Auditory Scene Analysis in Humans and Machines

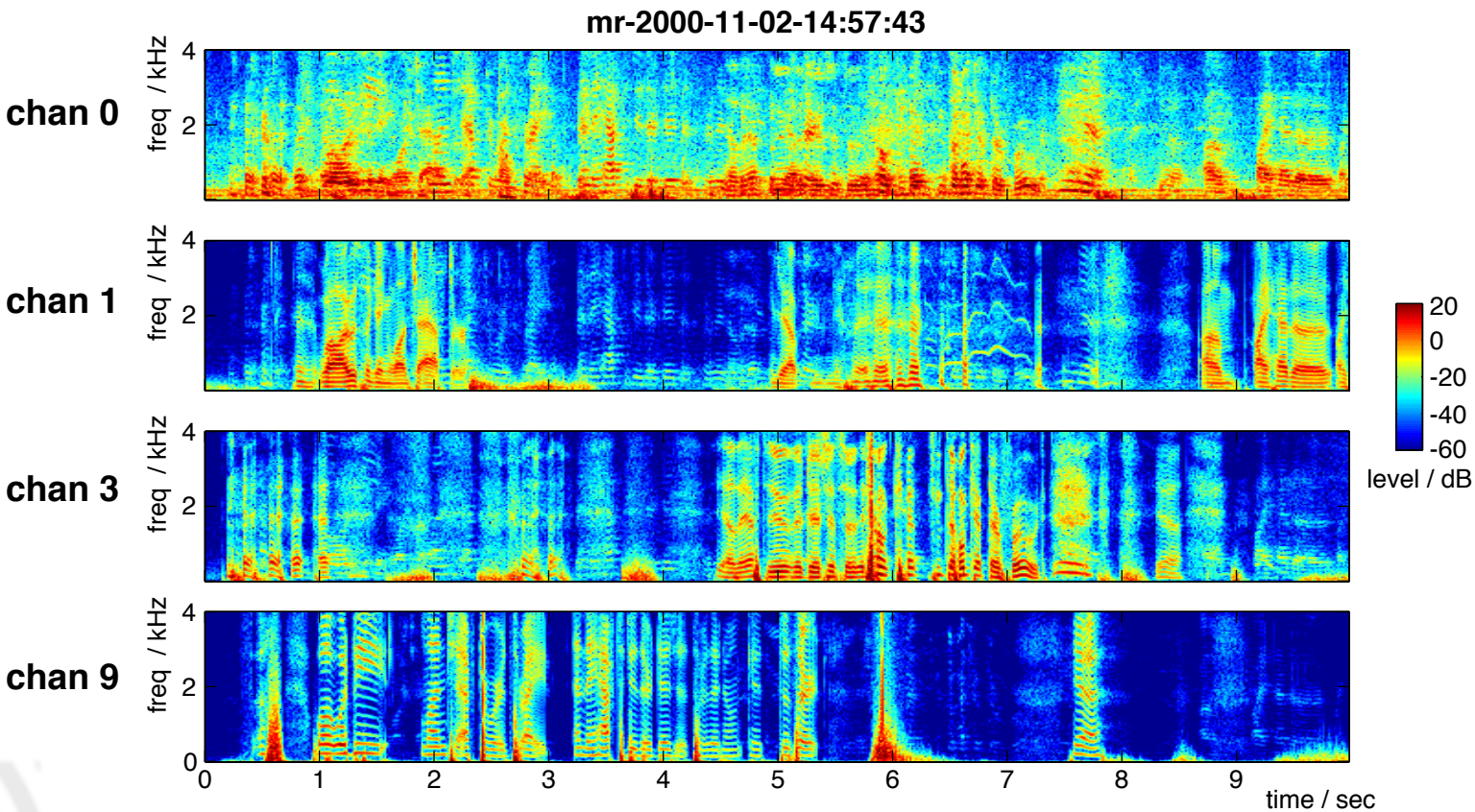
Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Eng., Columbia Univ., NY USA

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

1. The ASA Problem
2. Human ASA
3. Machine Source Separation
4. Systems & Examples
5. Concluding Remarks

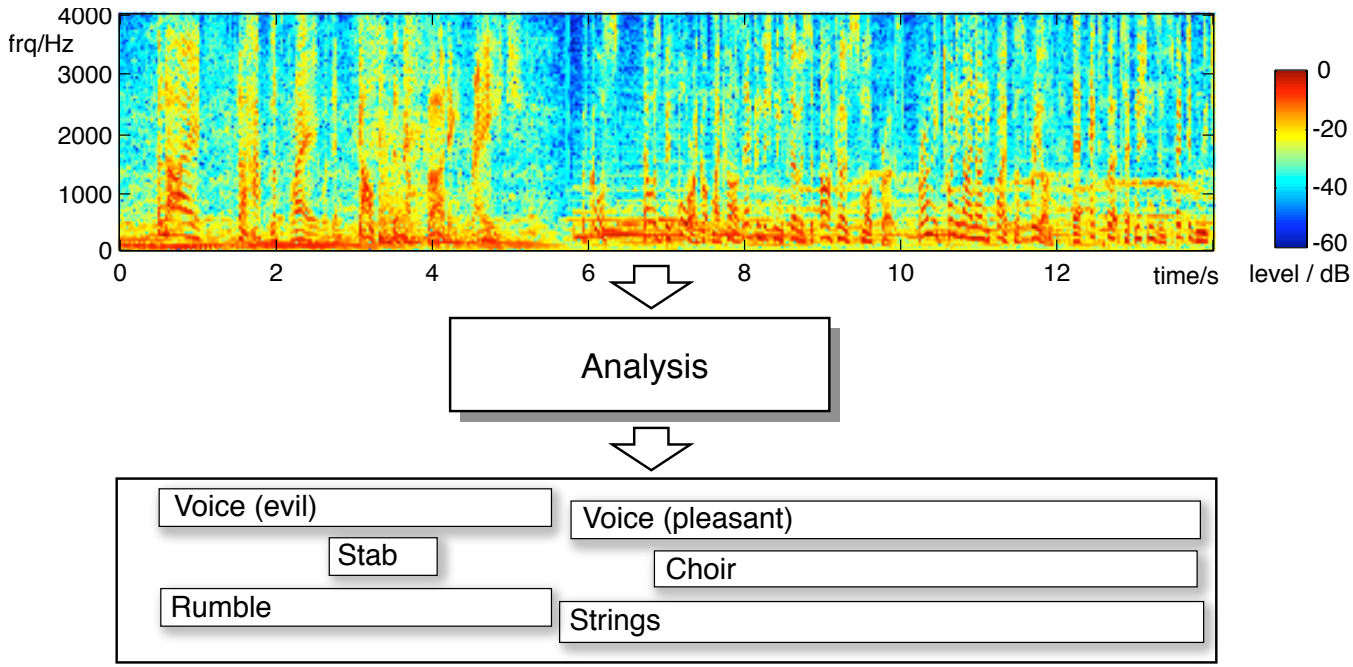
1. Auditory Scene Analysis

- Sounds rarely occurs in isolation
  - .. but recognizing sources in mixtures is a problem
  - .. for humans and machines



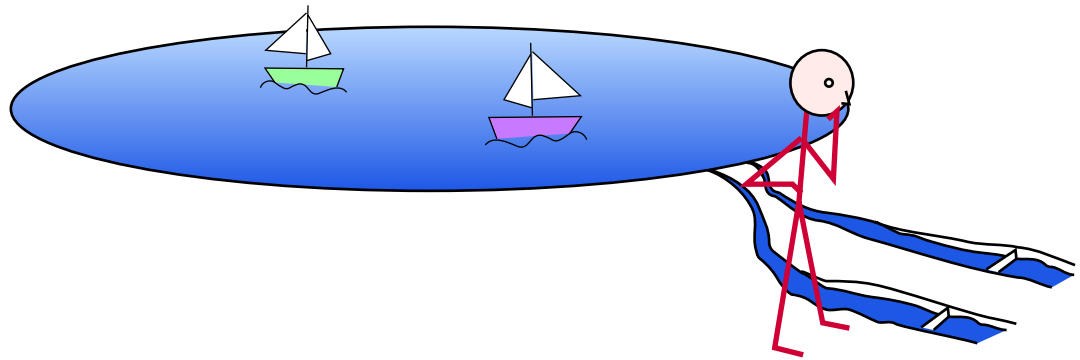
Sound Mixture Organization

- Goal: recover individual sources from scenes
  - .. duplicating the perceptual effect



- Problems: competing sources, channel effects
- Dimensionality loss
  - need additional constraints

The Problem of Mixtures

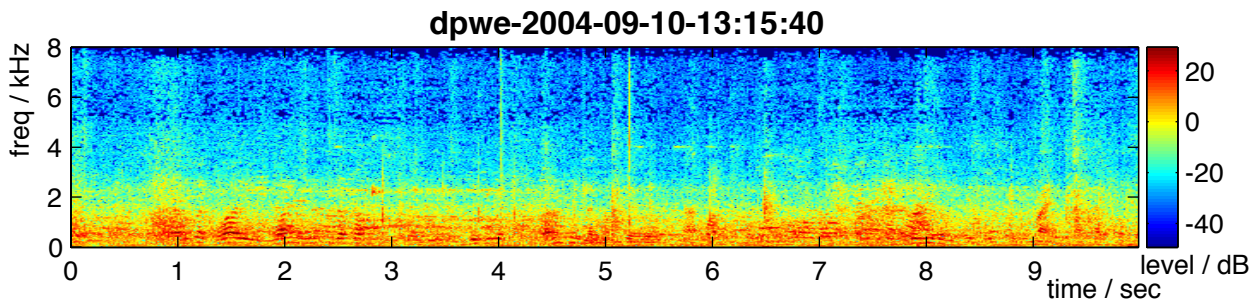


"Imagine two narrow channels dug up from the edge of a lake, with handkerchiefs stretched across each one. Looking only at the motion of the handkerchiefs, you are to answer questions such as: How many boats are there on the lake and where are they?" (after Bregman'90)

- **Received waveform is a mixture**
  - 2 sensors, N sources - underconstrained
- **Undoing mixtures: hearing's primary goal?**
  - .. by any means available

Source Separation Scenarios

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



pa-2004-09-10-131540.wav

- Unmixing/remixing/enhancement...

How Can We Separate?

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

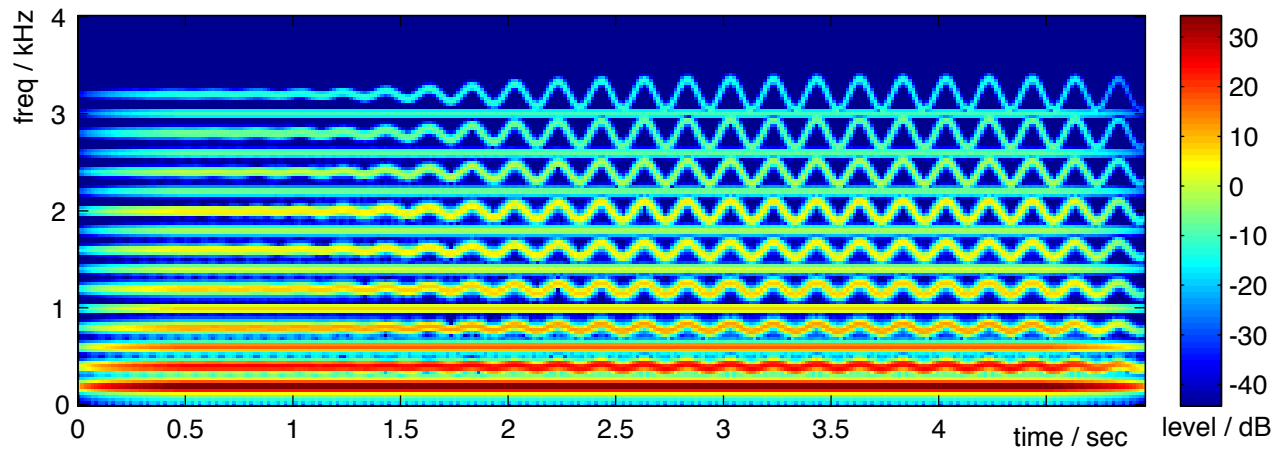
    → use part to guess the remainder

Outline

1. The ASA Problem
2. Human ASA
   - scene analysis
   - separation by location
   - separation by source characteristics
3. Machine Source Separation
4. Systems & Examples
5. Concluding Remarks

