Speech Separation in Humans and Machines

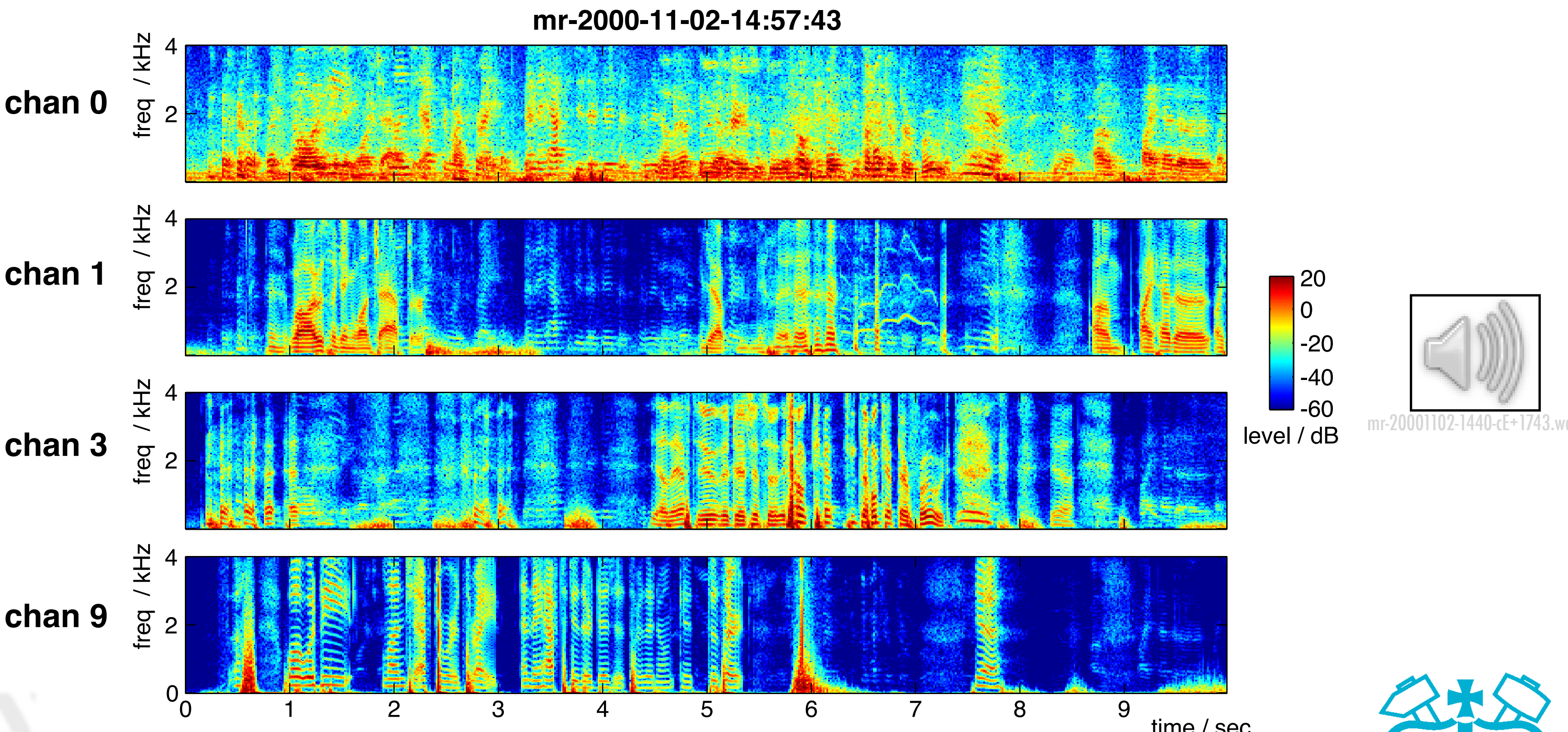
Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

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

1. The Speech Separation Problem
2. Human Performance
3. Source Separation
4. Source Inference
5. Concluding Remarks

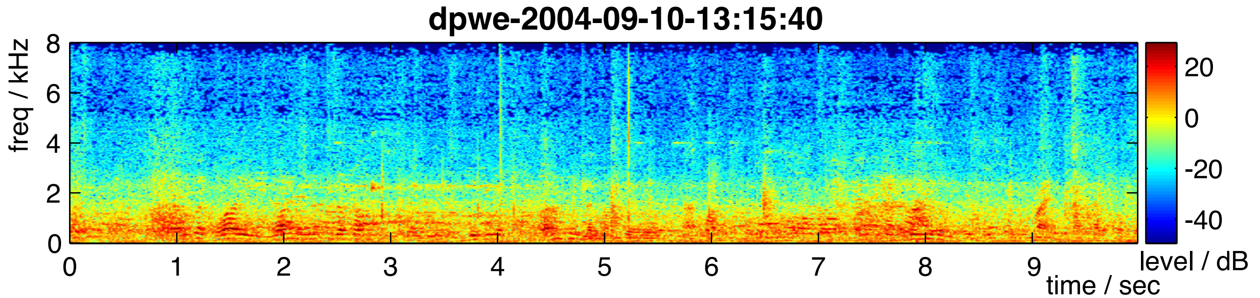
1. Speech Separation

- Speech rarely occurs in isolation
  - .. but recognizing mixed speech is a problem
  - .. for humans and machines



Speech Separation Scenarios

- Interactive voice systems
  - human-level understanding is expected
- Speech prostheses
  - crowds: #1 complaint of hearing aid users
- Archive analysis
  - identifying and isolating speech



pa-2004-09-10-131540.wav

- Surveillance...

How Can We Separate?

- By between-sensor differences (spatial cues)
  - 'steer a null' onto a compact interfering source
- By finding a 'separable representation'
  - spectral? but speech is broadband
  - periodicity? maybe – for voiced speech
  - something more signal-specific...
- By inference (based on knowledge/models)
  - speech is redundant

    → use part to guess the remainder

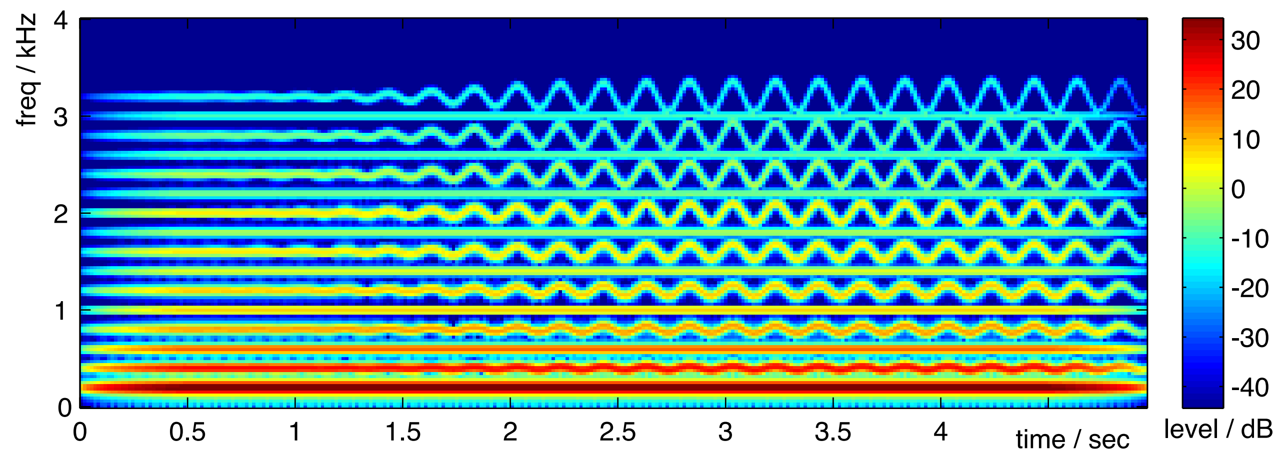
Outline

1. The Speech Separation problem
2. Human Performance
   - scene analysis
   - speech separation by location
   - speech separation by voice characteristics
3. Source Separation
4. Source Inference
5. Concluding Remarks

