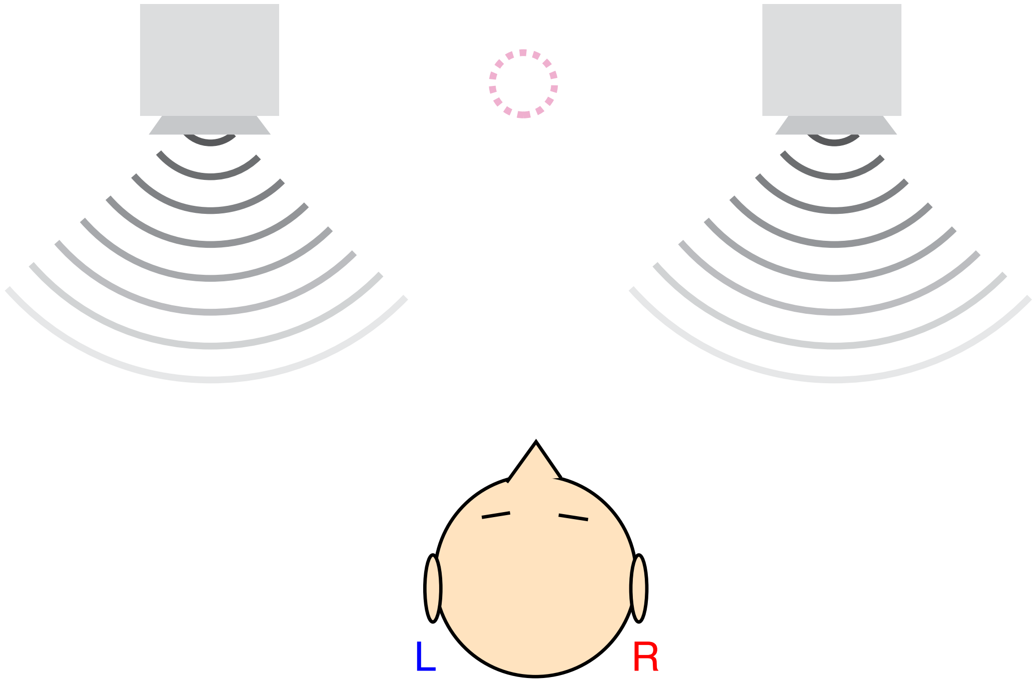
- Quite a challenge!

Audio Mixing

- Studio recording combines separate tracks into, e.g., 2 channels (stereo)
  - different levels
  - panning
  - other effects



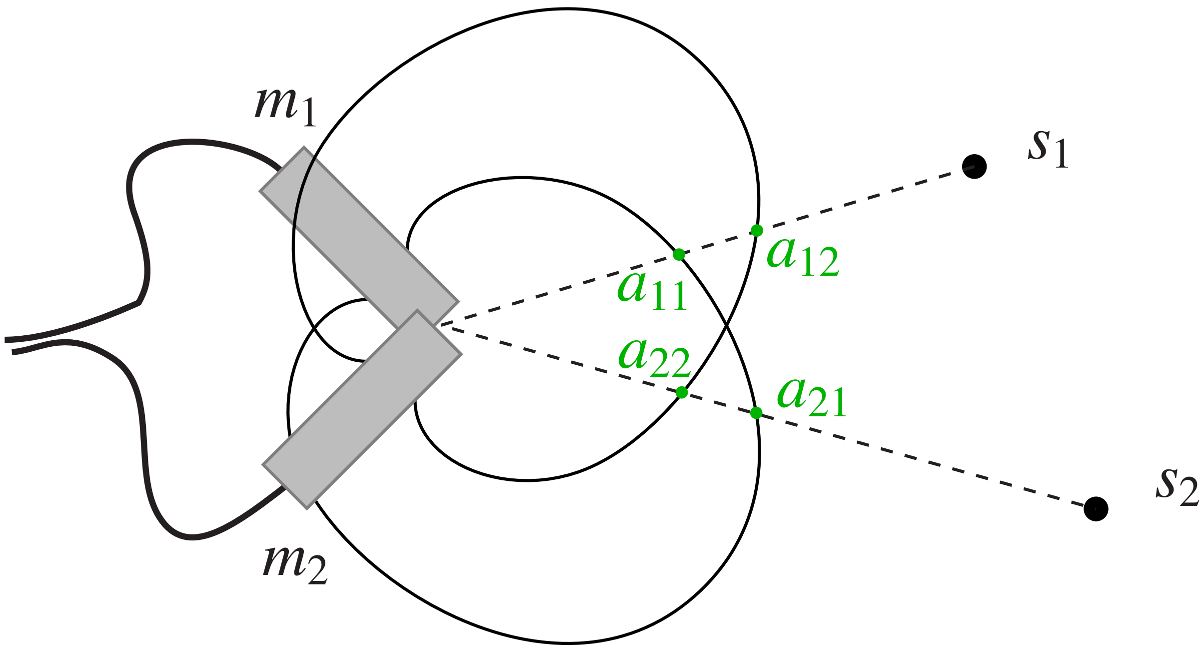
- Stereo Intensity Panning
  - manipulating ILD only
  - constant power
  - more channels: use just nearest pair?