Auditory Scene Analysis

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

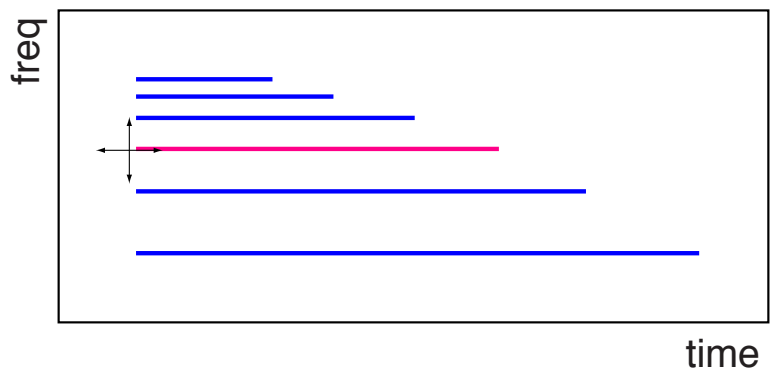


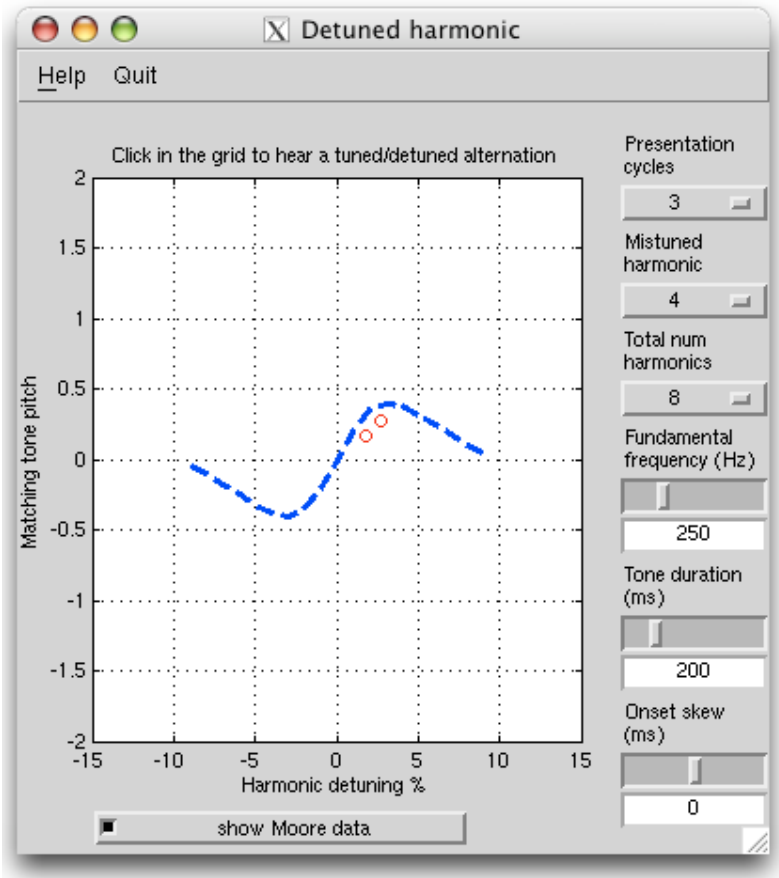
Reynolds-McAdams oboe

reynolds-mcadams-dpwe.wav

Perceiving Sources

- Harmonics distinct in ear, but perceived as one source ("fused"):



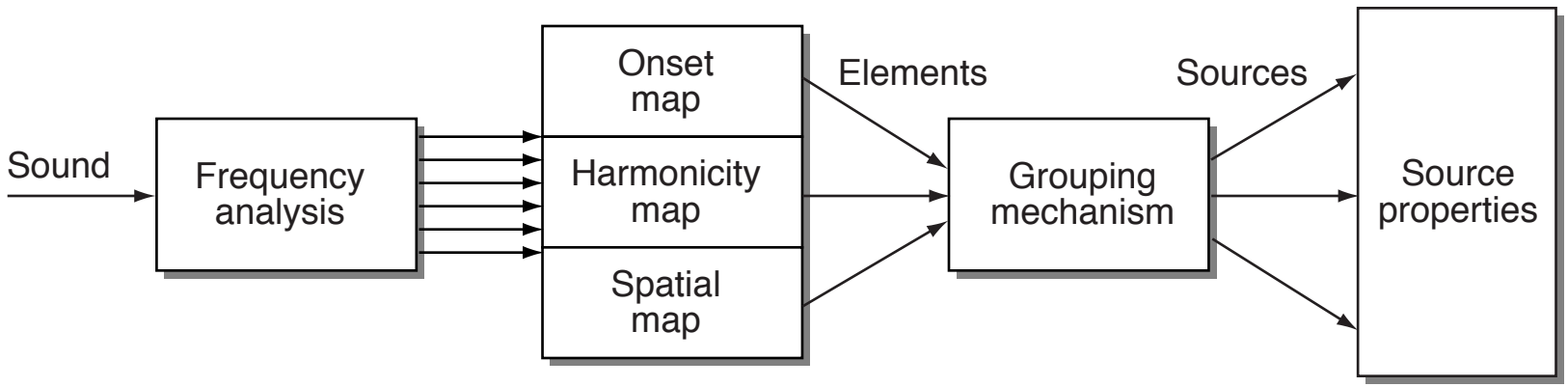


  - depends on common onset
  - depends on harmonics
- Experimental techniques
  - ask subjects "how many"
  - match attributes e.g. pitch, vowel identity
  - brain recordings (EEG "mismatch negativity")

Auditory Scene Analysis

Bregman'90
Darwin & Carlyon'95

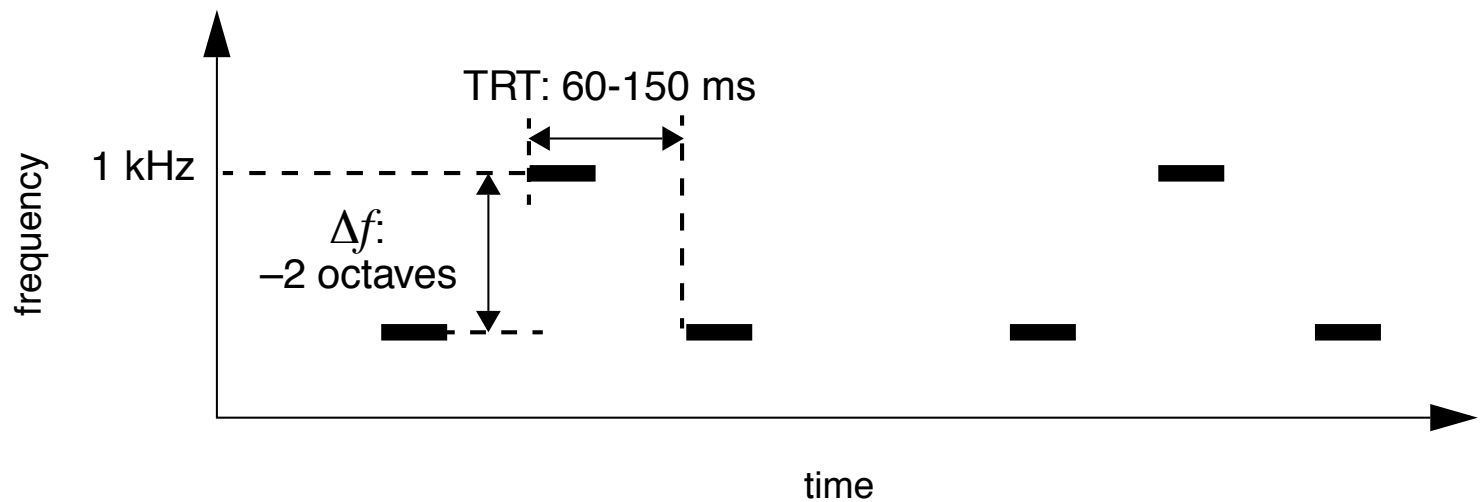
- How do people analyze sound mixtures?
  - break mixture into small elements (in time-freq)
  - elements are grouped in to sources using cues
  - sources have aggregate attributes
- Grouping rules (Darwin, Carlyon, ...):
  - cues: common onset/offset/modulation, harmonicity, spatial location, ...



(after Darwin 1996)

Streaming

- Sound event sequences are organized into streams
  - i.e. distinct perceived sources
  - difficult to make comparisons between streams
- Two-tone streaming experiments:

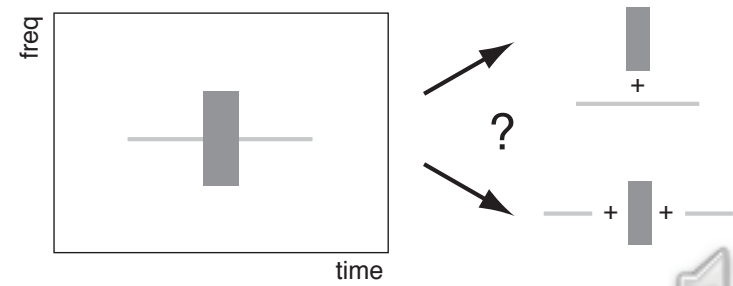


asacd-36-streaming-50s

  - ecological relevance?

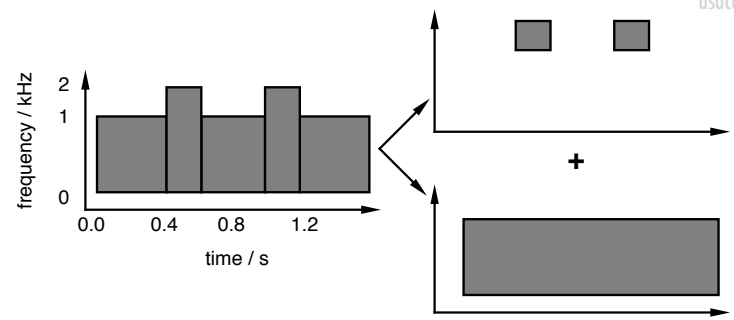
Illusions & Restoration

- Illusion = hearing more than is "there"
  - e.g. "pulsation threshold" example - tone is masked



asacd-26-pulsation

  - "old-plus-new" heuristic: existing sources continue



bregnoise

- Need to infer most likely real-world events
  - observation equally good match to either case
  - prior likelihood of continuity much higher

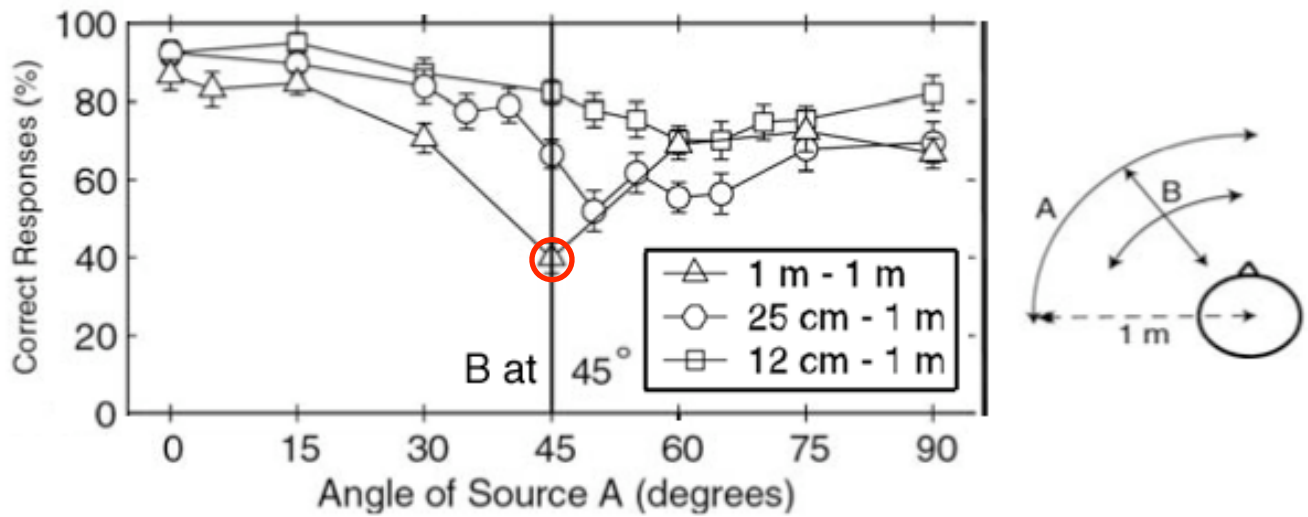
Human Performance: Spatial Separation

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:

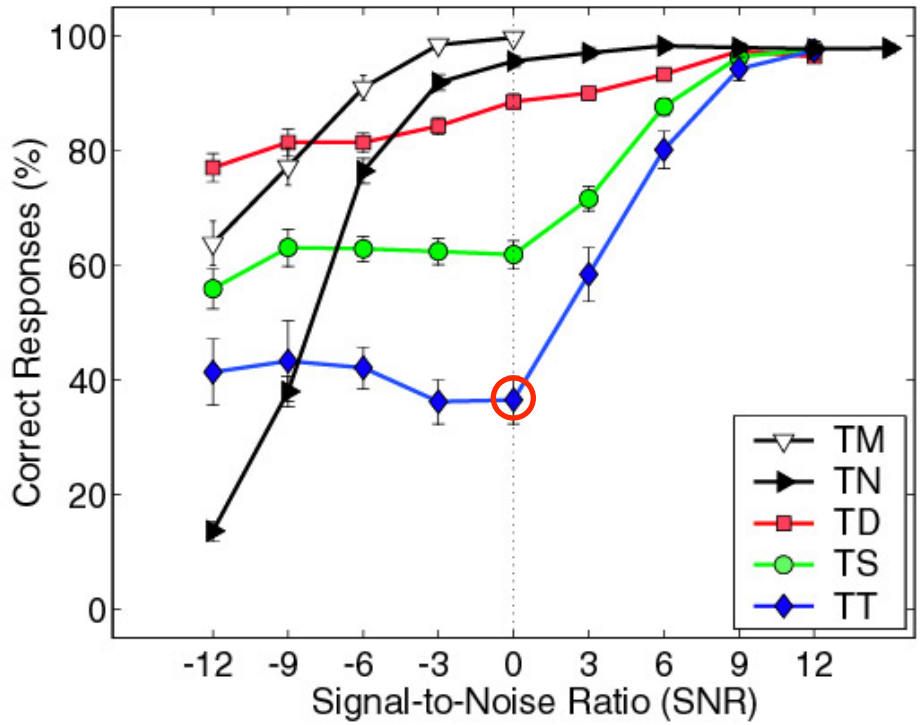


  - A, B same speaker
  - Range effect

Separation by Vocal Differences

Brungart et al.'01

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

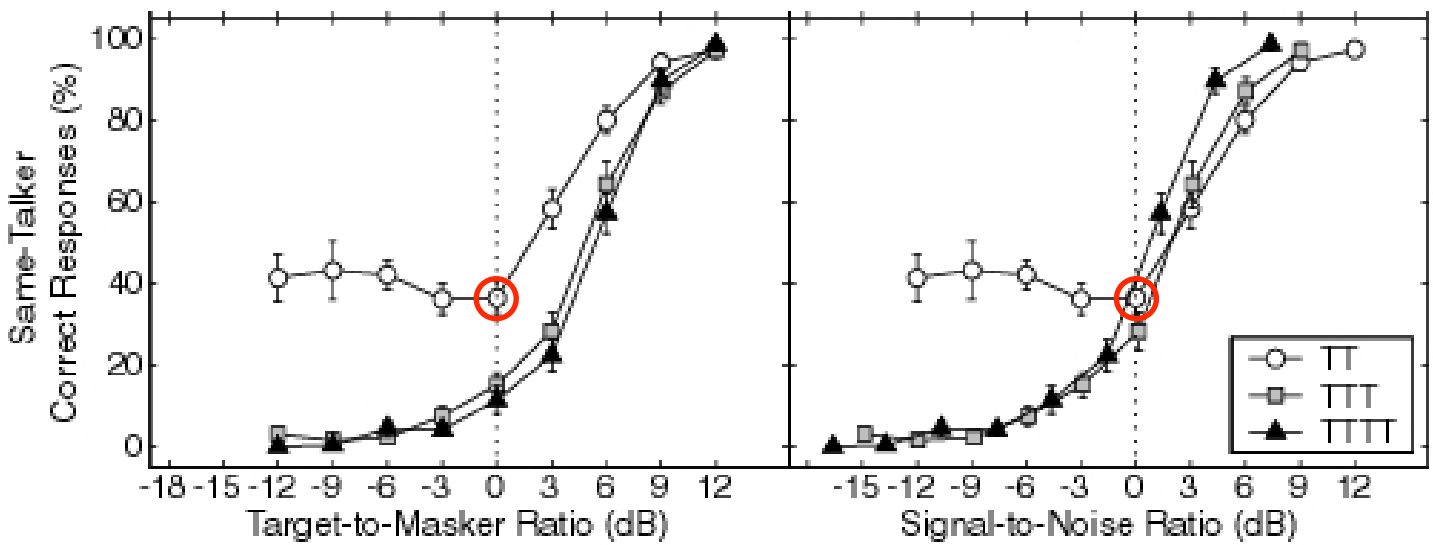


  - energetic vs. informational masking

Varying the Number of Voices

Brungart et al.'01

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



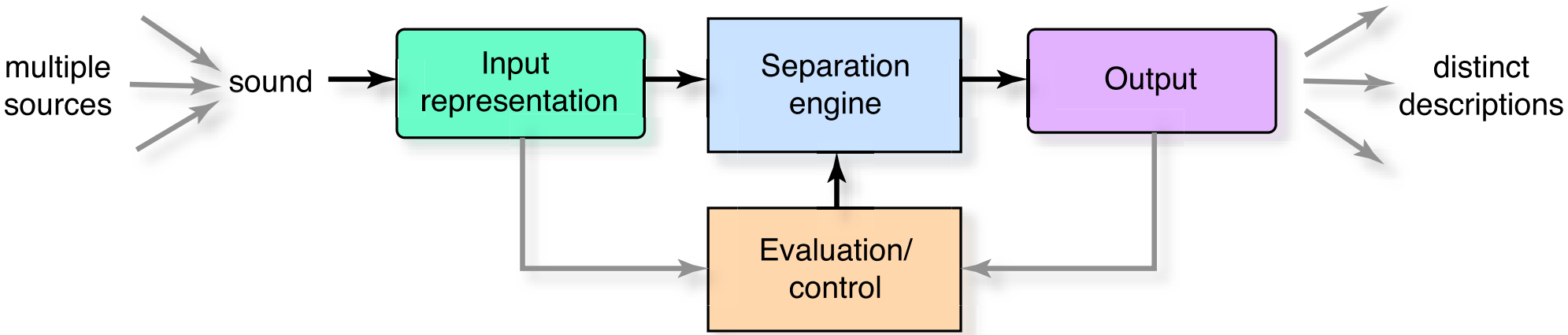
  - mix of *N* voices tends to speech-shaped noise...

Outline

1. The ASA Problem
2. Human ASA
3. **Machine Source Separation**
   - Independent Component Analysis
   - Computational Auditory Scene Analysis
   - Model-Based Separation
4. Systems & Examples
5. Concluding Remarks

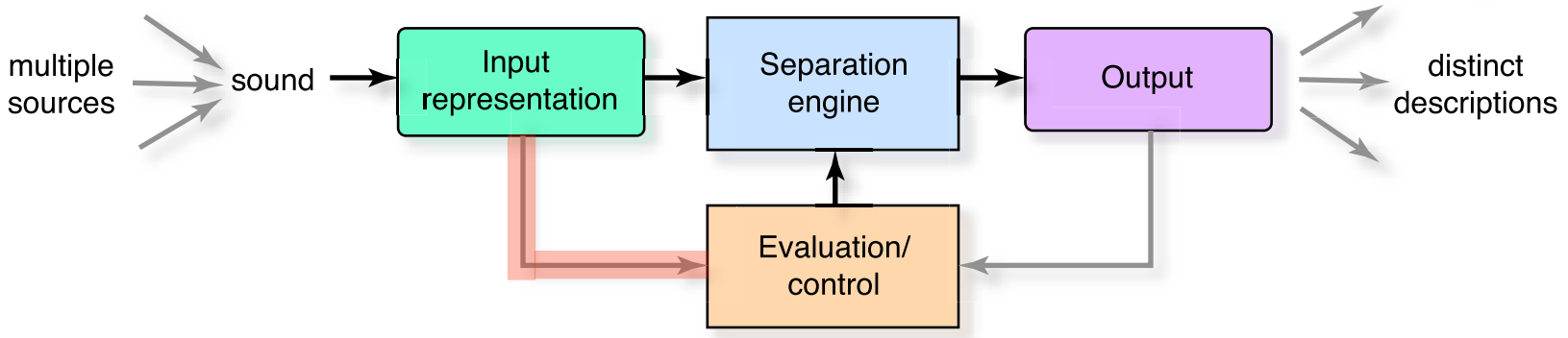
Scene Analysis Systems

- "Scene Analysis"
  - not necessarily separation, recognition, ...
  - scene = overlapping objects, ambiguity
- General Framework:



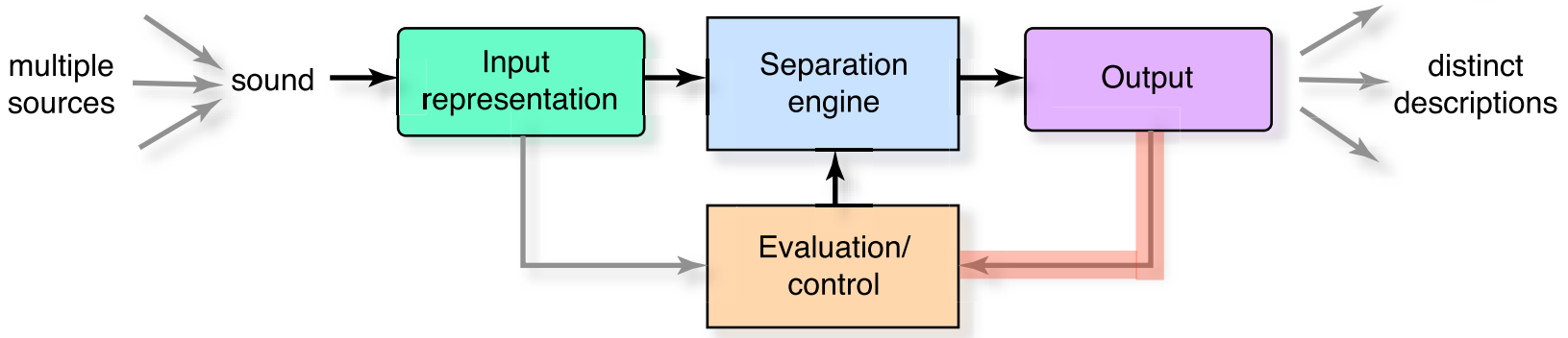
  - distinguish input and output representations
  - distinguish engine (algorithm) and control (constraints, "computational model")

Human and Machine Scene Analysis



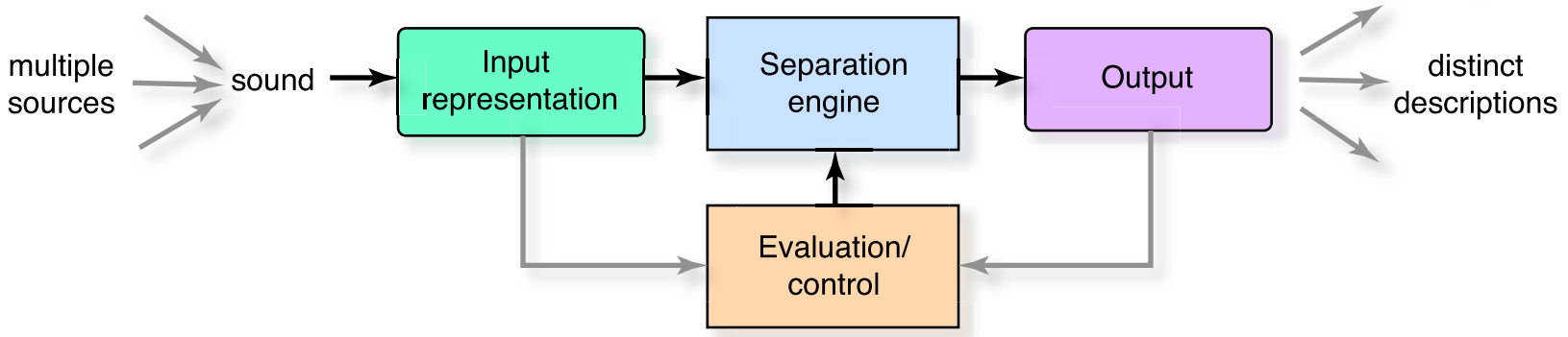
- CASA (e.g. Brown'92):
  - Input: Periodicity, continuity, onset "maps"
  - Output: Waveform (or mask)
  - Engine: Time-frequency masking
  - Control: "Grouping cues" from input
    - or: spatial features (Roman, ...)