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 oboe

reynolds-mcadams-dpwe.wav

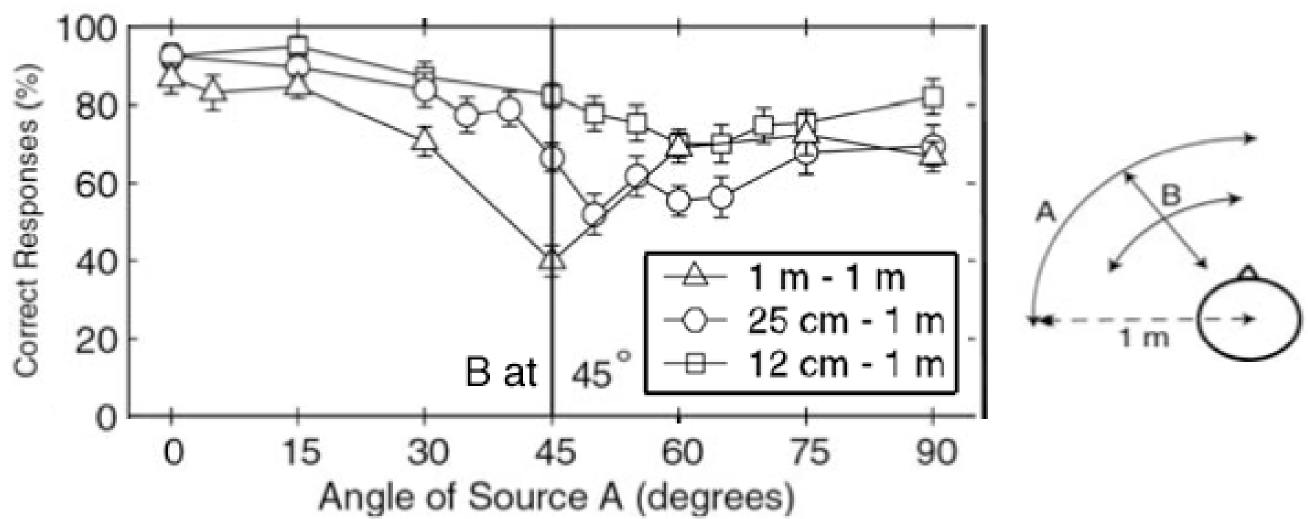
Speech Mixtures: Spatial Separation

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:**

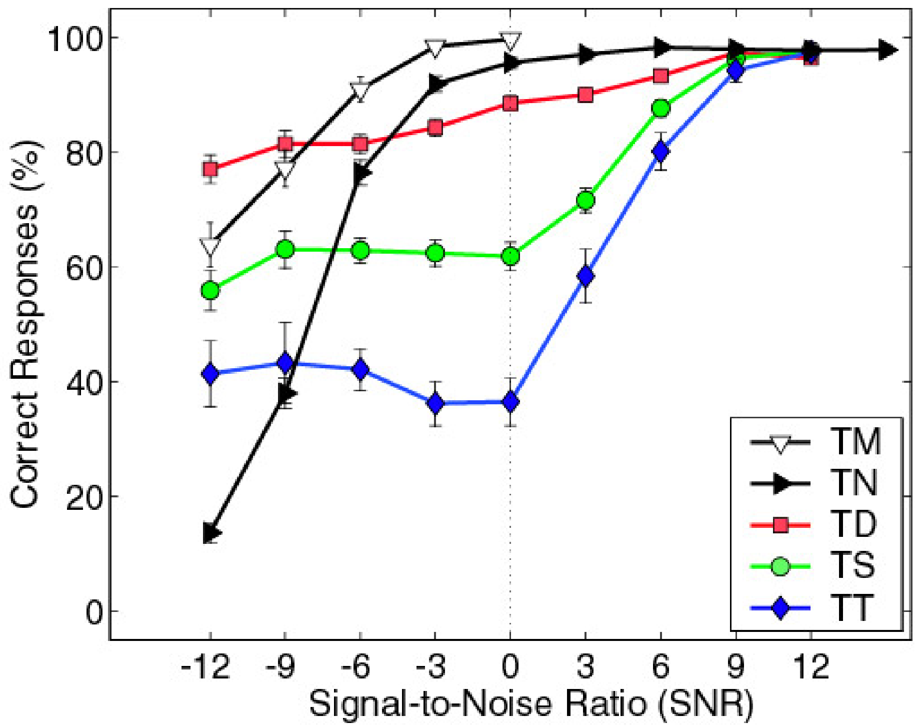


  - A, B same speaker

Separation by Vocal Differences

Brungart et al.'01

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

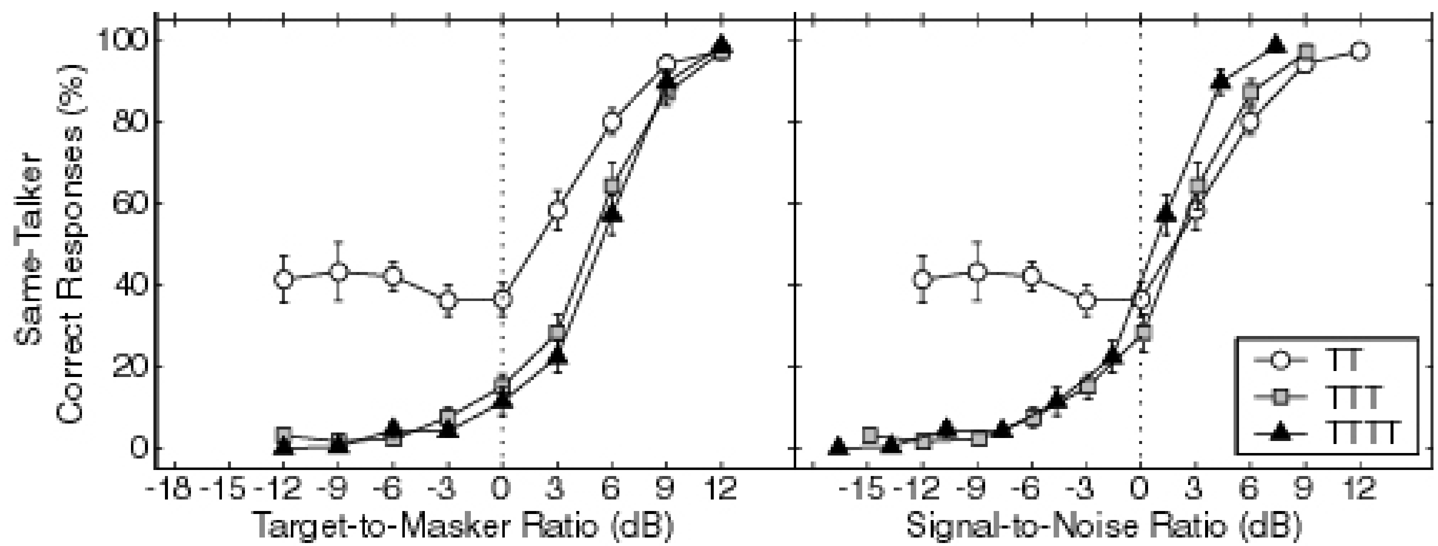


  - energetic vs. informational masking

Varying the Number of Voices

Brungart et al.'01

- Two voices OK;
  More than two voices harder
  - (same spatial origin)



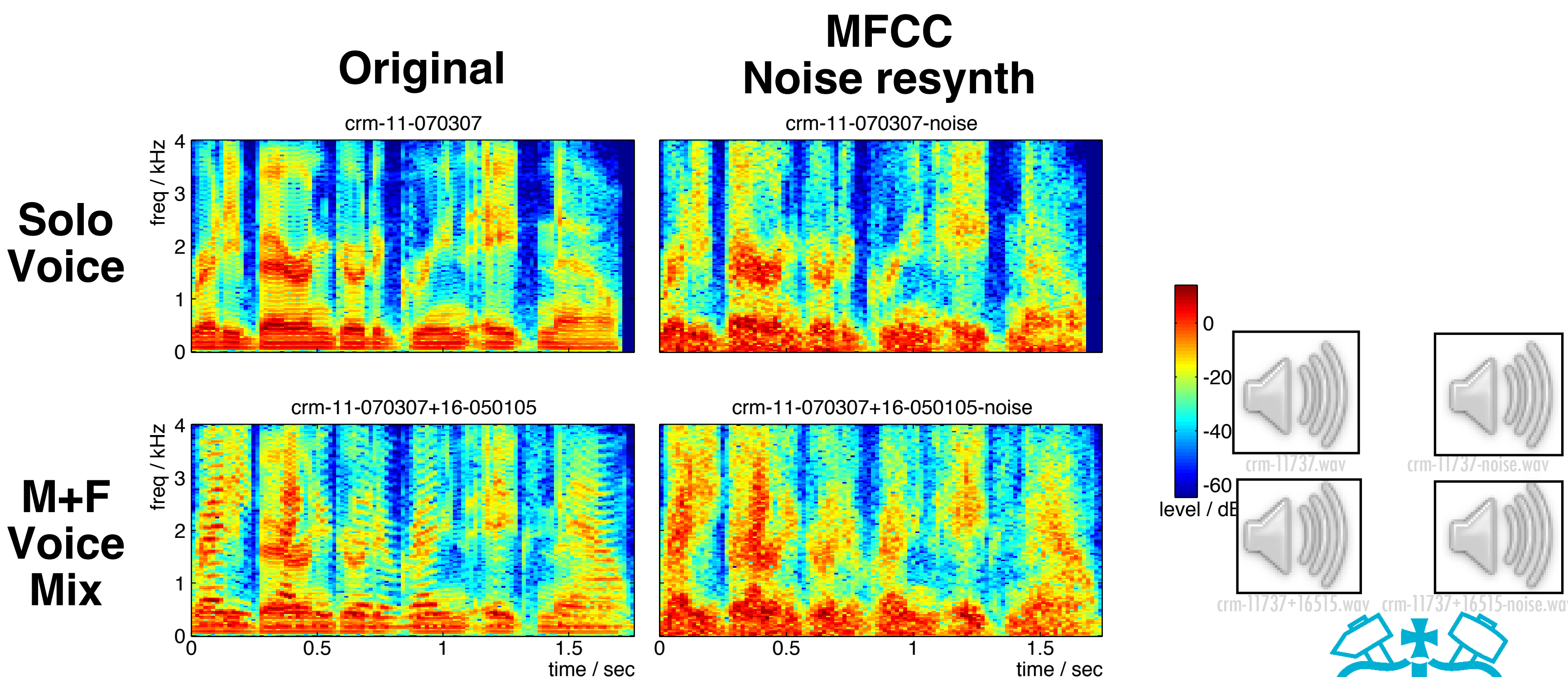
  - mix of *N* voices tends to speech-shaped noise...