2. Spatial Filtering

- N sources detected by M sensors
  - degrees of freedom
  - (else need other constraints)

- Consider 2 x 2 case:
  - directional mics



  - → mixing matrix:

$$\begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \Rightarrow \begin{bmatrix} \hat{s}_1 \\ \hat{s}_2 \end{bmatrix} = \hat{A}^{-1} m$$

Source Cancelation

- Simple 2 x 2 case example:

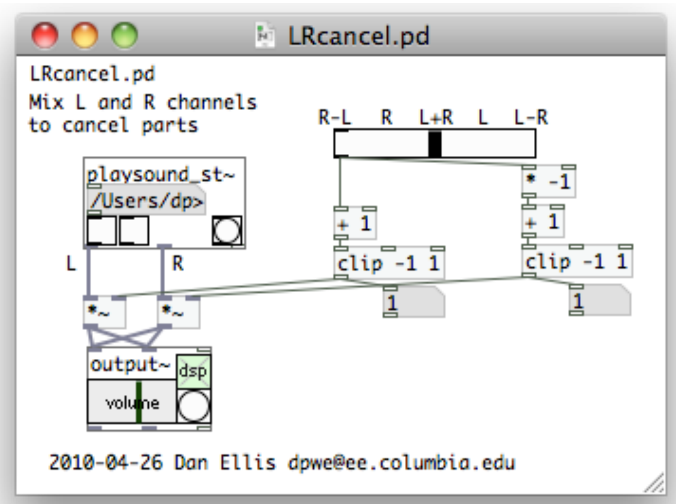
$$\begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 \\ 0.8 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$$

$$m_1(t) = s_1(t) + 0.5 s_2(t)$$

$$m_2(t) = 0.8 s_1(t) + s_2(t)$$

$$\Rightarrow m_1(t) - 0.5 m_2(t) = 0.6 s_1(t)$$

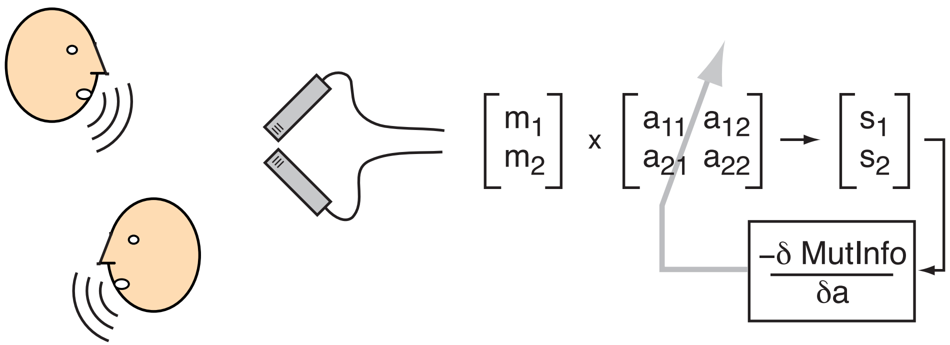
- if no delay and linearly-independent sums, can cancel one source per combination