Human and Machine Scene Analysis



- CASA (e.g. Brown'92):
- **ICA (Bell & Sejnowski et seq.):**
  - Input: waveform (or STFT)
  - Output: waveform (or STFT)
  - Engine: cancellation
  - Control: statistical independence of outputs
    - or energy minimization for beamforming

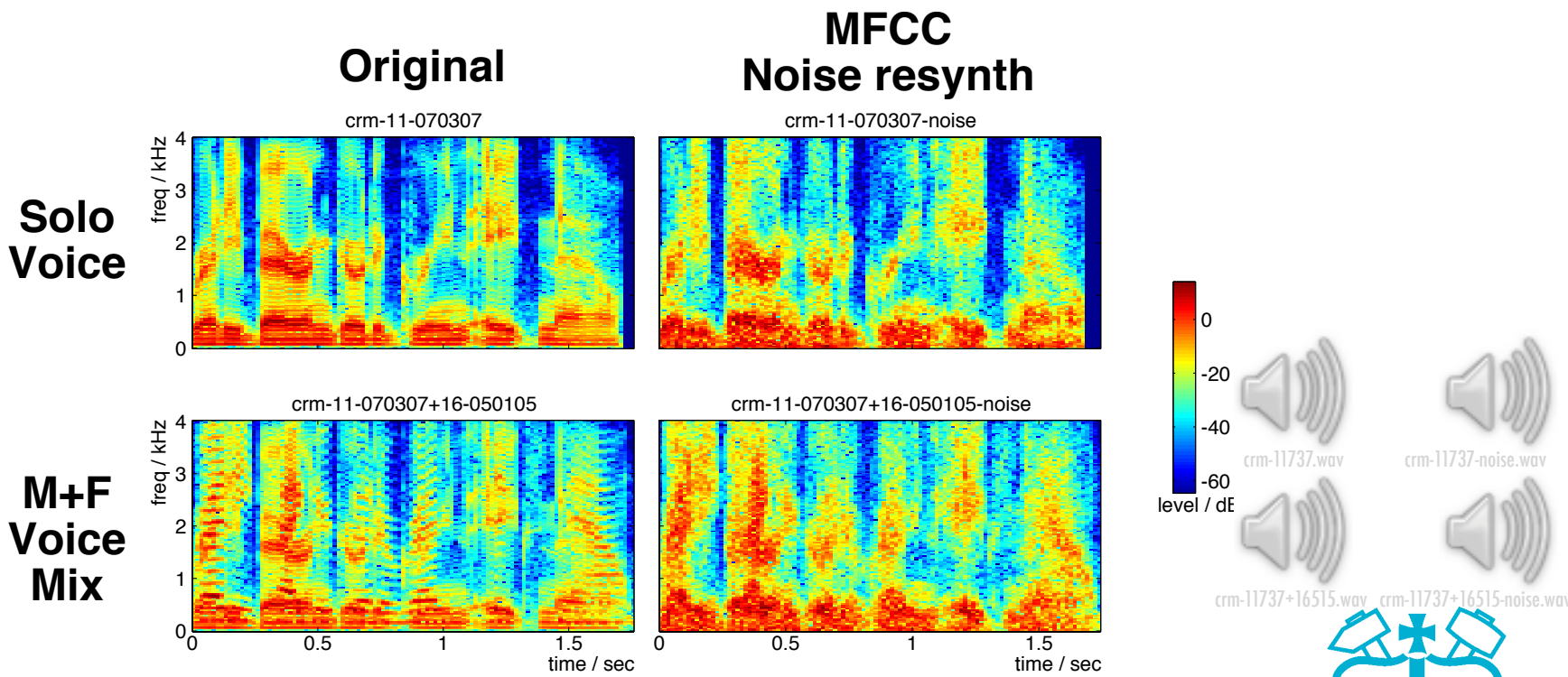
Human and Machine Scene Analysis



- CASA (e.g. Brown'92):
- ICA (Bell & Sejnowski et seq.):
- **Human Listeners:**
  - Input: excitation patterns ...
  - Output: percepts ...
  - Engine: ?
  - Control: find a plausible explanation

Machine Separation

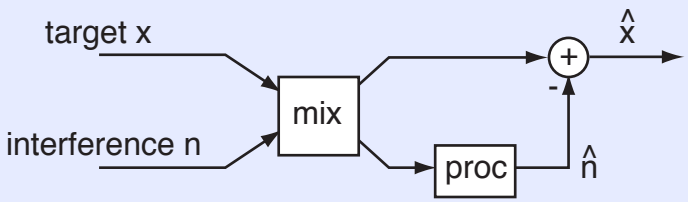
- Problem: Features of combinations are not combinations of features
  - voice is easy to characterize when in isolation
  - redundancy needed for real-world communication



Separation Approaches

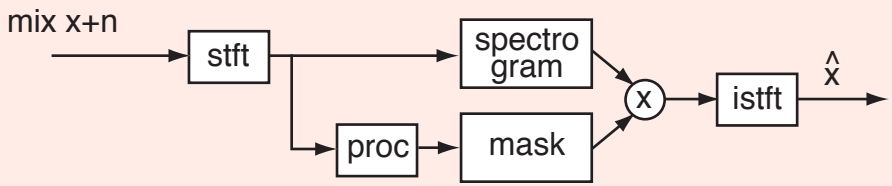
ICA

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



CASA / Model-based

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

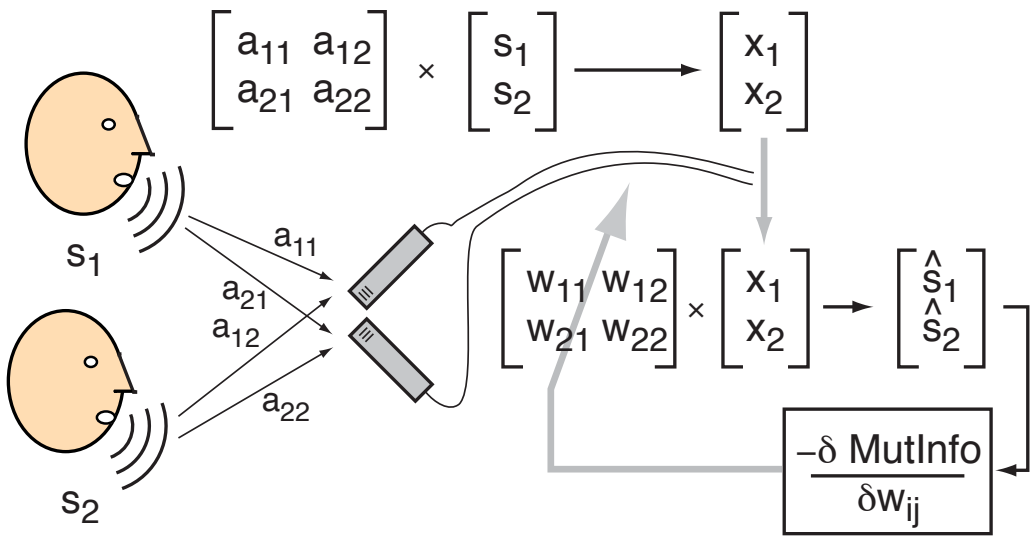


- Very different approaches!

Independent Component Analysis

Bell & Sejnowski'95
Smaragdis'98
...

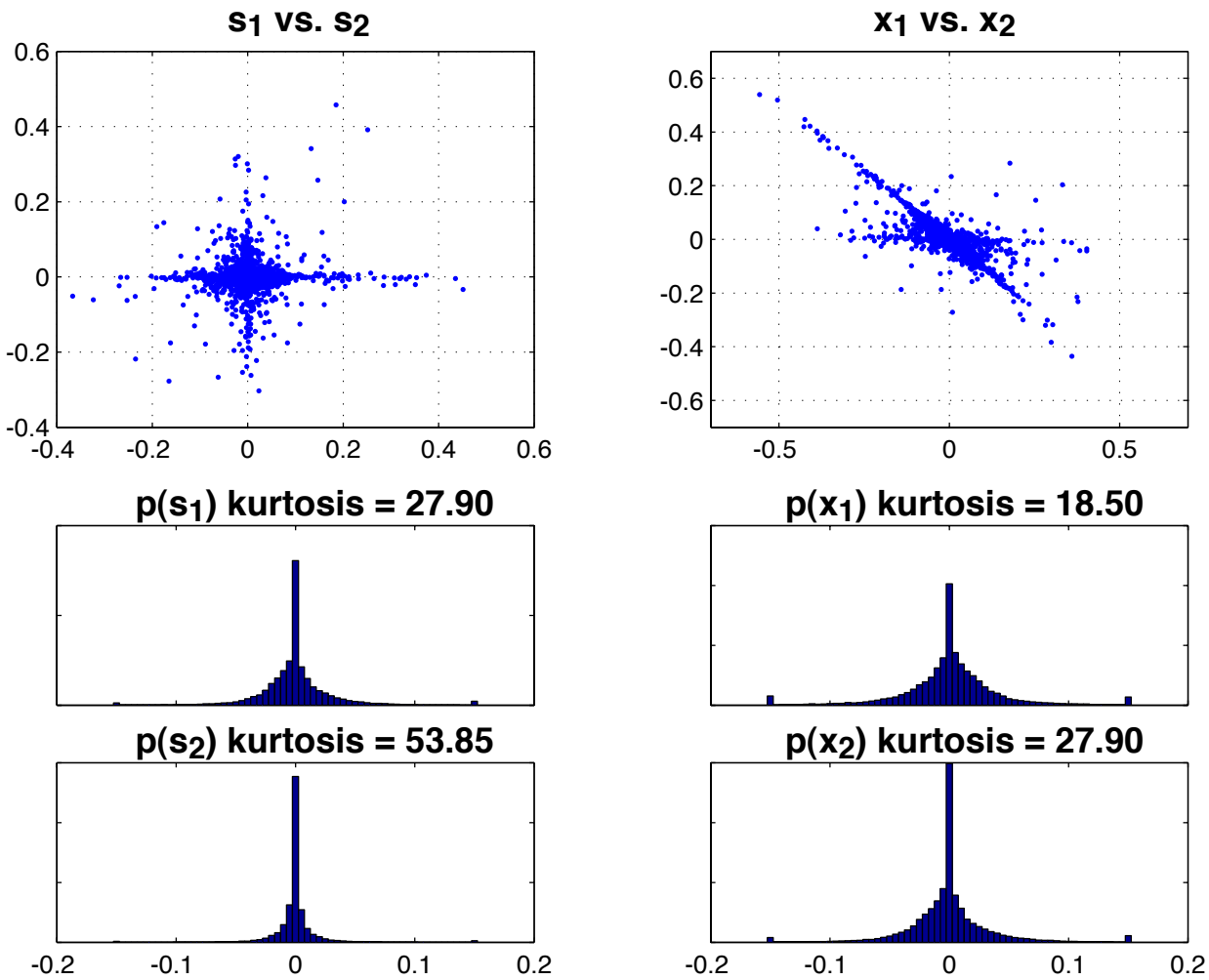
- Central idea:
  Search unmixing space
  to maximize independence of outputs



- simple mixing
  → a good solution (usually) exists

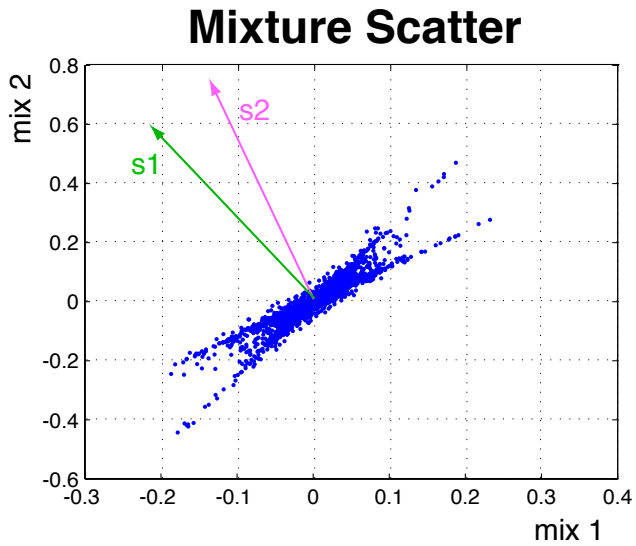
Mixtures, Scatters, Kurtosis

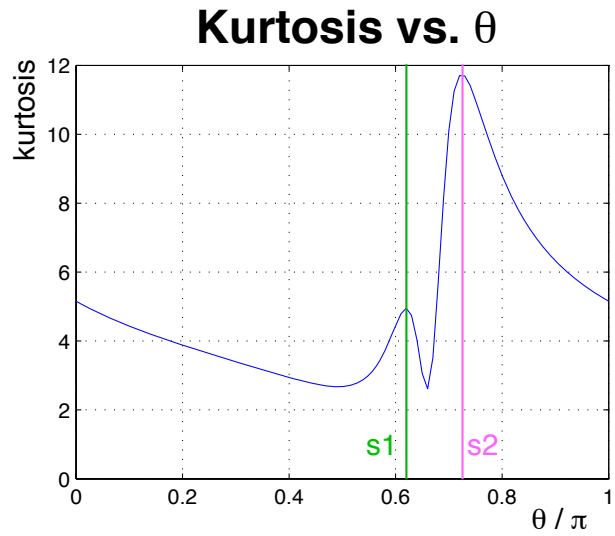
- Mixtures of sources become more Gaussian
  - can measure e.g. via 'kurtsosis' (4th moment)