Outline

1. The Speech Separation problem
2. Human Performance
3. **Source Separation**
   - Independent Component Analysis
   - Computational Auditory Scene Analysis
4. Source Inference
5. Concluding Remarks

Machine Separation

- Problem: Features of combinations are not combinations of features
  - voice is easy to characterize when in isolation
  - redundancy needed for real-world communication



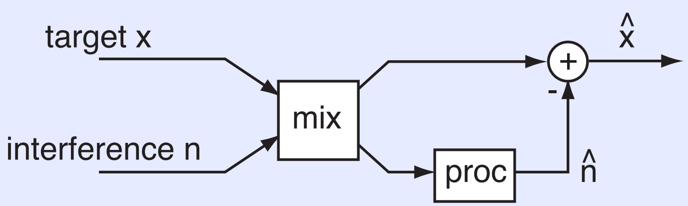
Lab ROSA
Laboratory for the Recognition and Organization of Speech and Audio

Columbia University in the City of New York

Separation Approaches

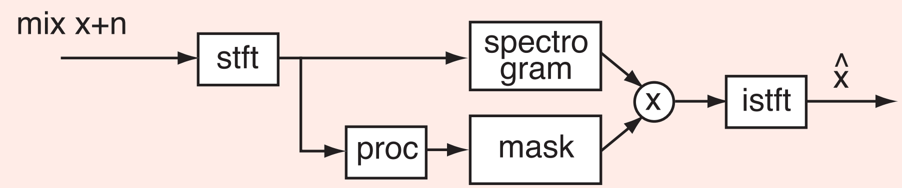
ICA

- Multi-channel
- Fixed filtering
- Perfect separation
  - maybe!



CASA / Model-based

- Single-channel
- Time-varying filtering
- Approximate Separation

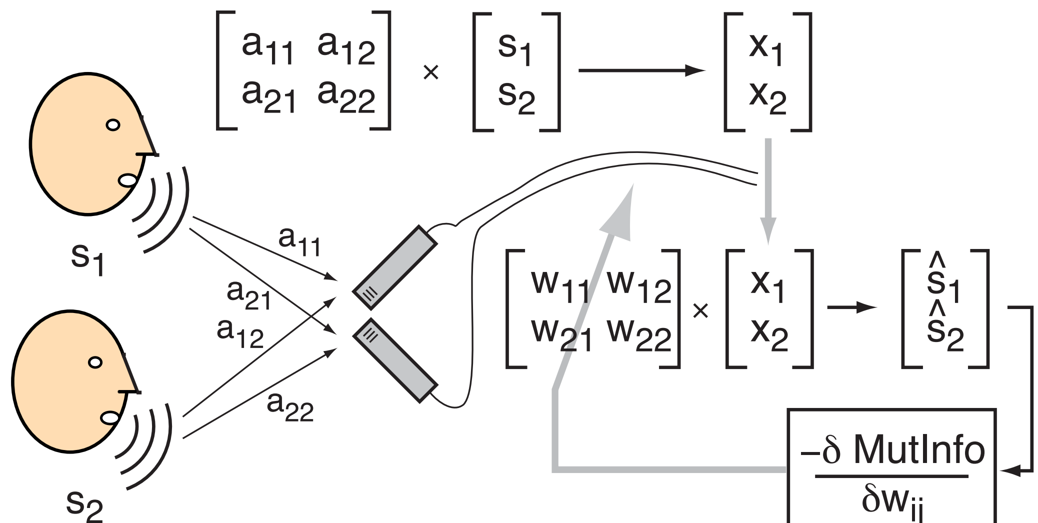


- **Very different approaches!**

Independent Component Analysis

Bell & Sejnowski'95
Smaragdis'98
...

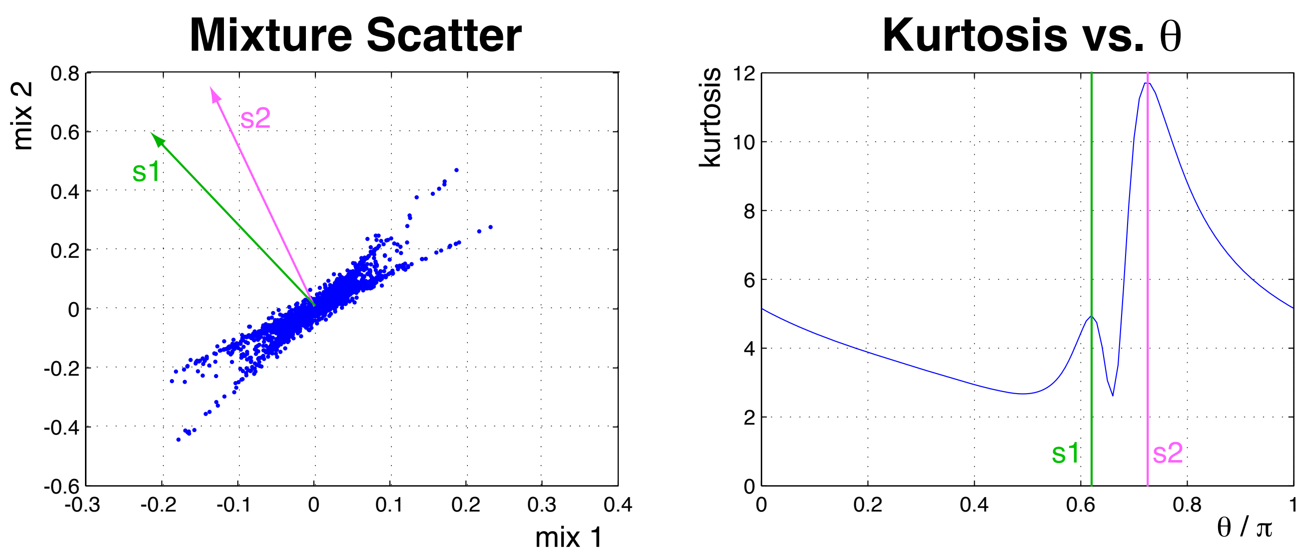
- Central idea:
  Search unmixing space
  to maximize independence of outputs



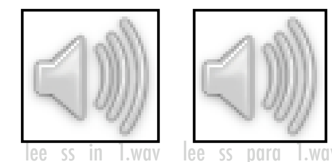
- simple mixing
  → a good solution (usually) exists

ICA Limitations

- Cancellation is very finicky
  - hard to get more than ~ 10 dB rejection



from Parra & Spence'00

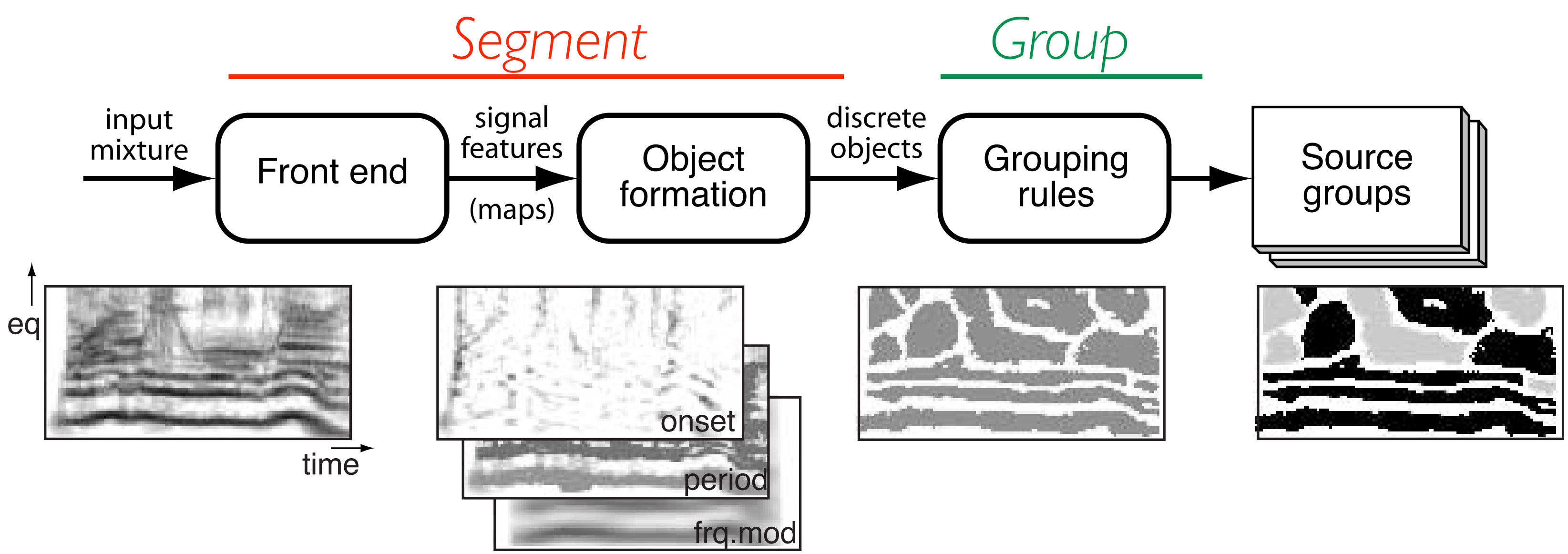


- The world is not instantaneous, fixed, linear
  - subband models for reverberation
  - continuous adaptation
- Needs spatially-compact interfering sources