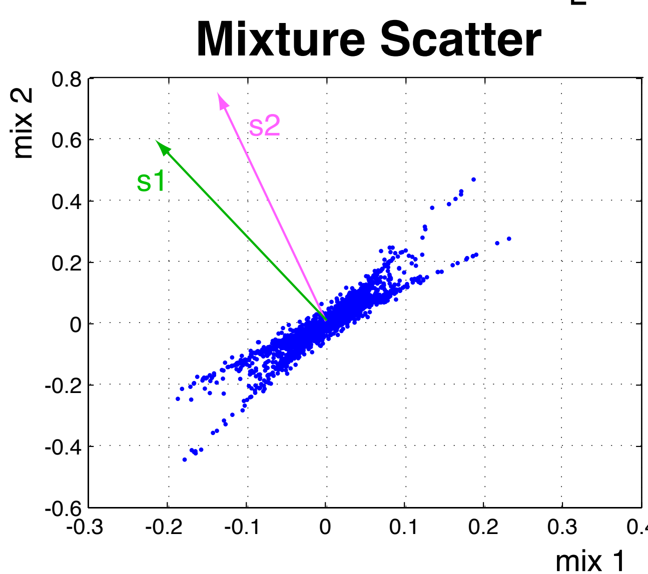
Independent Component Analysis

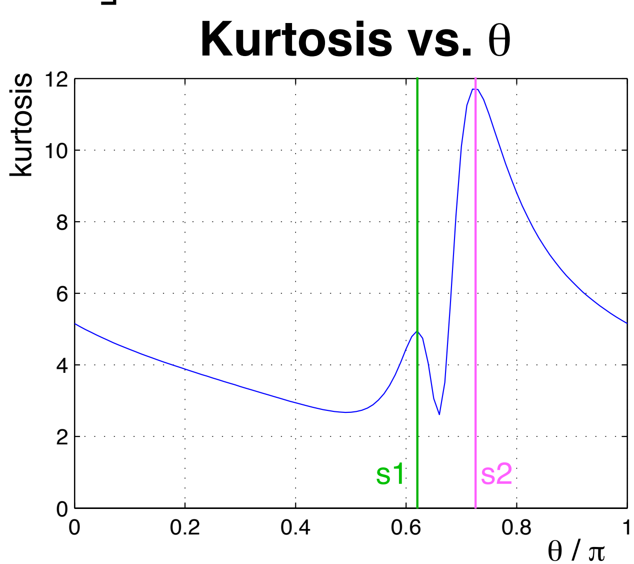
Bell & Sejnowski '95

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



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

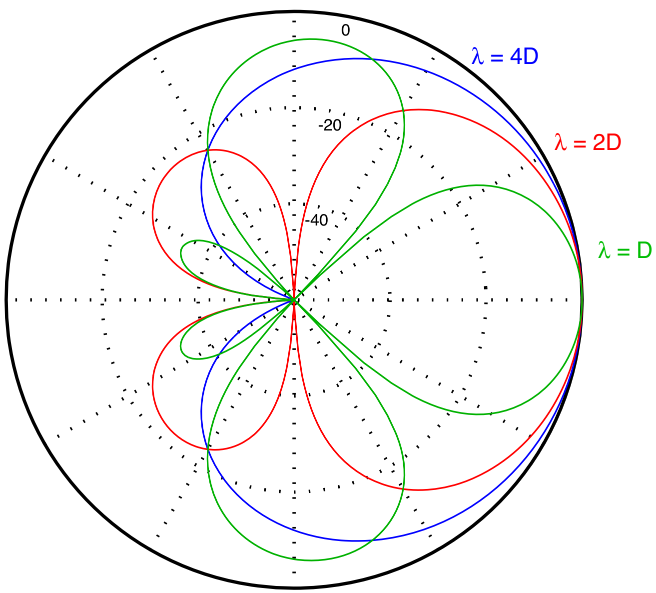


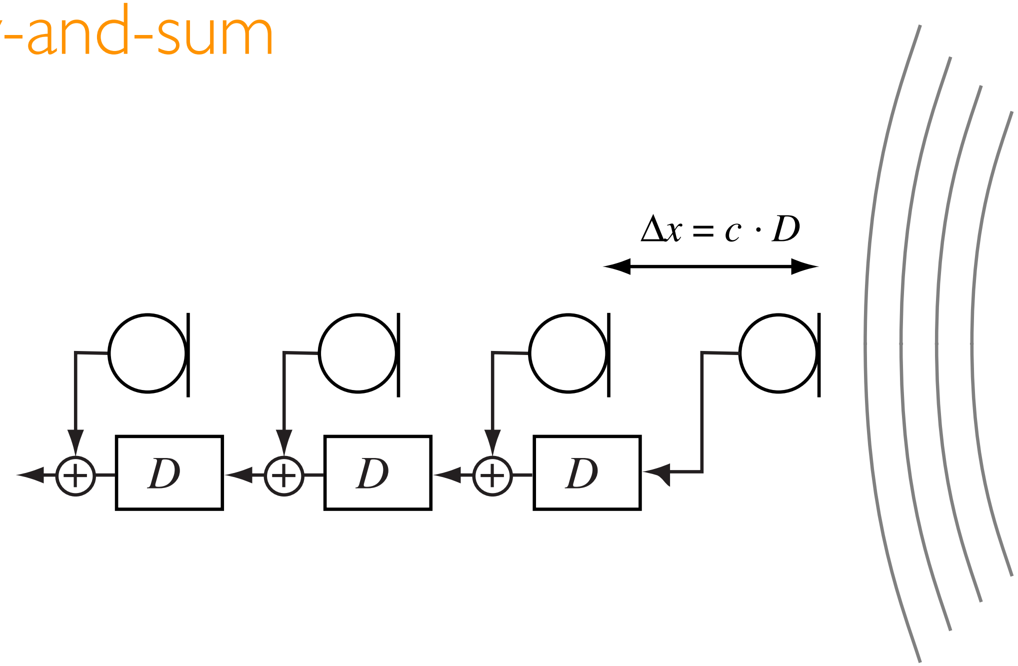


Microphone Arrays

Benesty, Chen, Huang '08

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