ICA Limitations

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





from Parra & Spence'00


lee_ss_in_1.wav

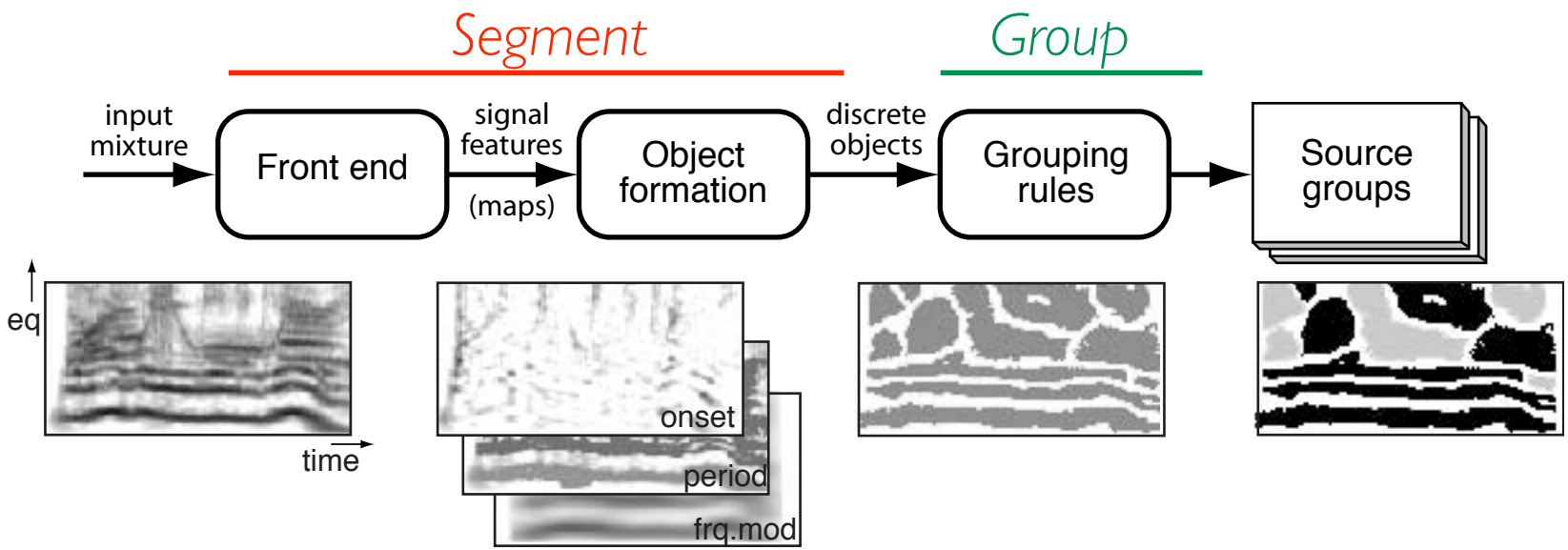

lee_ss_para_1.wav

- 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

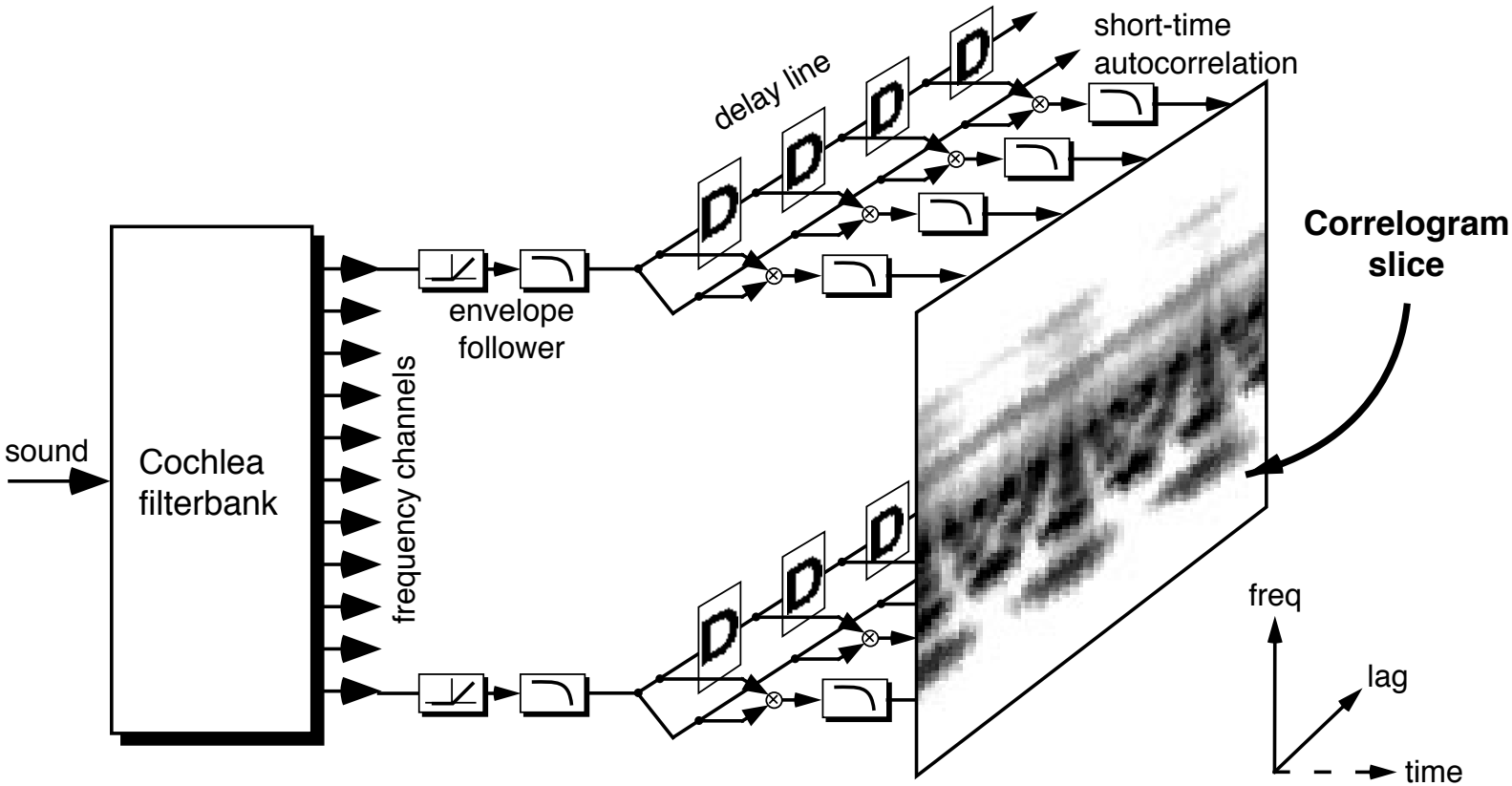


- ... principal cue is harmonicity

CASA Preprocessing

Slaney & Lyon '90

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

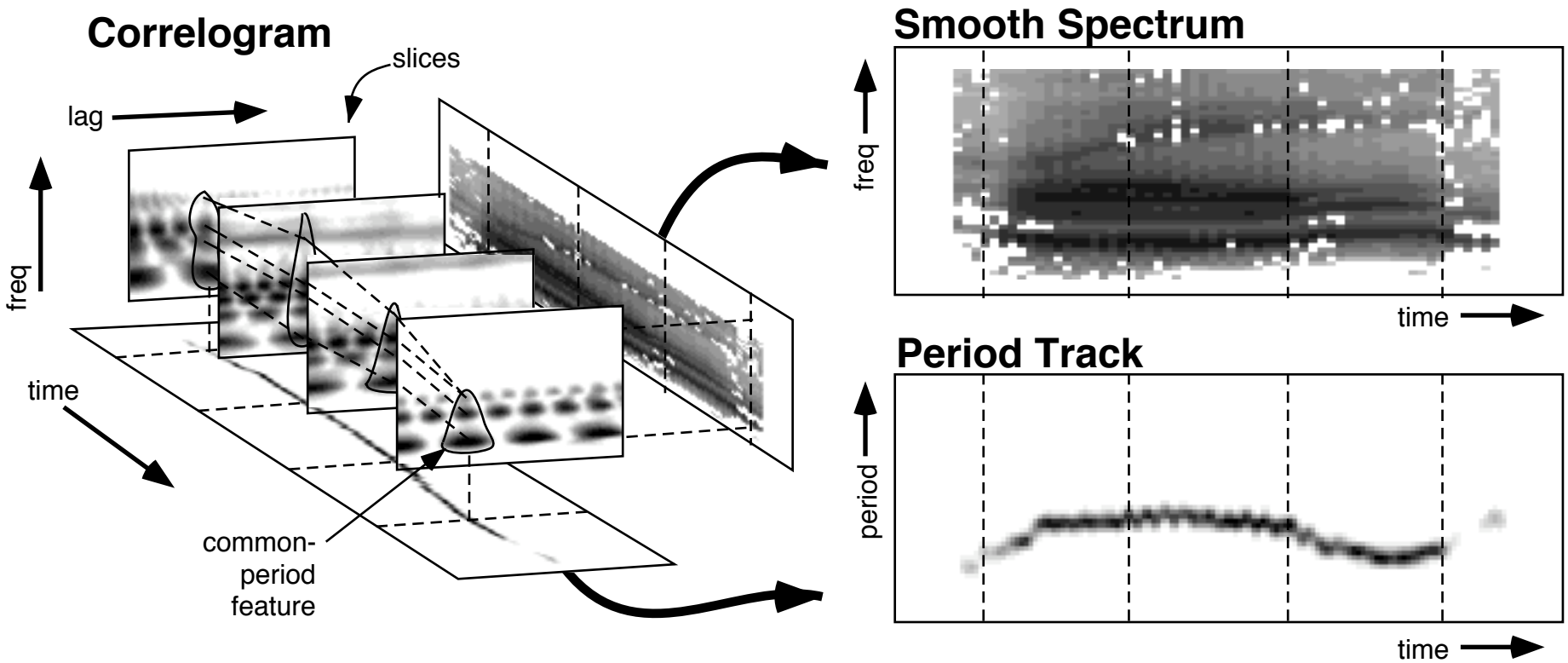


  - c/w Modulation Filtering *[Schimmel & Atlas '05]*

"Weft" Periodic Elements

Ellis '96

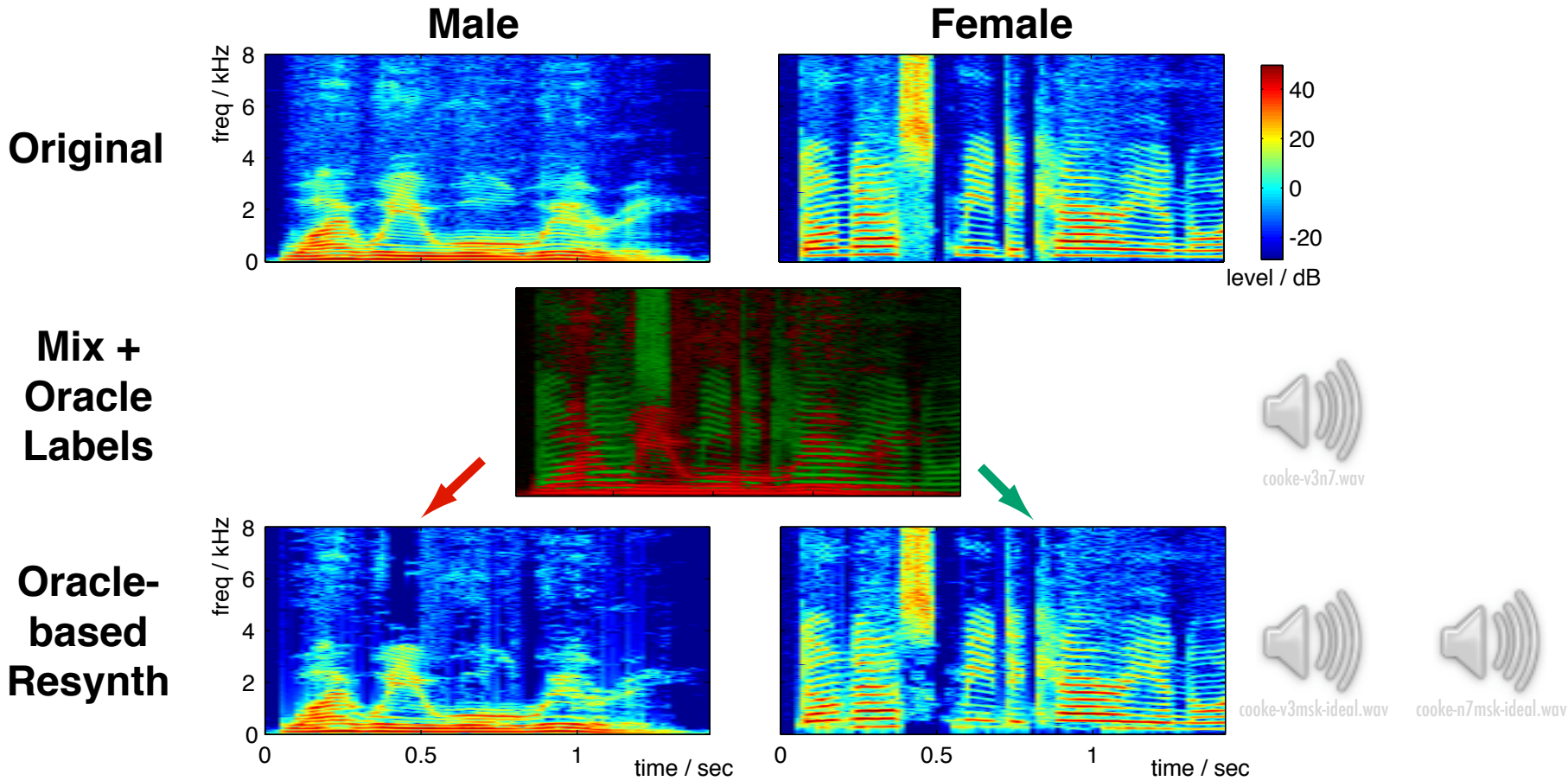
- Represent harmonics without grouping?