Computational Auditory Scene Analysis

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

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

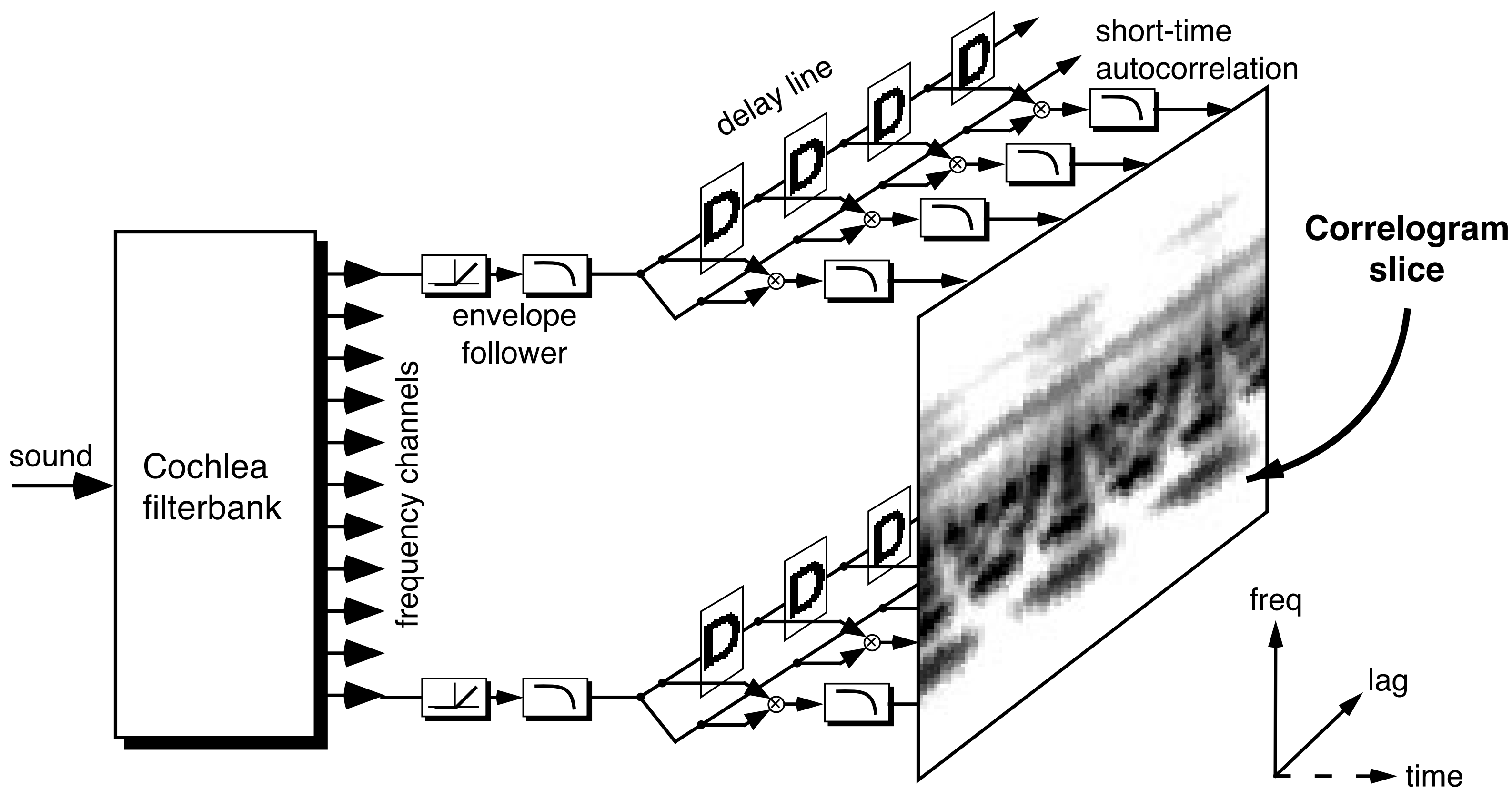


  - ... principal cue is harmonicity

CASA Preprocessing

Slaney & Lyon '90

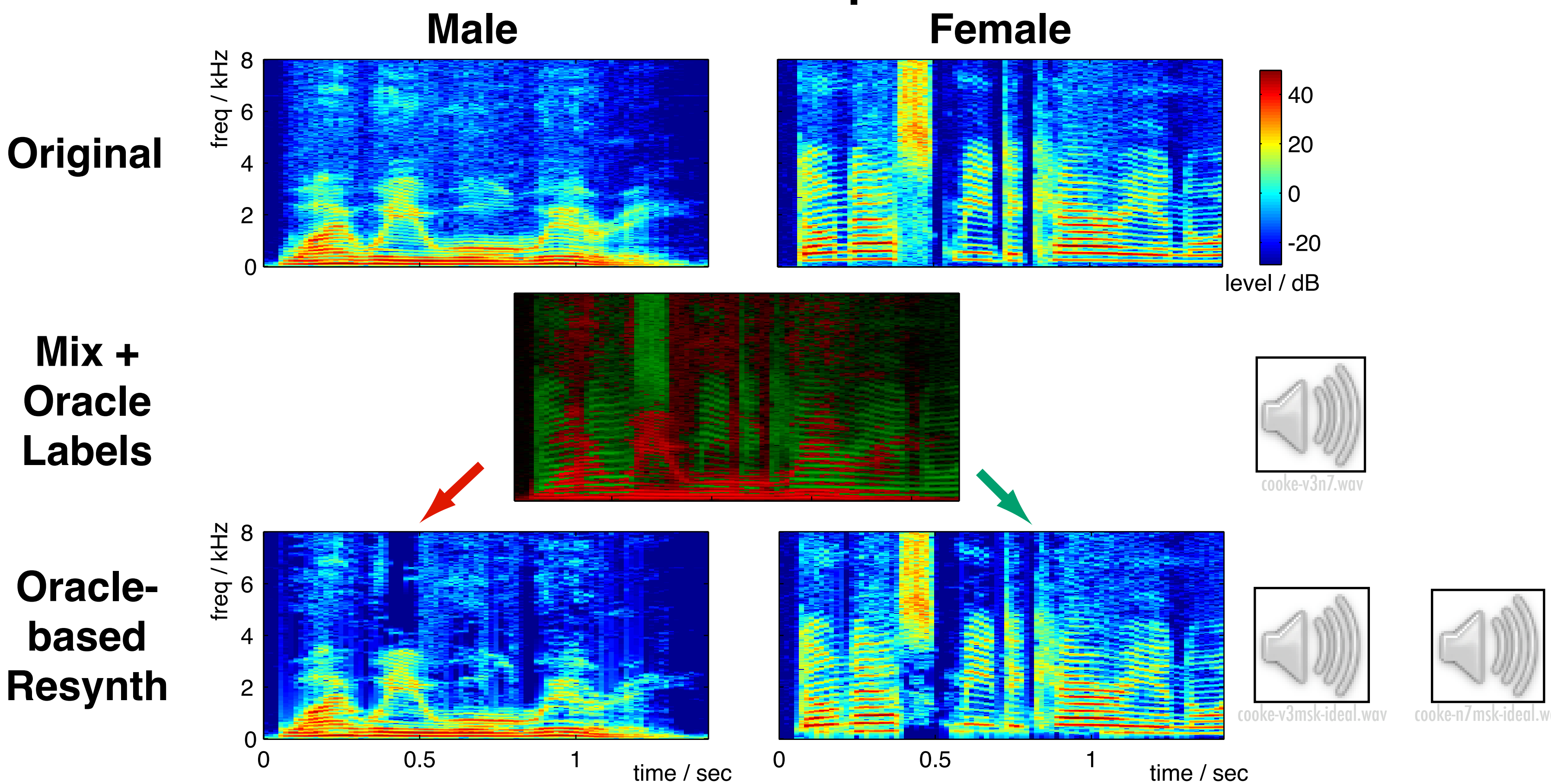
- **Correlogram**: a 3rd "periodicity" axis
  - envelope of wideband channels follows pitch