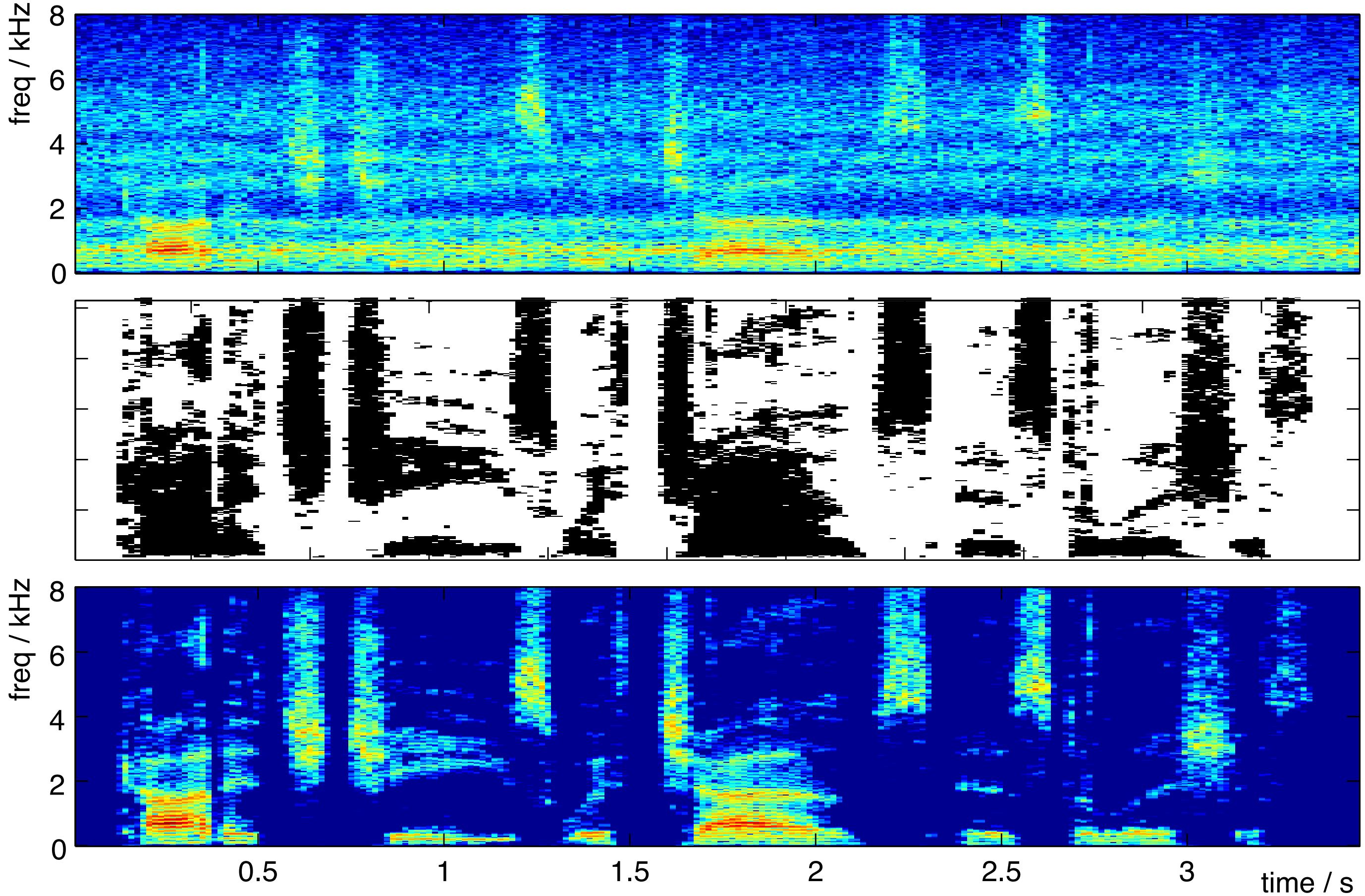


  - off-axis spectral coloration
  - many variants - filter & sum, sidelobe cancelation ...

3. Time-Frequency Masking

Brown & Cooke '94
Roweis '01

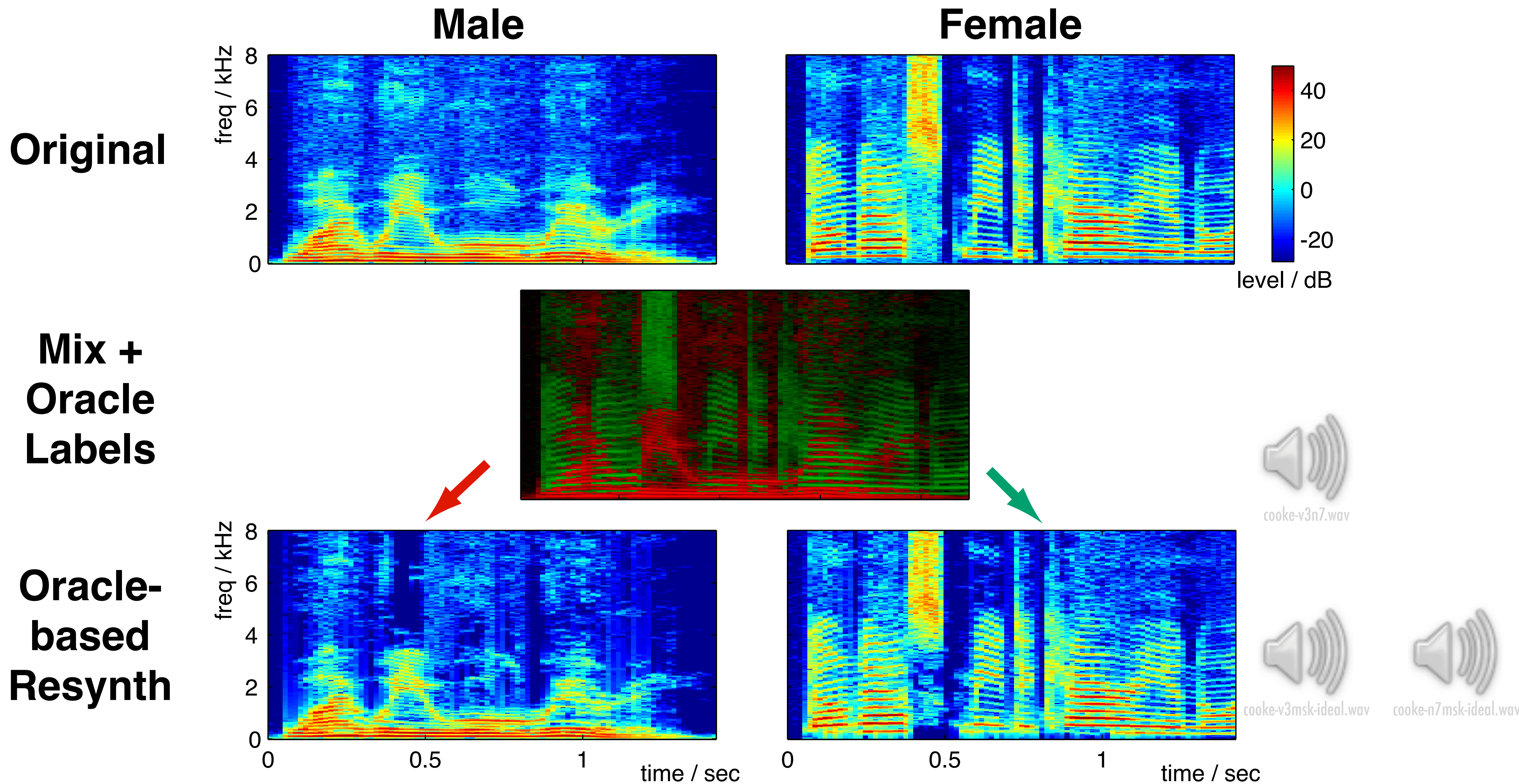
- What if there is only one channel?
  - cannot have fixed cancellation
  - but could have fast time-varying filtering:



- The trick is finding the right mask...