- hard to separate multiple pitch tracks

Time-Frequency (T-F) Masking

- "Local Dominance" assumption



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

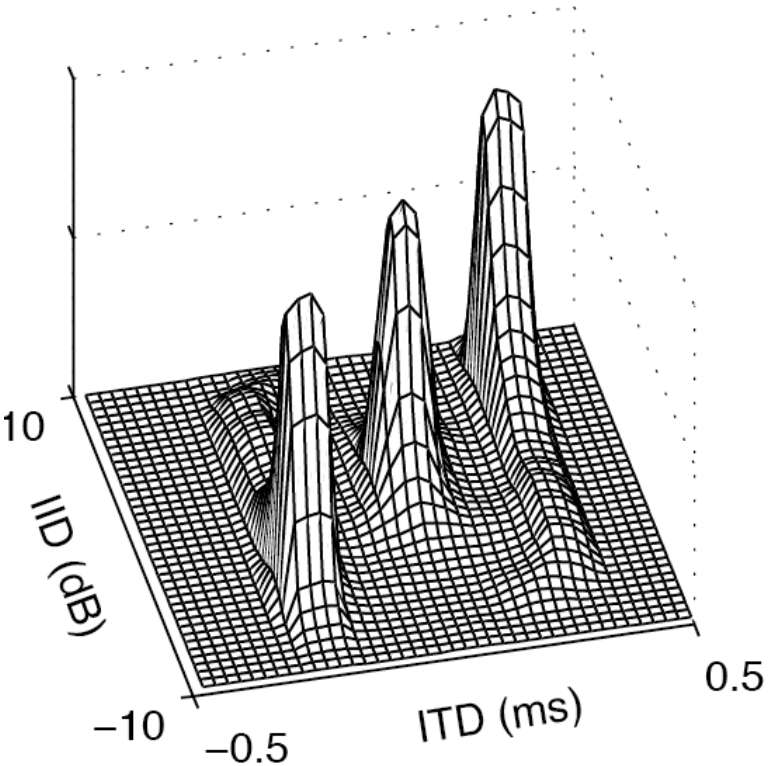
Combining Spatial + T-F Masking

- 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



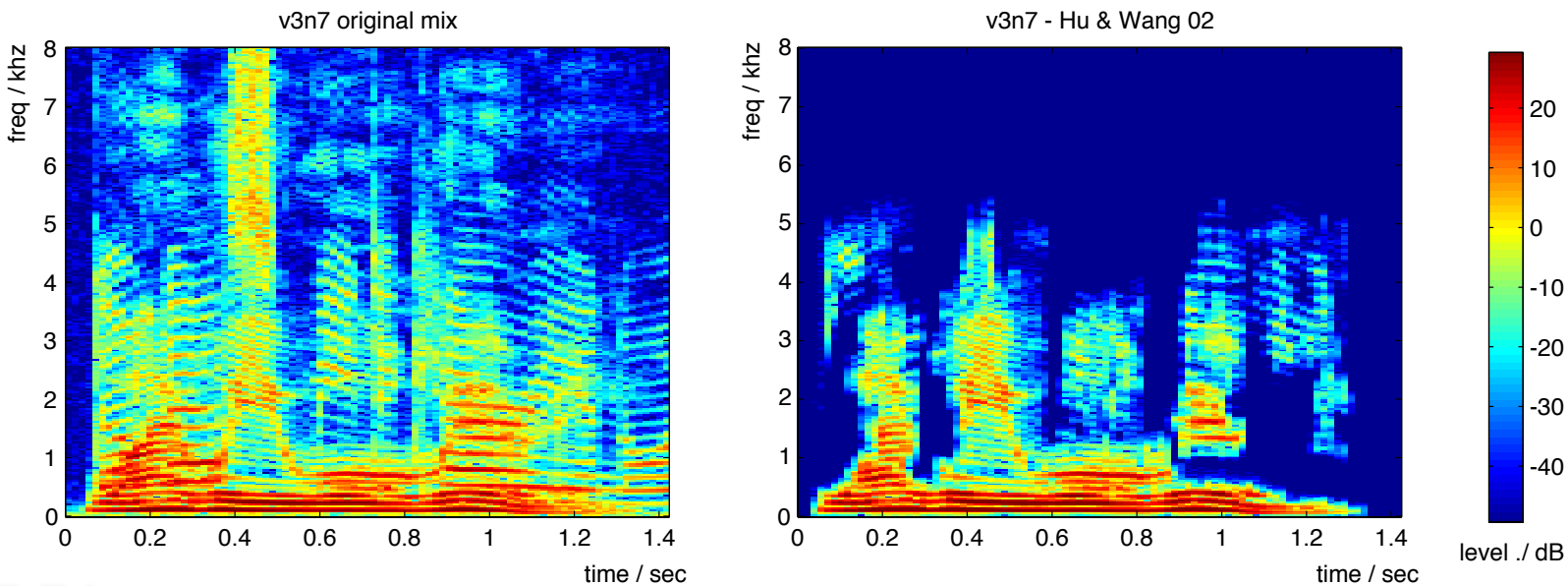
- T-F masking after ICA

  [Blin et al. '04]

  - cancellation can remove energy within T-F cells

CASA limitations

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

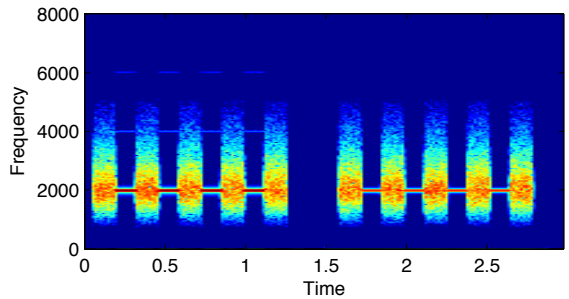


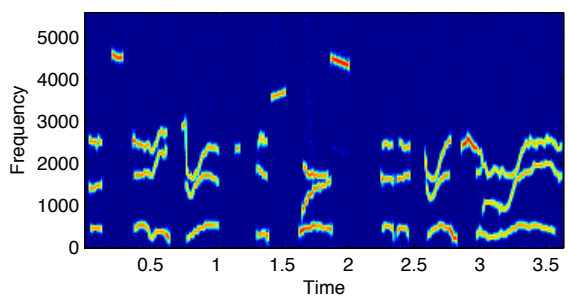
from Hu & Wang '04

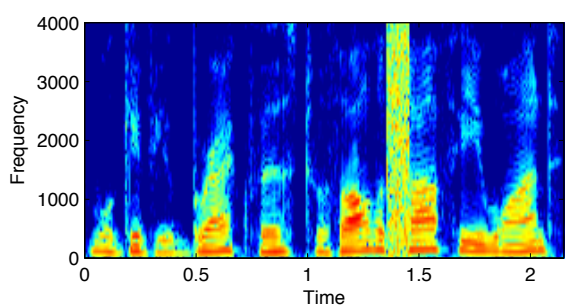
huwang-v3n7.wav

Auditory "Illusions"

- How do we explain illusions?
  - pulsation threshold
  - sinewave speech
  - phonemic restoration





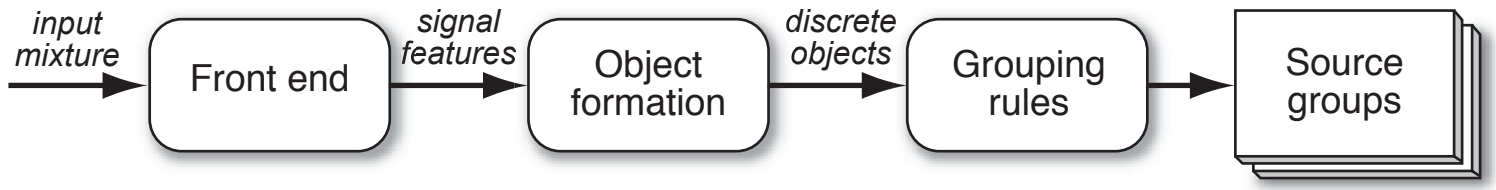


- Something is providing the missing (illusory) pieces ... source models

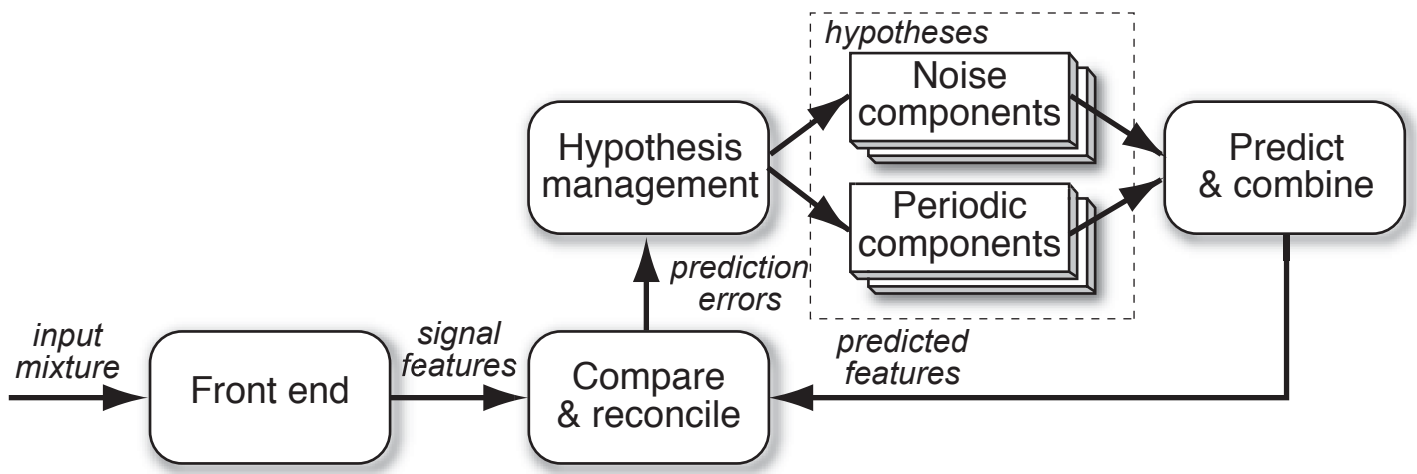
Adding Top-Down Constraints

Ellis '96

- Bottom-up CASA: limited to what's "there"



- Top-down predictions allow illusions



- match observations to a "world-model"...