  - c/w Modulation Filtering *[Schimmel & Atlas '05]*

Time-Frequency (T-F) Masking

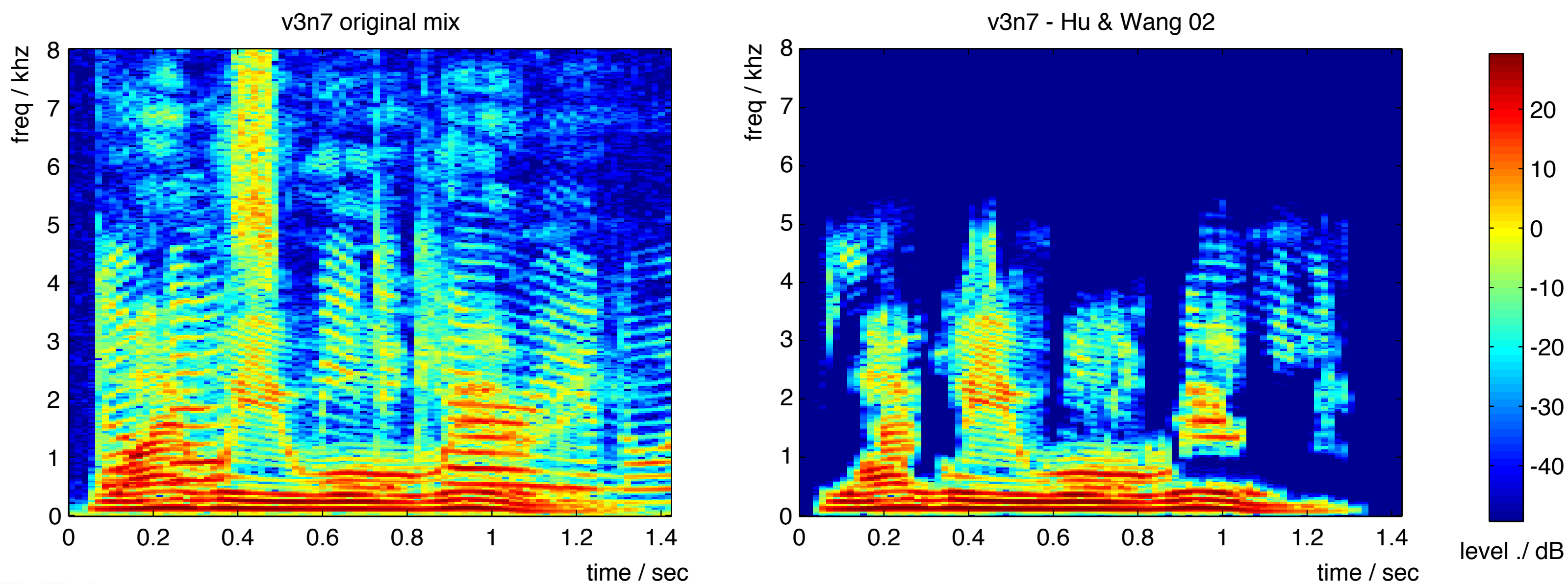
- "Local Dominance" assumption



  - oracle masks are remarkably effective!
  - $|mix - \max(male, female)| < 3$dB for ~80% of cells

CASA limitations

- **Driven by local features**
  - problems with aperiodic sources...
- **Limitations of T-F masking**
  - need to identify single-source regions
  - cannot undo overlaps – leaves gaps



from Hu & Wang '04

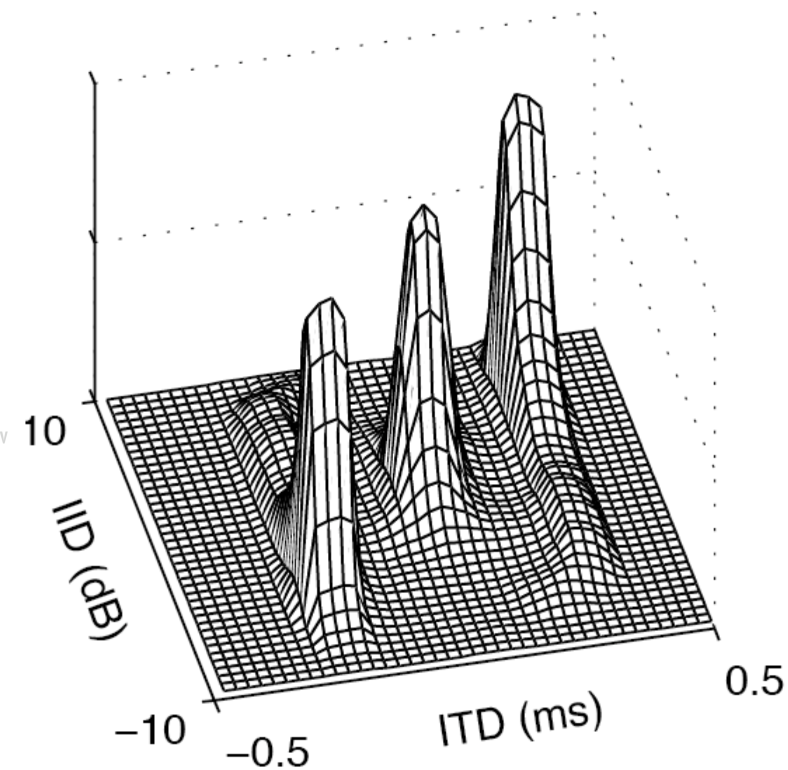
huwang-v3n7.wav

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

Columbia University in the City of New York

Combining Spatial + T-F Masking

- T-F masks based on inter-channel properties
  [Roman et al. '02], [Yilmaz & Rickard '04]
  lee_ss_in_1.wav lee_ss_duet_1.wav
  - multiple channels make CASA-like masks better



- T-F masking after ICA
  [Blin et al. '04]
  - cancellation can remove energy within T-F cells

Outline

1. The Speech Separation problem
2. Human Performance
3. Source Separation
4. **Source Inference**
   - Separation vs. inference
   - Model-based separation
   - Speech Fragment Decoding
5. Concluding Remarks

Separation vs. Inference

Ellis'96

- Ideal separation is rarely possible
  - i.e. no projection can completely remove overlaps
- Overlaps ⇒ Ambiguity
  - scene analysis = find "most reasonable" explanation
- Ambiguity can be expressed probabilistically
  - i.e. posteriors of sources $\{S_i\}$ given observations X:

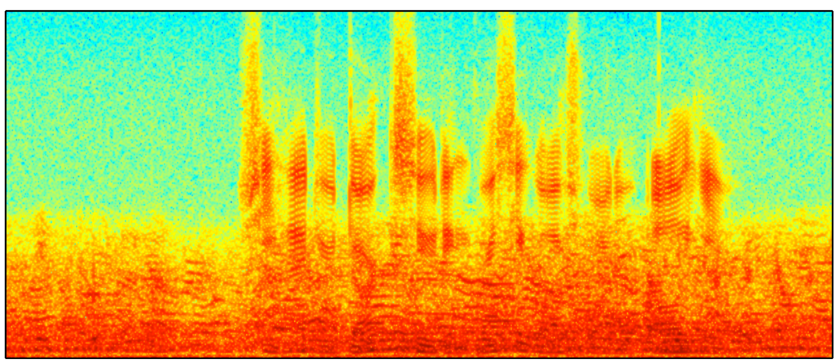
$$P(\{S_i\}|\, X) \propto \underbrace{P(X\,|\{S_i\})}_{\text{combination physics}}\ \underbrace{P(\{S_i\})}_{\text{source models}}$$

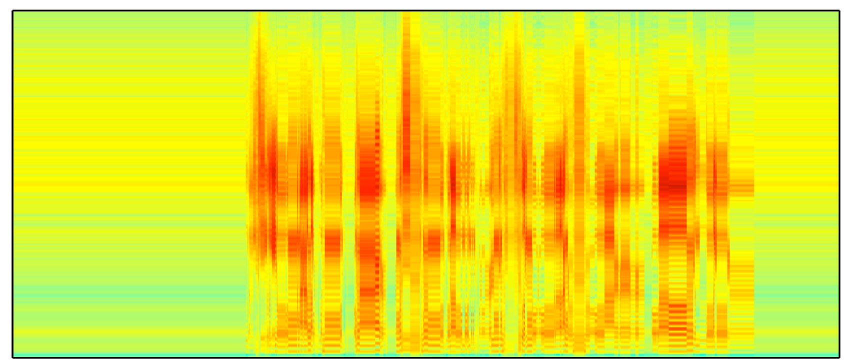
- Better source models → better inference
  - .. learn from examples?

Model-Based Separation

Varga & Moore'90
Roweis'03...

- Central idea:
  Employ strong learned constraints
  to disambiguate possible sources
  - $\{S_i\} = \text{argmax}_{S_i} P(X \mid \{S_i\})$
- e.g. fit speech-trained Vector-Quantizer
  to mixed spectrum:



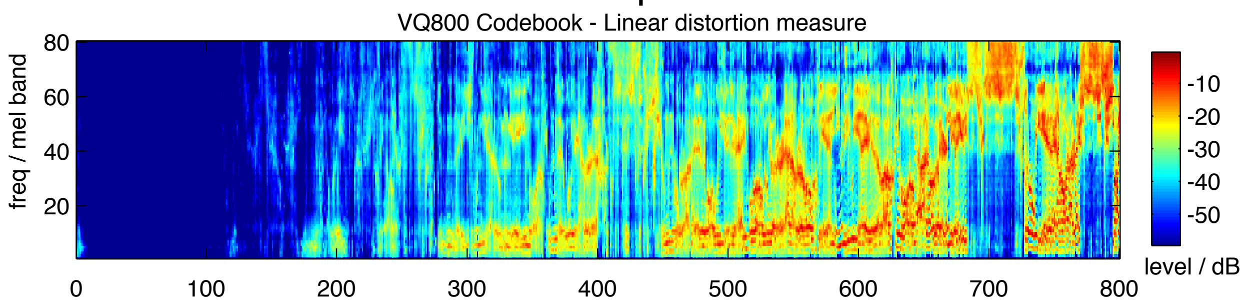


from Roweis'03

- separate via T-F mask (again)

Can Models Do CASA?