Time-Frequency Masking

- Works well for overlapping voices

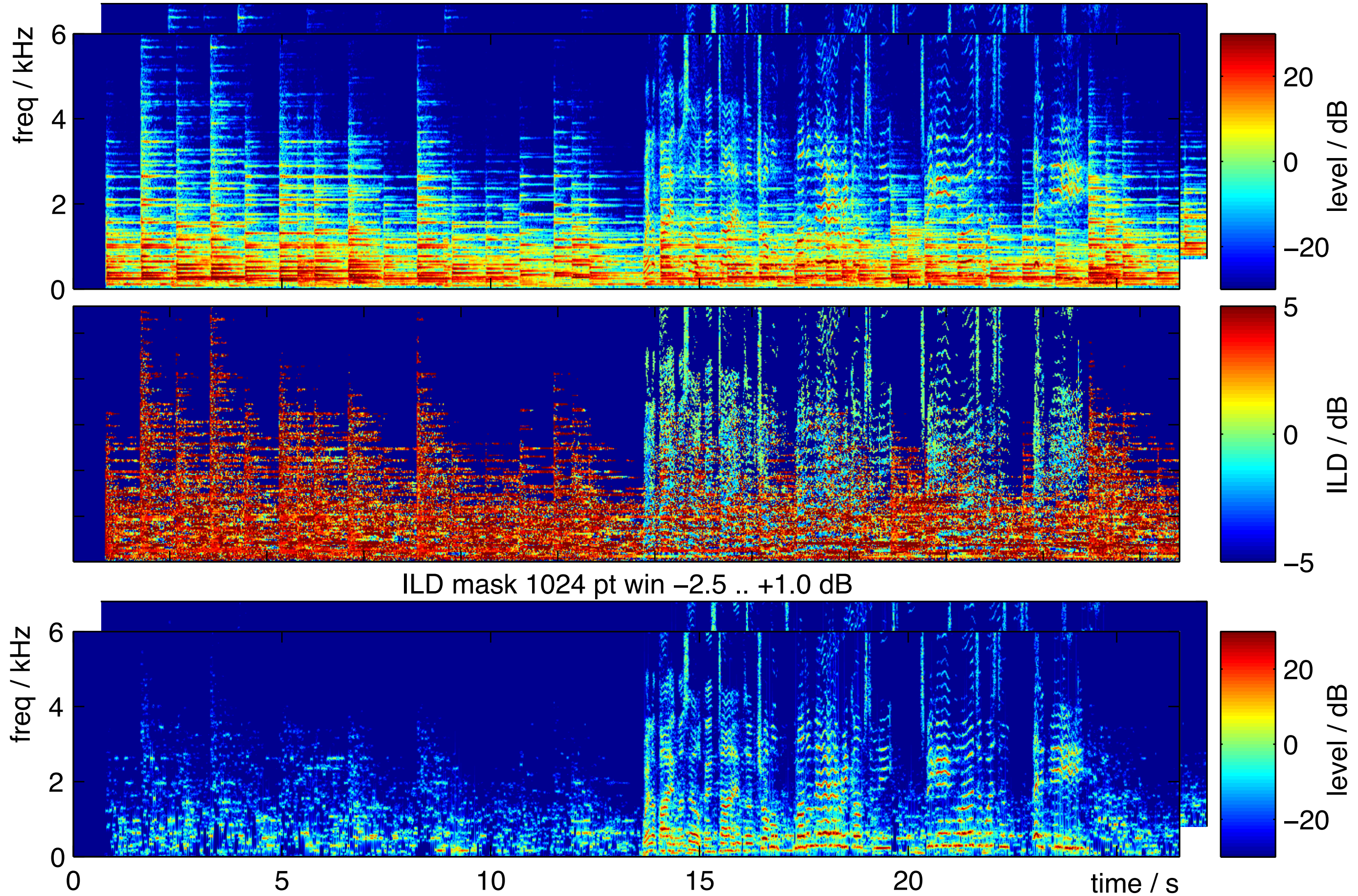


  - time-frequency resolution?

Pan-Based Filtering

Avendano 2003

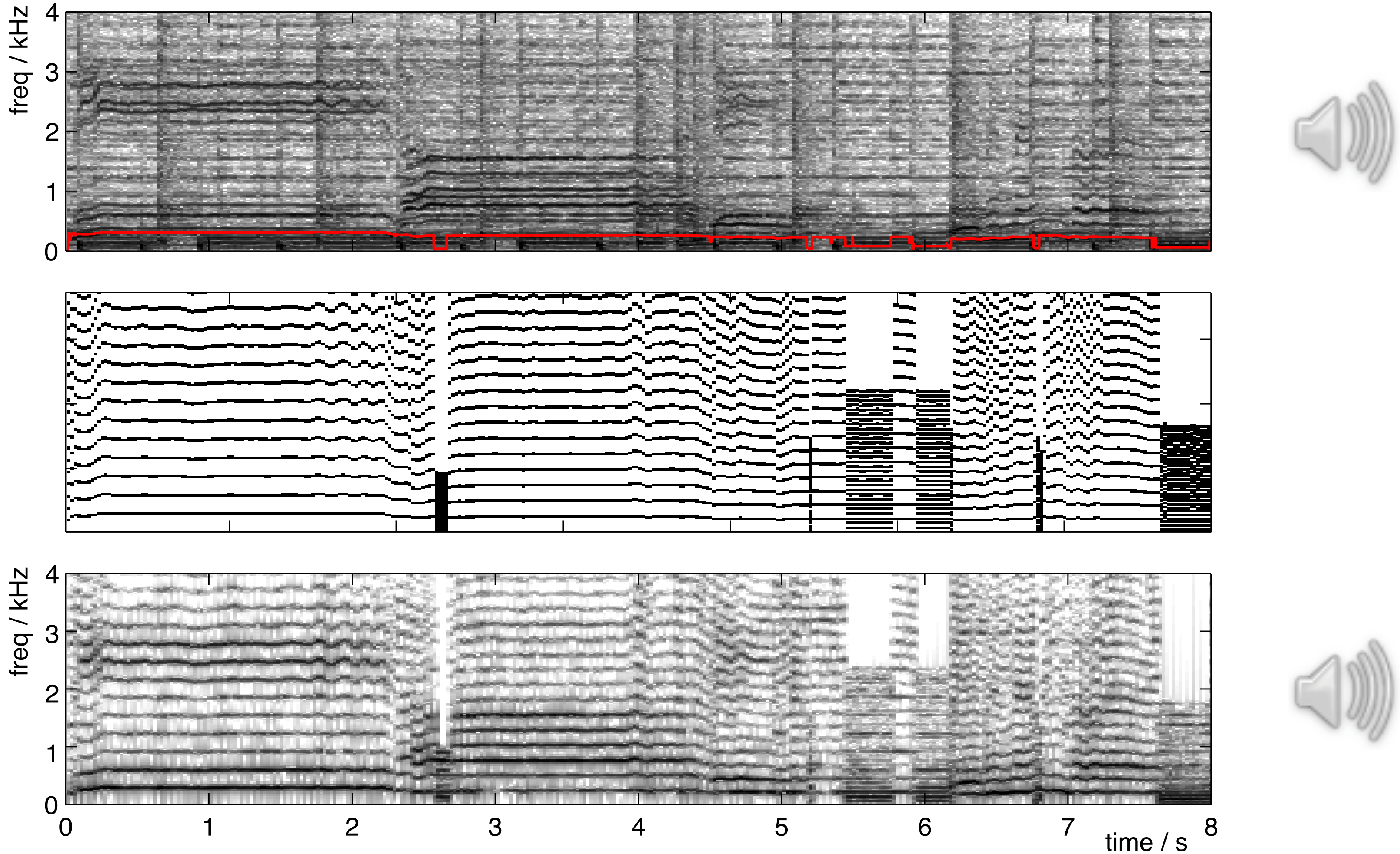
- Can use time-frequency masking even for stereo
  - e.g. calculate "panning index" as ILD
  - mask cells matching that ILD