Separation vs. Inference

- 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:

$$P(\{S_i\}|\, X) \propto P(X\,|\{S_i\})\; P(\{S_i\})$$

combination physics *source models*

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

Simple Source Separation

Roweis '01, '03
Kristjannson '04, '06

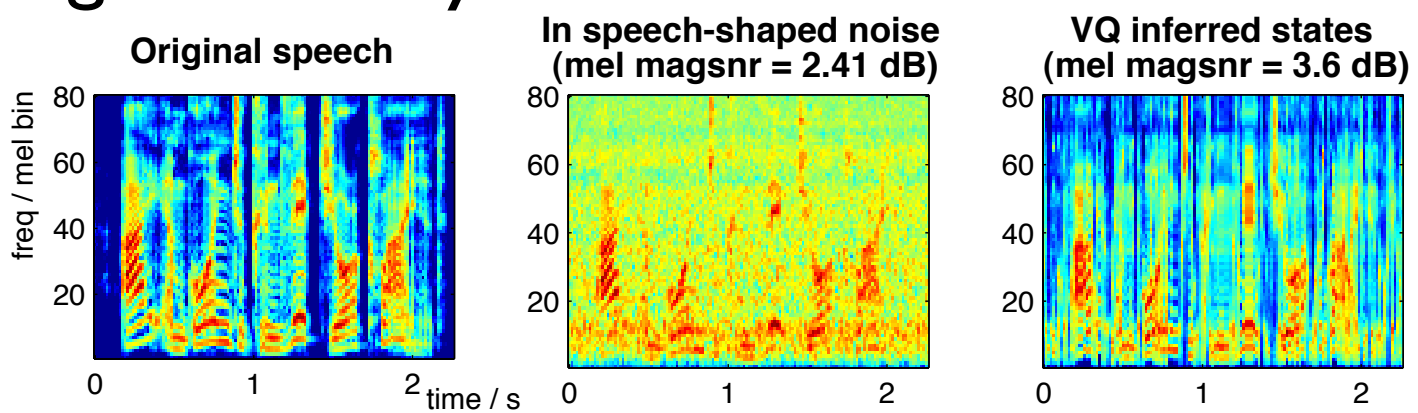
- Given models for sources, find "best" (most likely) states for spectra:

$$p(\mathbf{x}|i_1,i_2) = \mathcal{N}(\mathbf{x};\mathbf{c}_{i1}+\mathbf{c}_{i2},\Sigma)$$ *combination model*

$$\{i_1(t),i_2(t)\} = argmax_{i_1,i_2} p(\mathbf{x}(t)|i_1,i_2)$$ *inference of source state*

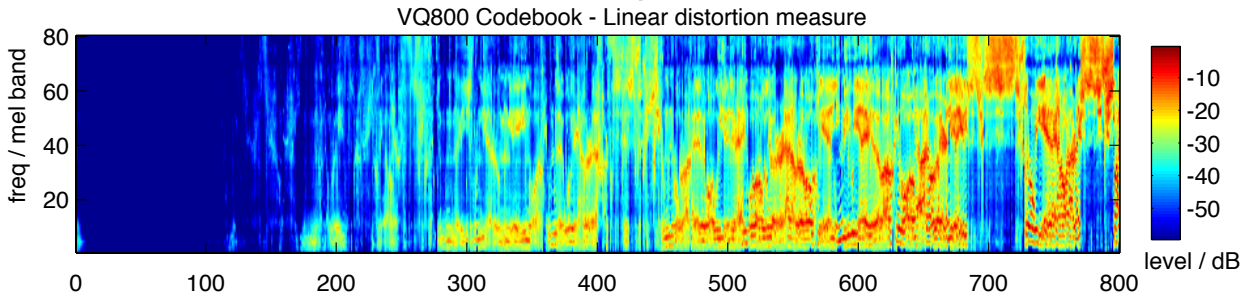
  - can include sequential constraints...
  - different domains for combining $\mathbf{c}$ and defining Σ

- E.g. stationary noise:



Can Models Do CASA?

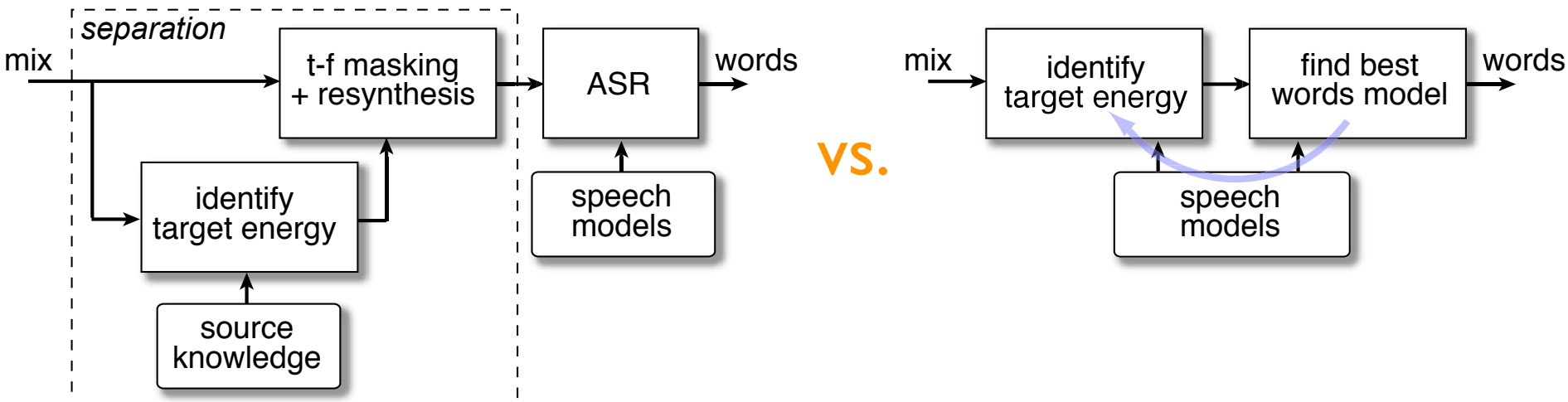
- Source models can learn harmonicity, onset
  - ... to subsume rules/representations of CASA



  - 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 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

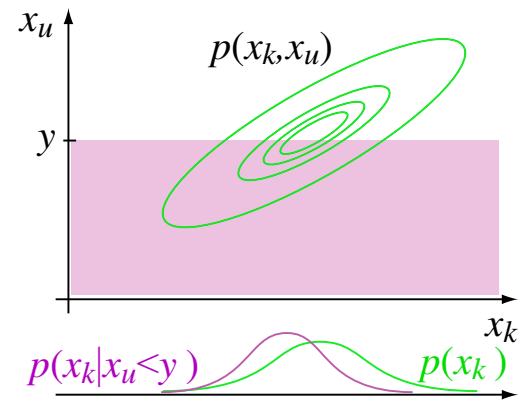


  - words output = abstract description of signal

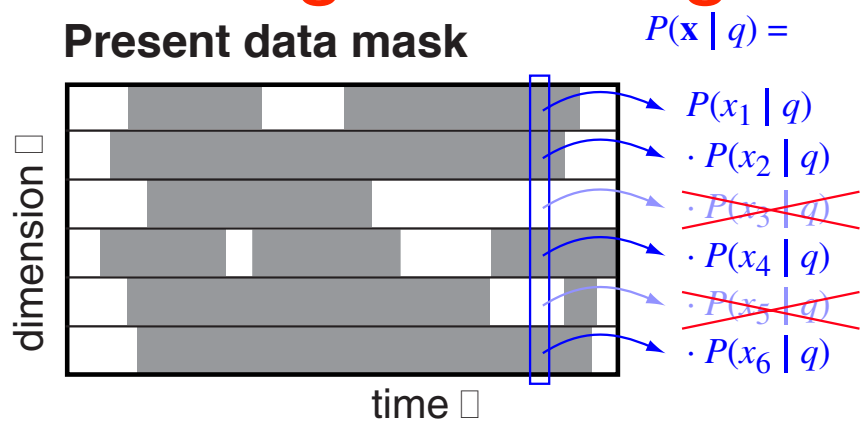
Missing Data Recognition

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:

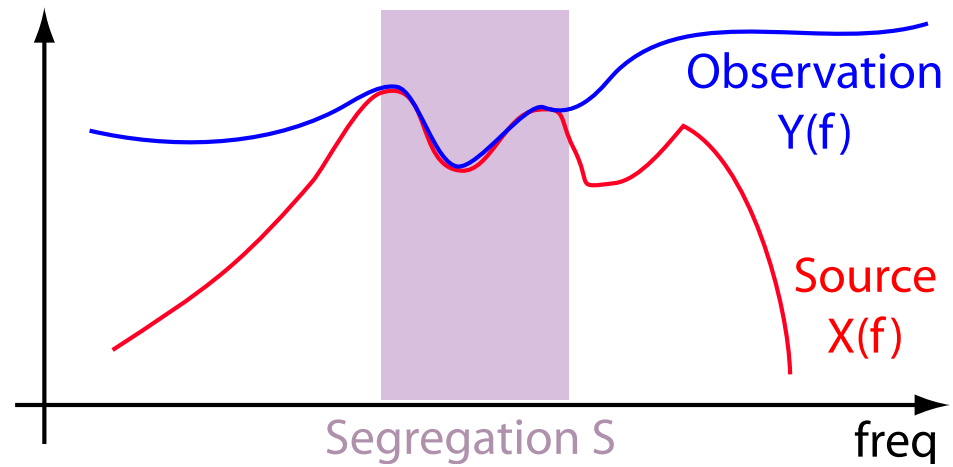


  - hard part is finding the mask (segregation)

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:

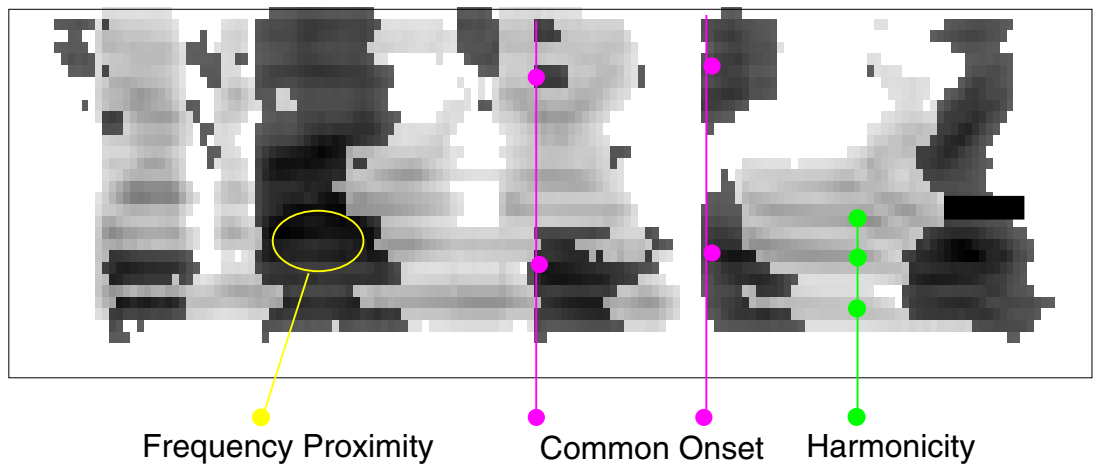
$$P(M,S|Y) = P(M)\int P(X|M)\cdot\frac{P(X|Y,S)}{P(X)}dX\cdot P(S|Y)$$

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:



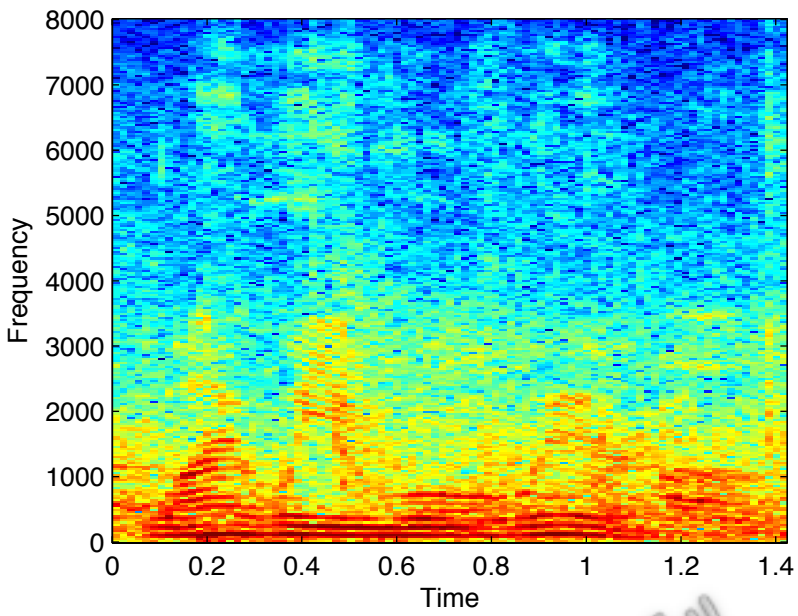
Outline

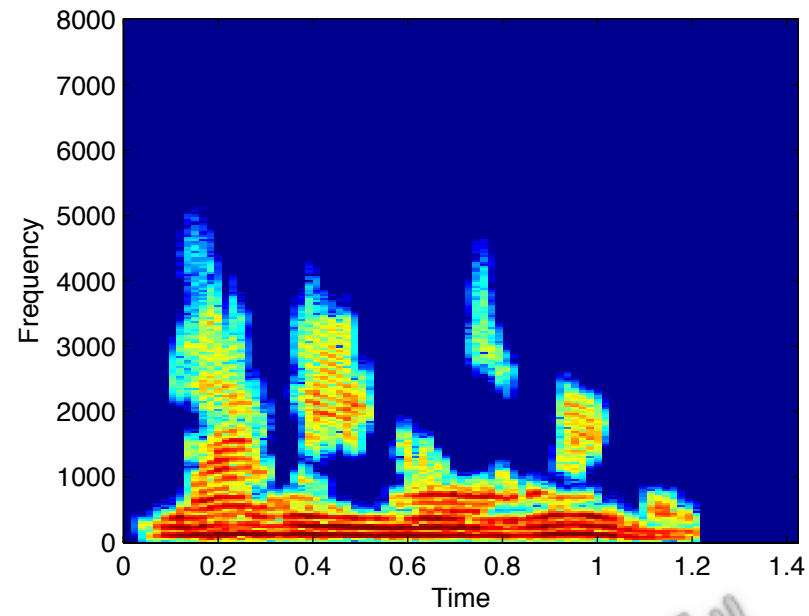
1. The ASA Problem
2. Human ASA
3. Machine Source Separation
4. **Systems & Examples**
   - Periodicity-based
   - Model-based
   - Music signals
5. Concluding Remarks