- **Source models can learn harmonicity, onset**
  - ... to subsume rules/representations of CASA

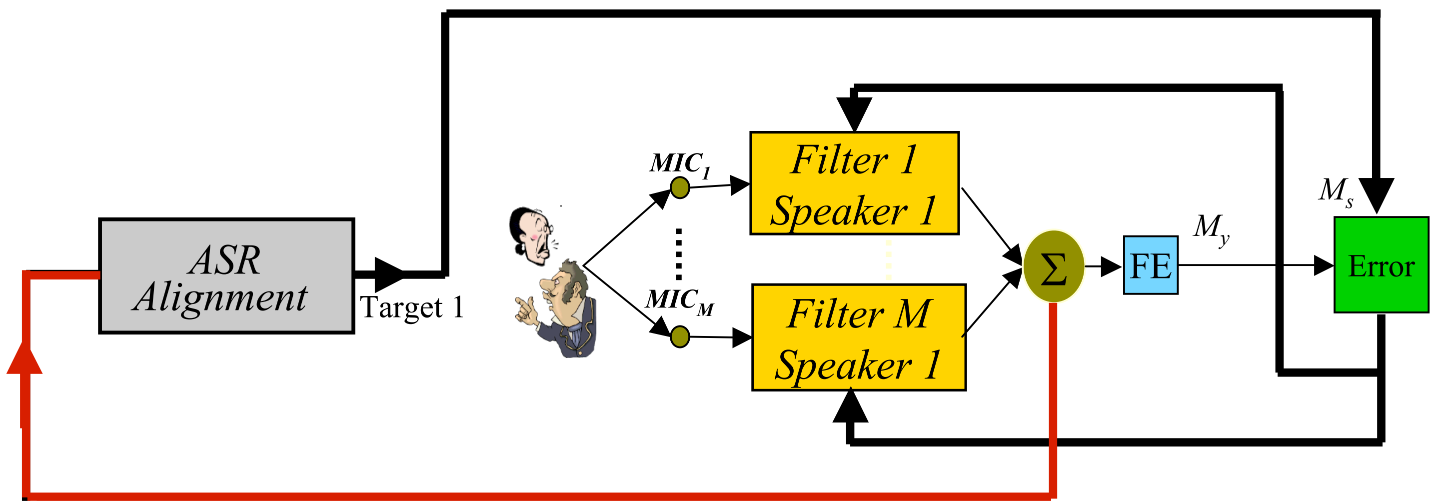


  - can capture spatial info too *[Pearlmutter & Zador'04]*
- **Can also capture sequential structure**
  - e.g. consonants follow vowels
  - ... like people do?
- **But: need source-specific models**
  ... for every possible source
  - use model adaptation? *[Ozerov et al. 2005]*

Separation with ASR Models

Seltzer et al. '02
Reyes et al. '03

- Drive separation engine
  to match outputs to existing speech models

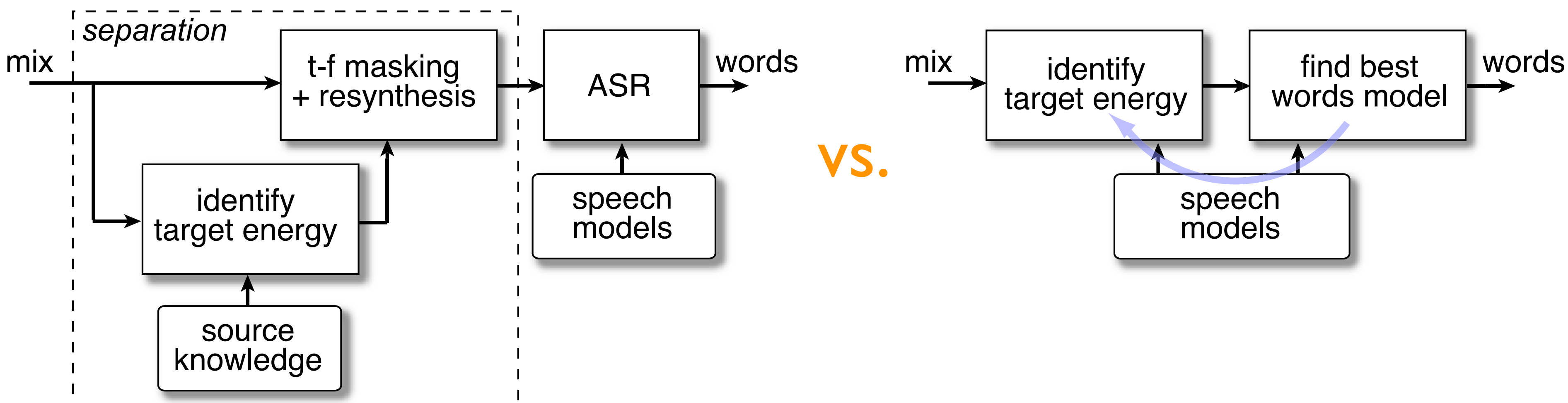


from Manuel Reyes's WASPAA 2003 presentation

  - ASR includes a very detailed source model

Separation or Description?

- **Are isolated waveforms required?**
  - clearly sufficient, but may not be necessary
  - not part of perceptual source separation!
- **Integrate separation with application?**
  - e.g. speech recognition

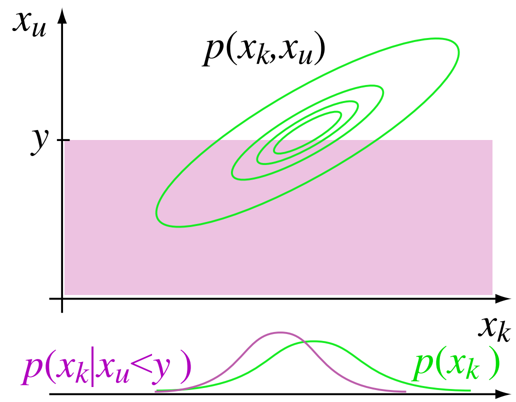


  - words output = abstract description of signal

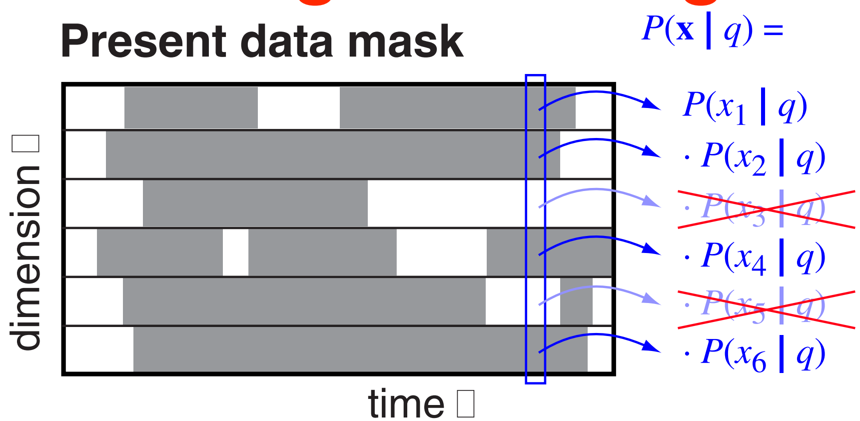
Missing Data Recognition

Cooke et al. '01

- Speech models $p(x|M)$ are multidimensional...
  - need values for all dimensions to evaluate $p(\bullet)$
- But: can make inferences given just a subset of dimensions x_k
  - $$p(x_k|M) = \int p(x_k, x_u|M)\,dx_u$$



- Hence, missing data recognition:

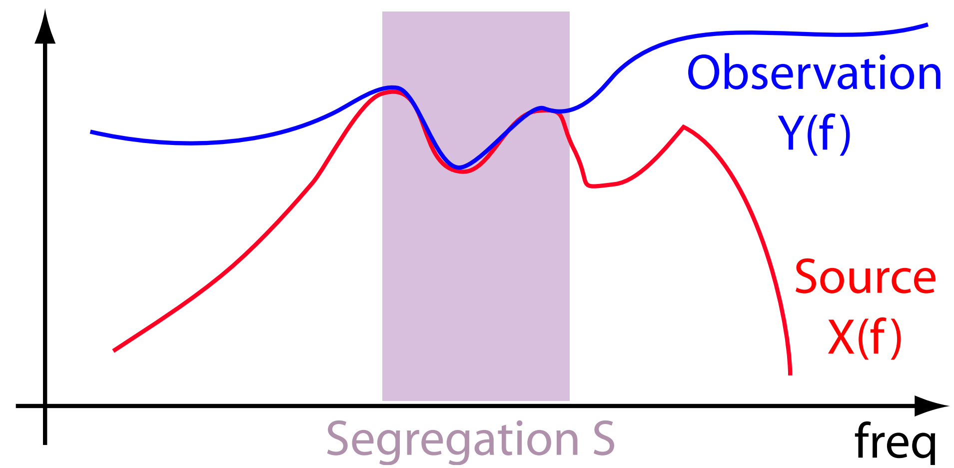


  - hard part is finding the mask (segregation)

The Speech Fragment Decoder

Barker et al. '05

- Match 'uncorrupt' spectrum to ASR models using missing data



- Joint search for model M and segregation S to maximize:

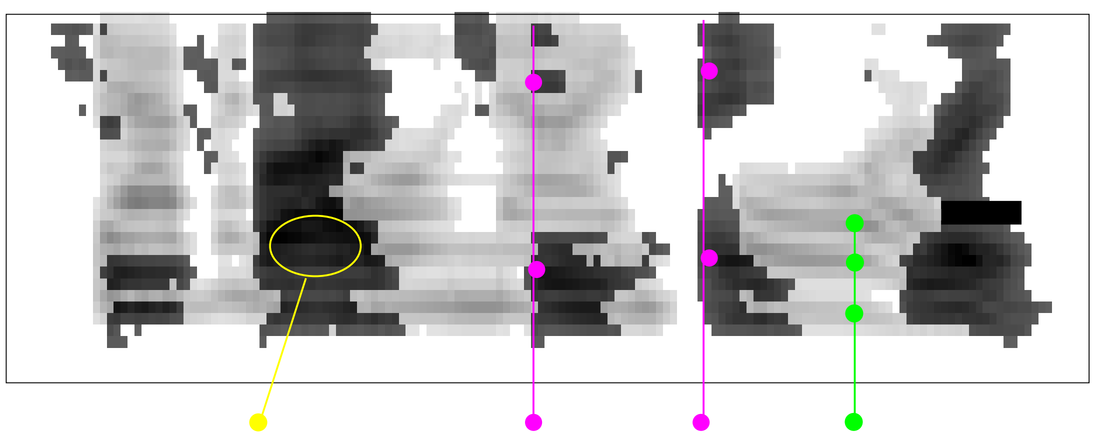
$$P(M,S|Y) = P(M)\int P(X|M)\cdot\frac{P(X|Y,S)}{P(X)}dX\cdot P(S|Y)$$

Isolated Source Model *Segregation Model*

Using CASA cues

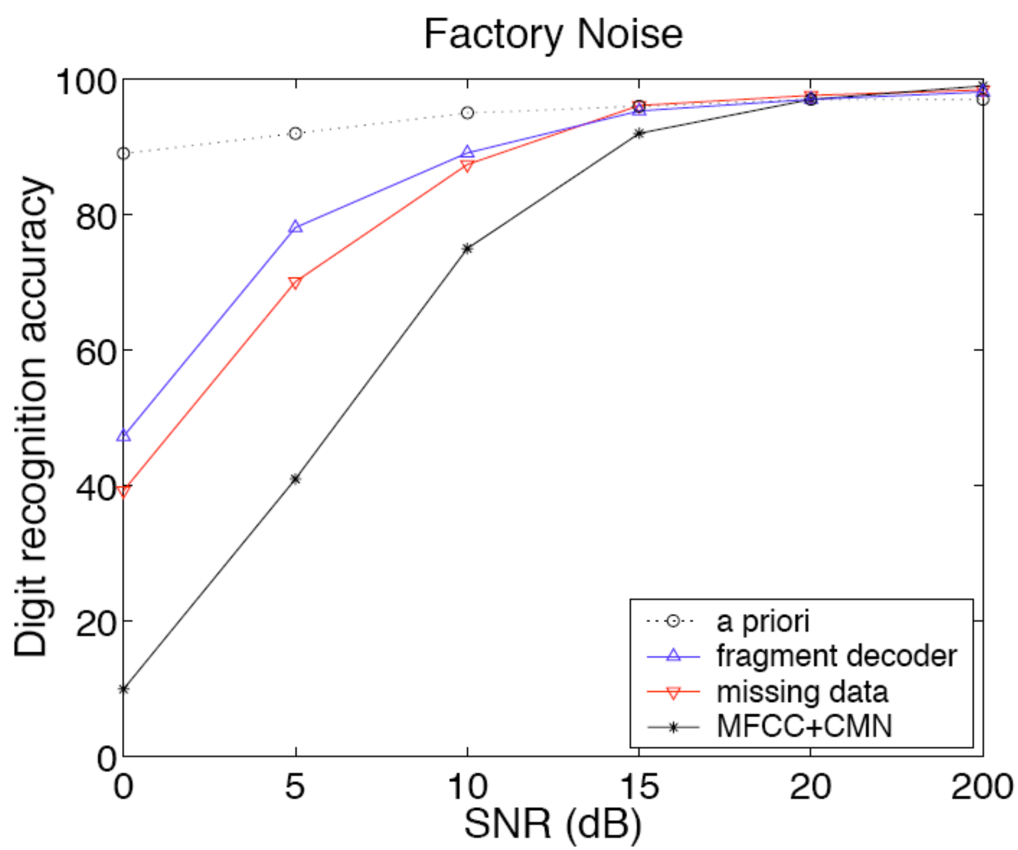
$$P(M,S|Y) = P(M)\int P(X|M)\cdot\frac{P(X|Y,S)}{P(X)}dX\cdot P(S|Y)$$

- CASA can help **search**
  - consider only segregations made from CASA chunks
- CASA can rate **segregation**
  - construct $P(S|Y)$ to reward CASA qualities:



Speech-Fragment Recognition

- CASA-based fragments give extra gain over missing-data recognition



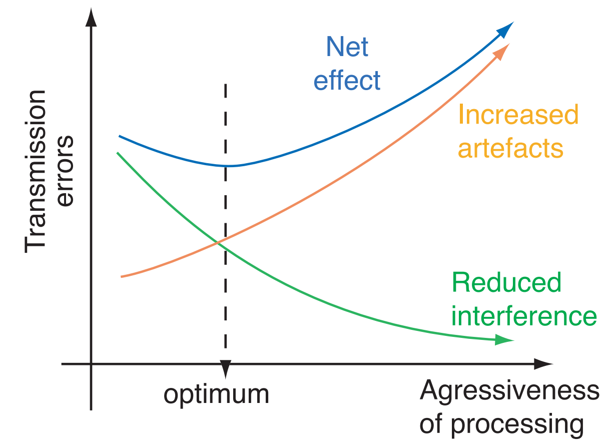
from Barker et al. '05

Outline

1. The Speech Separation problem
2. Human Performance
3. Source Separation
4. Source Inference
5. **Concluding Remarks**
    - Evaluation
    - Connecting to Perception

Evaluation

- How to measure separation performance?
  - depends what you are trying to do
- SNR?
  - energy (and distortions) are not created equal
  - different nonlinear components [Vincent et al. '06]
- Intelligibility?
  - rare for nonlinear processing to improve intelligibility
  - listening tests expensive



- ASR performance?
  - separate-then-recognize too simplistic; ASR needs to accommodate separation

“Speech Separation Challenge”

- Mixed and Noisy Speech ASR task defined by Martin Cooke and Te-Won Lee
    - short, grammatically-constrained utterances:

<command:4><color:4><preposition:4><letter:25><number:10><adverb:4>

e.g. "bin white at M 5 soon"

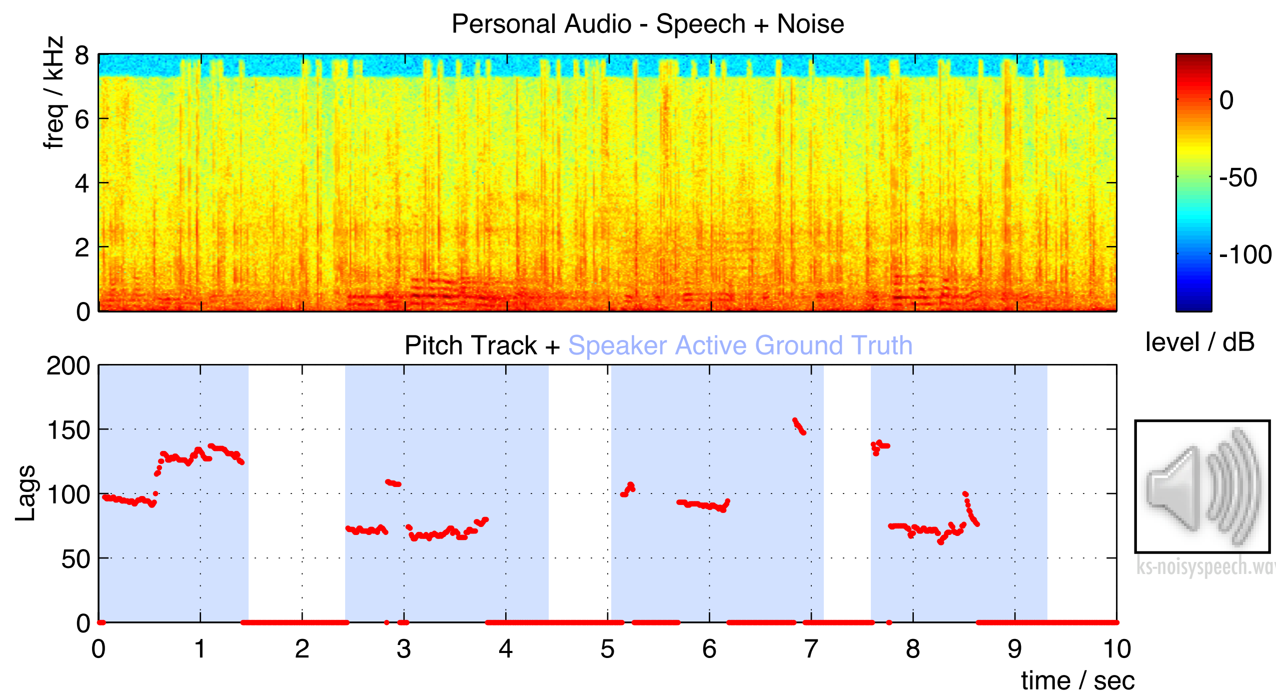
t5_bwam5s_m5_bbilzp_6p1.wav

- Results to be presented at Interspeech’06
    - http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm
- See also “Statistical And Perceptual Audition” workshop
    - http://www.sapa2006.org/