Harmonic-based Masking

Denbigh & Zhao 1992

- Time-frequency masking
  can be used to pick out harmonics
  - given pitch track, know where to expect harmonics



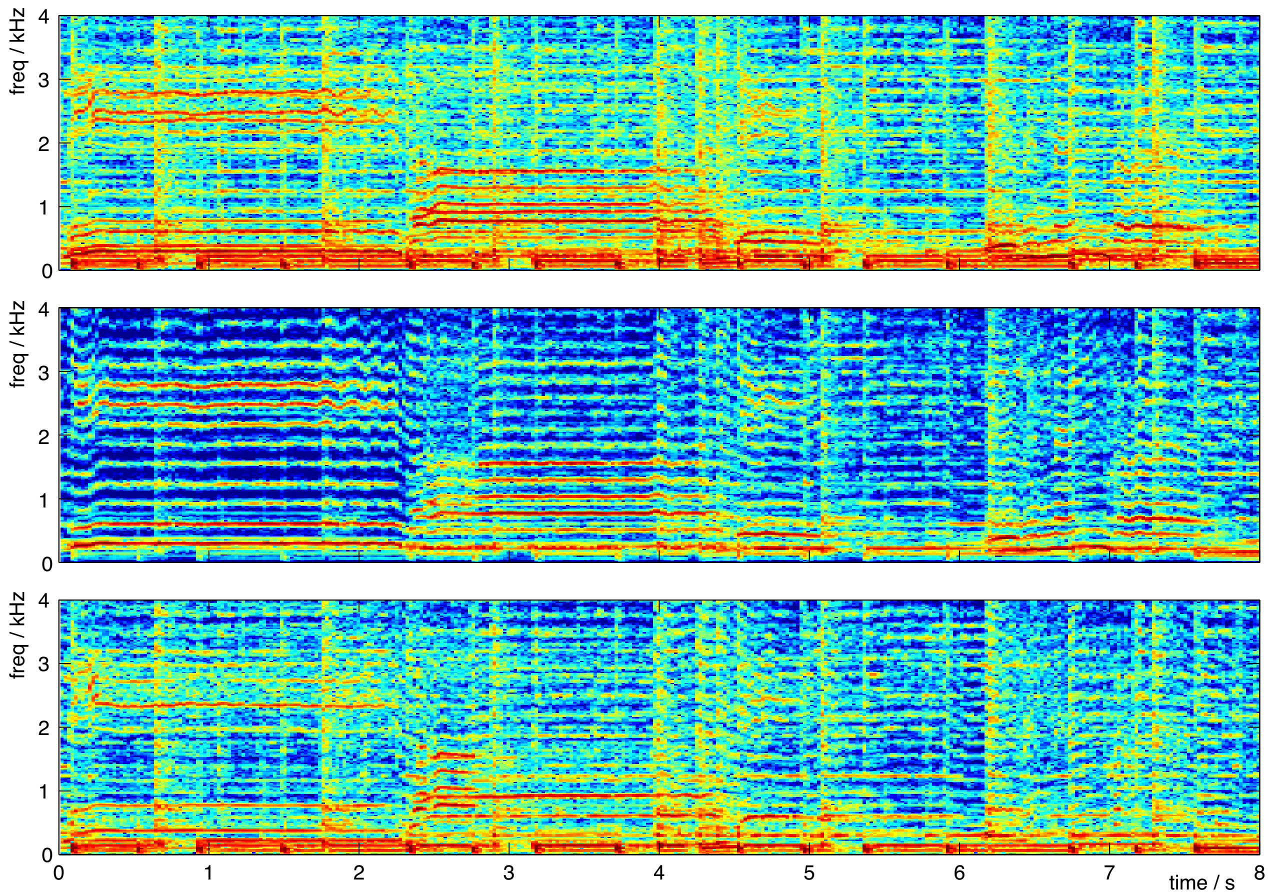
Harmonic Filtering

Avery Wang 1995

- Given pitch track, could use time-varying comb filter to get harmonics
  - or: isolate each harmonic by heterodyning:

$$\hat{x}(t) = \sum_k \hat{a}_k(t) cos(k\hat{\omega}_0(t)t)$$

$$\hat{a}_k(t) = LPF\{|x(t)e^{-jk\hat{\omega}_0(t)t}|\}$$



Nonnegative Matrix Factorization

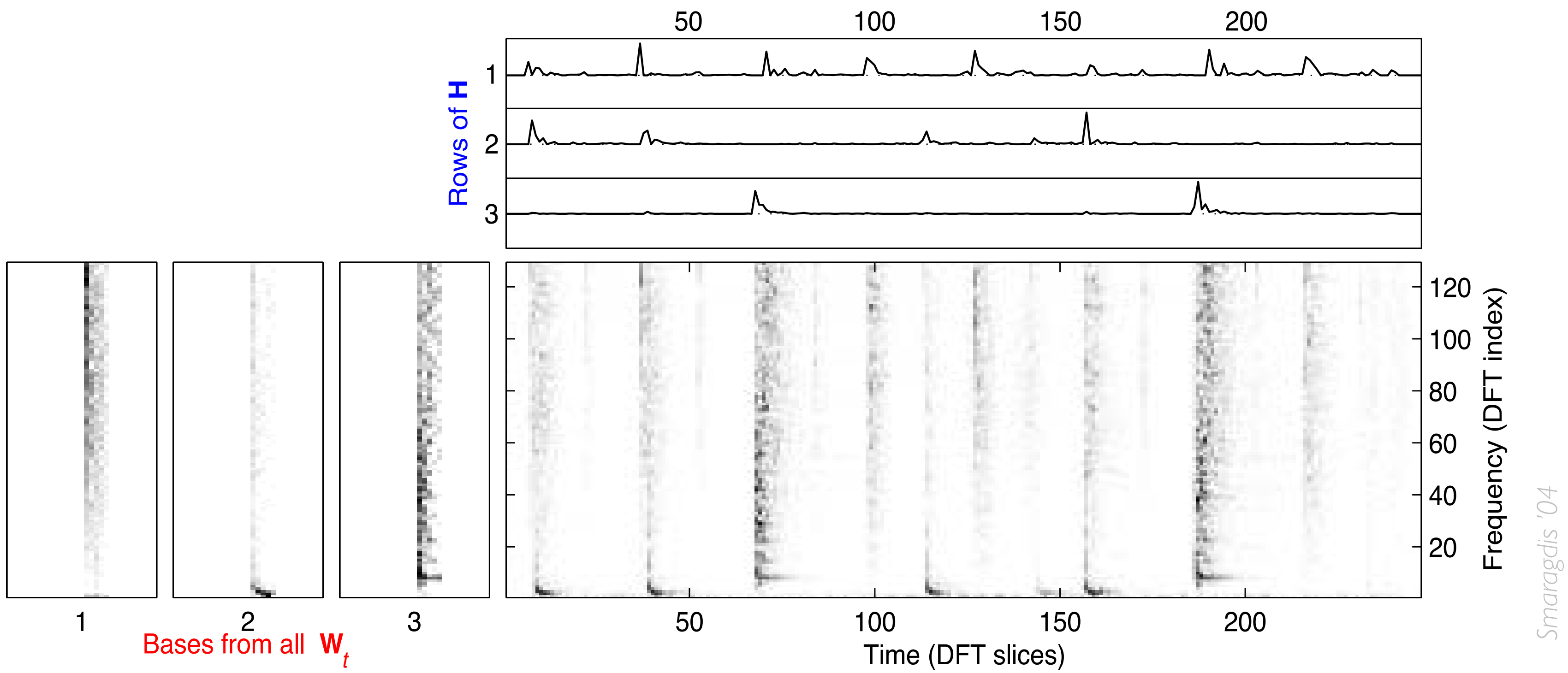
Lee & Seung '99
Abdallah & Plumbley '04
Smaragdis & Brown '03
Virtanen '07

- 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



4. Model-Based Separation

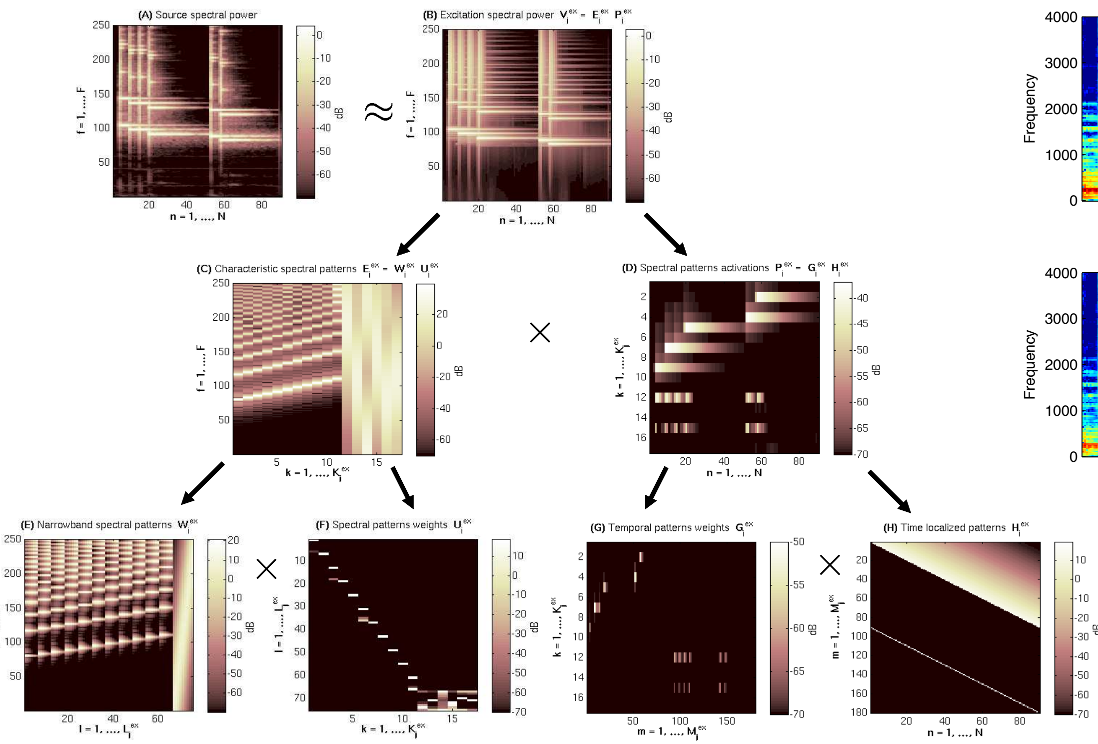
- When N (sources) > M (sensors), need additional constraints to solve problem
  - e.g. assumption of single dominant pitch
- Can assemble into a model M of source s_i
  - defines set of "possible" waveforms
  - ..probabilistically: $Pr(s_i|M)$
- Source separation from mixture as inference:
  - $$\mathbf{s} = \{s_i\} = \arg\max_{\mathbf{s}} Pr(\mathbf{x}|\mathbf{s}, A) P(A) \prod_i Pr(s_i|M)$$