More Realistic Evaluation

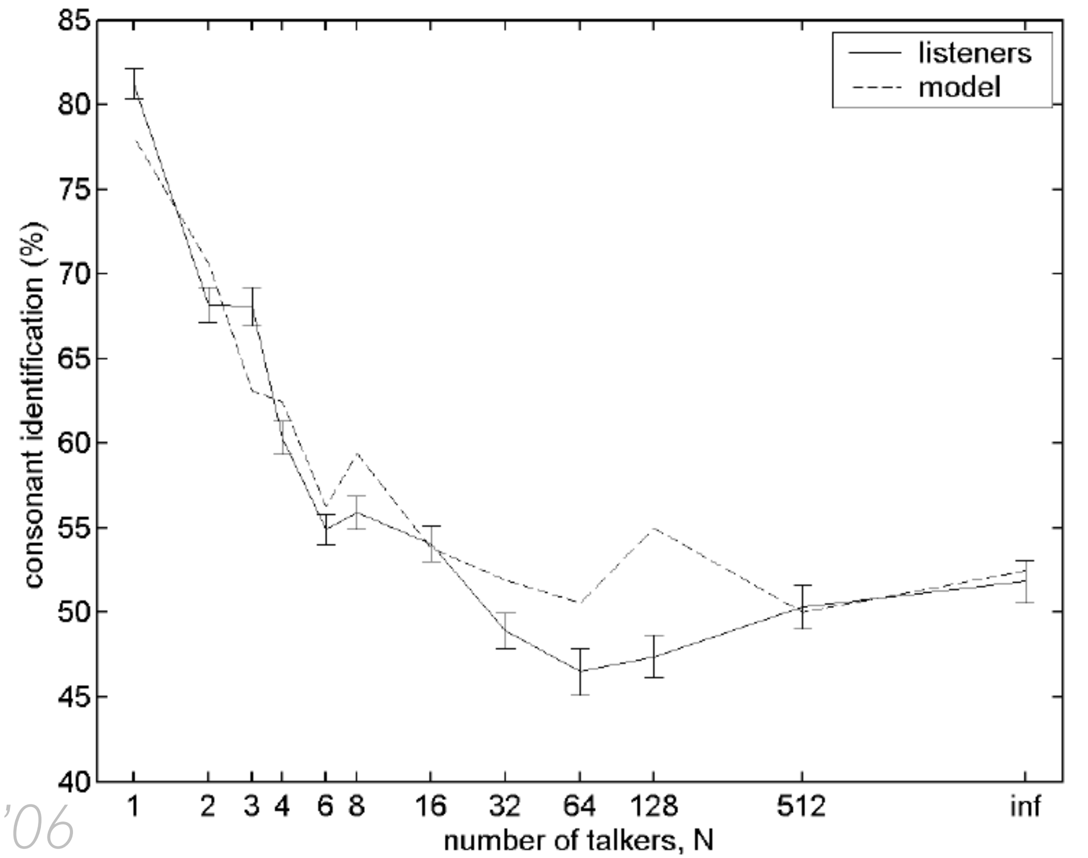
- Real-world **speech tasks**
    - crowded environments
    - applications: communication, command/control, transcription

- **Metric**
    - human intelligibility?
    - 'diarization' annotation (not transcription)



Reconnecting to Perception

- People are (still) much better at speech recognition, including mixtures
- Can we model human separation with ASR?
    - "Glimpse model": MD ASR using oracle local SNR
    - Listeners identify high SNR islands?



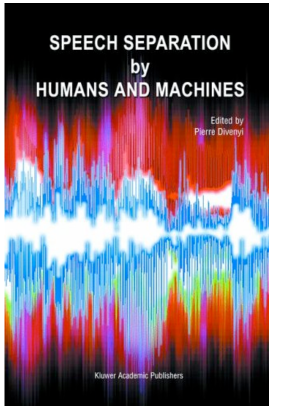
from Cooke'06

Summary & Conclusions

- Listeners do well separating speech
  - using spatial location
  - using source-property variations
- Machines do less well
  - difficult to apply enough constraints
  - need to exploit signal detail
- Models capture constraints
  - learn from the real world
  - adapt to sources
- Inferring state (≈ recognition) is a promising approach to separation

Sources / See Also

- NSF/AFOSR Montreal Workshops '03, '04
  - www.ebire.org/speechseparation/
  - labrosa.ee.columbia.edu/Montreal2004/
  - as well as the resulting book...



- Hanse meeting:
  - www.lifesci.sussex.ac.uk/home/Chris_Darwin/Hanse/
- DeLiang Wang's ICASSP'04 tutorial
  - www.cse.ohio-state.edu/~dwang/presentation.html
- Martin Cooke's NIPS'02 tutorial
  - www.dcs.shef.ac.uk/~martin/nips.ppt

References 1/2

[Barker et al. '05] J. Barker, M. Cooke, D. Ellis, "Decoding speech in the presence of other sources," Speech Comm. 45, 5-25, 2005.

[Bell & Sejnowski '95] A. Bell & T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," Neural Computation, 7:1129-1159, 1995.

[Blin et al.'04] A. Blin, S. Araki, S. Makino, "A sparseness mixing matrix estimation (SMME) solving the underdetermined BSS for convolutive mixtures," ICASSP, IV-85-88, 2004.

[Bregman '90] A. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[Brungart '01] D. Brungart, "Informational and energetic masking effects in the perception of two simultaneous talkers," JASA 109(3), March 2001.

[Brungart et al. '01] 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.

[Brungart et al. '02] 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.

[Brown & Cooke '94] G. Brown & M. Cooke, "Computational auditory scene analysis," Comp. Speech & Lang. 8 (4), 297–336, 1994.

[Cooke et al. '01] M. Cooke, P. Green, L. Josifovski, A. Vizinho, "Robust automatic speech recognition with missing and uncertain acoustic data," Speech Communication 34, 267-285, 2001.

[Cooke'06] M. Cooke, "A glimpsing model of speech perception in noise," submitted to JASA.

[Darwin & Carlyon '95] C. Darwin & R. Carlyon, "Auditory grouping" *Handbk of Percep. & Cogn. 6: Hearing*, 387–424, Academic Press, 1995.

[Ellis'96] D. Ellis, "Prediction-Driven Computational Auditory Scene Analysis," Ph.D. thesis, MIT EECS, 1996.

[Hu & Wang '04] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Tr. Neural Networks, 15(5), Sep. 2004.

[Okuno et al. '99] H. Okuno, T. Nakatani, T. Kawabata, "Listening to two simultaneous speeches," Speech Communication 27, 299–310, 1999.

References 2/2

[Ozerov et al. '05] A. Ozerov, P. Phillippe, R. Gribonval, F. Bimbot, "One microphone singing voice separation using source-adapted models," Worksh. on Apps. of Sig. Proc. to Audio & Acous., 2005.

[Pearlmutter & Zador '04] B. Pearlmutter & A. Zador, "Monaural Source Separation using Spectral Cues," Proc. ICA, 2005.

[Parra & Spence '00] L. Parra & C. Spence, "Convolutive blind source separation of non-stationary sources," IEEE Tr. Speech & Audio, 320-327, 2000.

[Reyes et al. '03] M. Reyes-Gómez, B. Raj, D. Ellis, "Multi-channel source separation by beamforming trained with factorial HMMs," Worksh. on Apps. of Sig. Proc. to Audio & Acous., 13–16, 2003.

[Roman et al. '02] N. Roman, D.-L. Wang, G. Brown, "Location-based sound segregation," ICASSP, I-1013-1016, 2002.

[Roweis '03] S. Roweis, "Factorial models and refiltering for speech separation and denoising," EuroSpeech, 2003.

[Schimmel & Atlas '05] S. Schimmel & L. Atlas, "Coherent Envelope Detection for Modulation Filtering of Speech," ICASSP, I-221-224, 2005.

[Slaney & Lyon '90] M. Slaney & R. Lyon, "A Perceptual Pitch Detector," ICASSP, 357-360, 1990.

[Smaragdis '98] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," Intl. Wkshp. on Indep. & Artif.l Neural Networks, Tenerife, Feb. 1998.

[Seltzer et al. '02] M. Seltzer, B. Raj, R. Stern, "Speech recognizer-based microphone array processing for robust hands-free speech recognition," ICASSP, I–897–900, 2002.

[Varga & Moore '90] A. Varga & R. Moore, "Hidden Markov Model decomposition of speech and noise," ICASSP, 845–848, 1990.

[Vincent et al. '06] E. Vincent, R. Gribonval, C. Févotte, "Performance measurement in Blind Audio Source Separation." IEEE Trans. Speech & Audio, in press.

[Yilmaz & Rickard '04] O. Yilmaz & S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Tr. Sig. Proc. 52(7), 1830-1847, 2004.