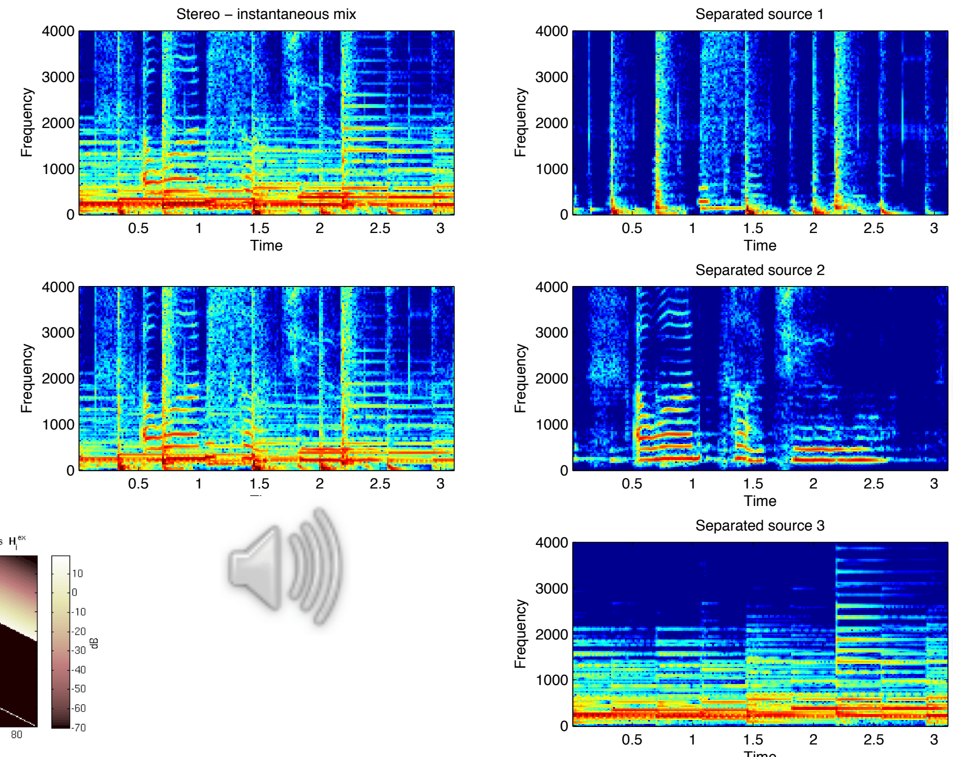
  where $Pr(\mathbf{x}|\mathbf{s}, A) = \mathcal{N}(\mathbf{x}|A\mathbf{s}, \nu)$

Source Models

Ozerov, Vincent & Bimbot 2010
http://bass-db.gforge.inria.fr/fasst/

- **Can constrain:**
    - source spectra (e.g. harmonic, noisy, smooth)
    - temporal evolution (piecewise-continuous)
    - spatial arrangements (point-source, diffuse)
- **Factored decomposition:**





Music: Shannon Hurley / Mix: Michel Desnoues & Alexey Ozerov / Separations: Alexey Ozerov

Summary

- Acoustic Source Mixtures
  The normal situation in real-world sounds
- Spatial filtering
  Canceling sources by subtracting channels
- Time-Frequency Masking
  Selecting spectrogram cells
- Model-Based Separation
  Exploiting regularities in source signals

References

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

C. Avendano, "Frequency-domain source identification and manipulation in stereo mixes for enhancement, suppression and re-panning applications," IEEE WASPAA, Mohonk, pp. 55-58, 2003.

A. Bell, T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7 no. 6, pp. 1129-1159, 1995.

J. Benesty, J. Chen, Y. Huang, *Microphone Array Signal Processing*, Springer, 2008.

J. Blauert, *Spatial Hearing*, MIT Press, 1996.

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

G. Brown & M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8 no. 4, pp. 297-336, 1994.

P. Denbigh & J. Zhao, "Pitch extraction and separation of overlapping speech," *Speech Communication*, vol. 11 no. 2-3, pp. 119-125, 1992.

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

A. Ozerov, E. Vincent, & F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation," INRIA Tech. Rep. 7453, Nov. 2010.

S. Roweis, "One microphone source separation," *Adv. Neural Info. Proc. Sys.*, pp. 793-799, 2001.

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

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, 2007.

Avery Wang, *Instantaneous and frequency-warped signal processing techniques for auditory source separation*, Ph.D. dissertation, Stanford CCRMA, 1995.