Current CASA

Hu & Wang'03

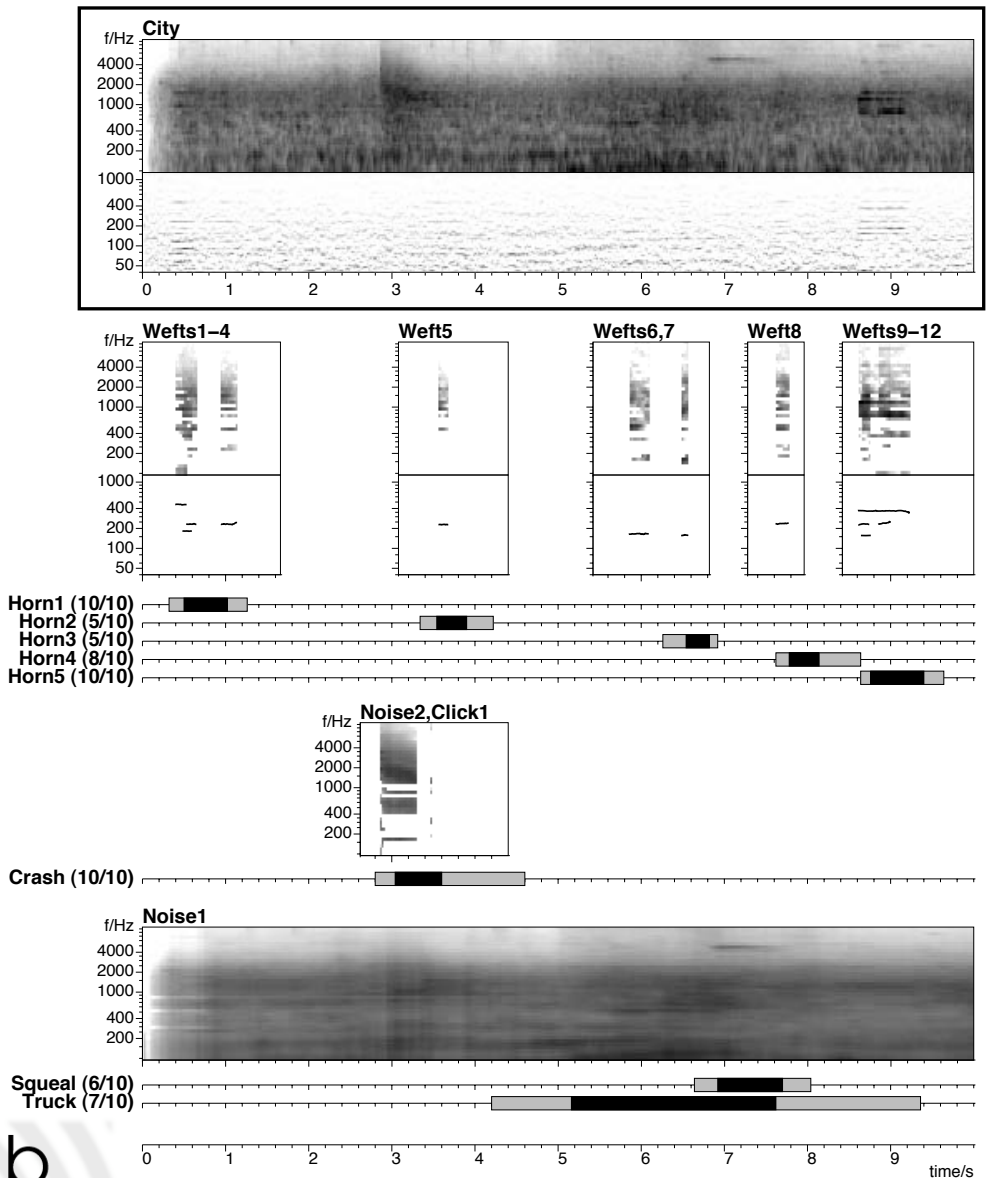
- **State-of-the-art bottom-up separation**
  - noise robust pitch track
  - label T-F cells by pitch
  - extensions to unvoiced transients by onset





Prediction-Driven CASA

Ellis'96

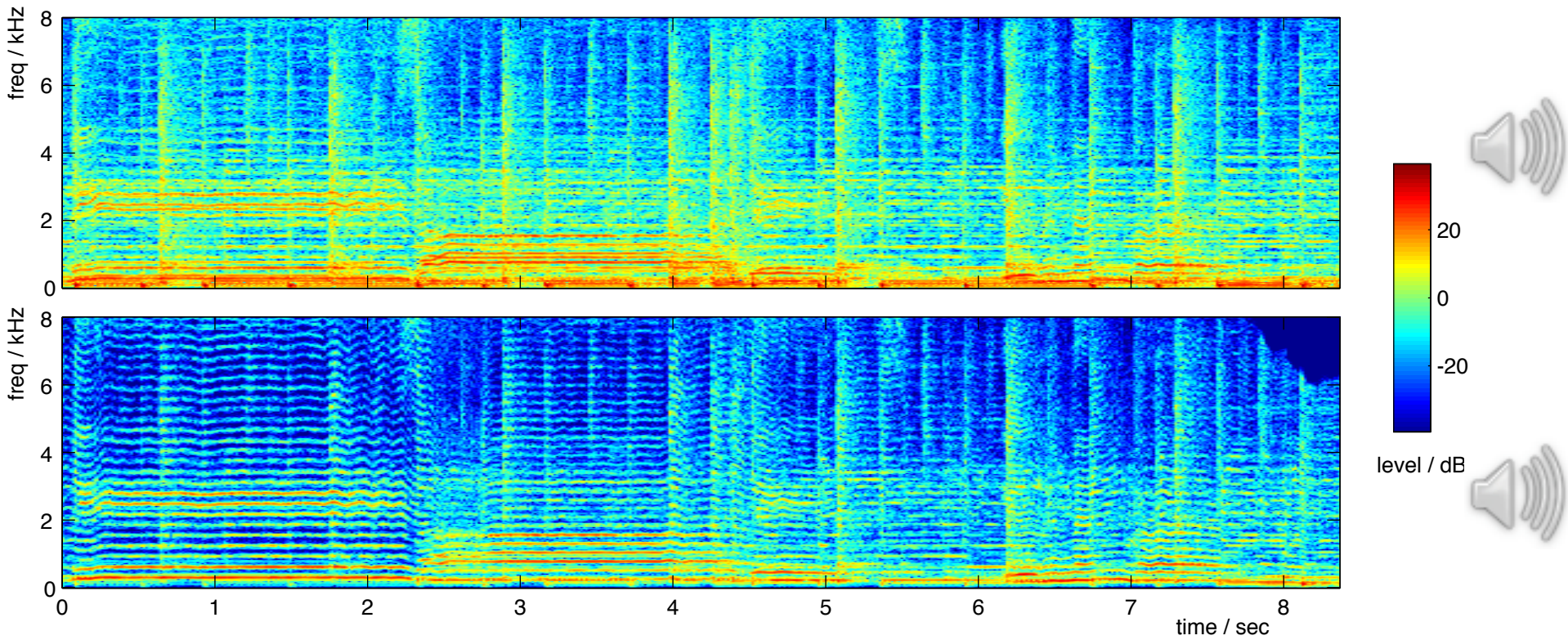


- Identify objects in real-world scenes
  - using "sound elements"

Singing Voice Separation

Avery Wang'94

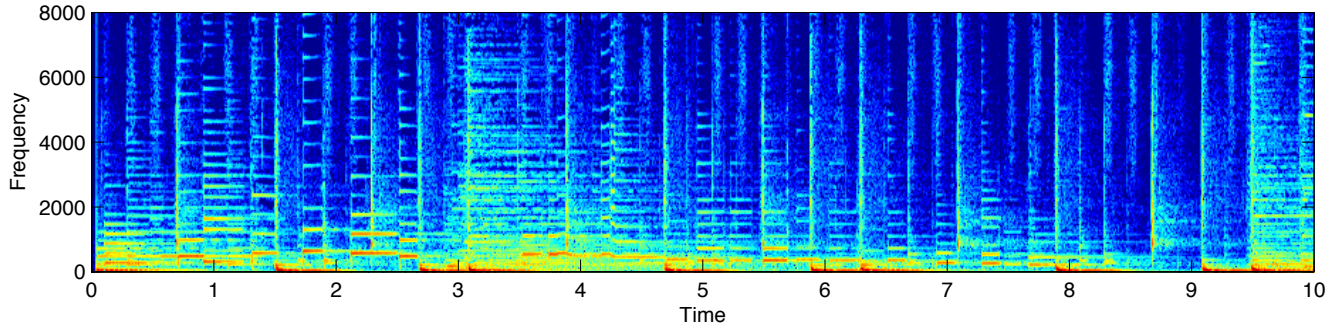
- Pitch tracking + harmonic separation

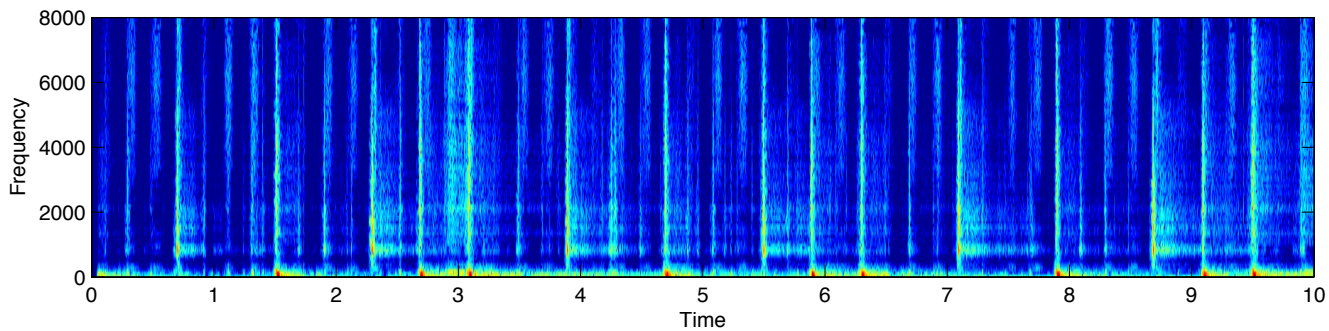


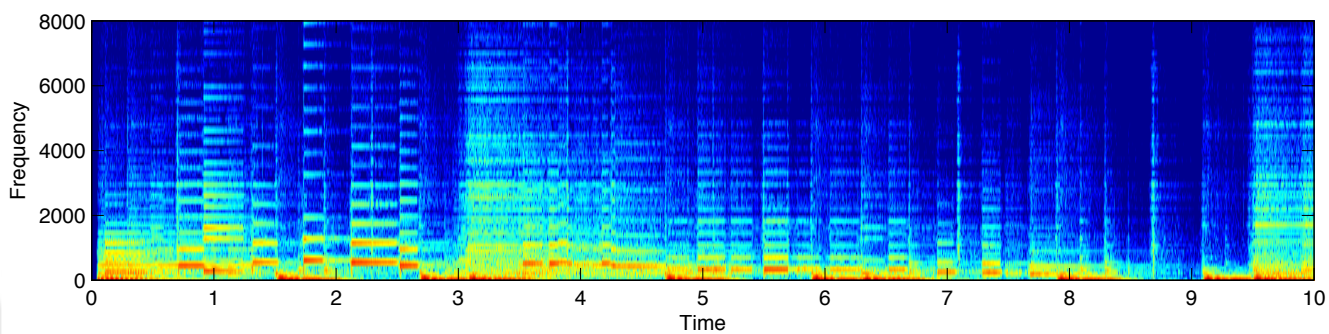
Periodic/Aperiodic Separation

Virtanen'03

- Harmonic structure + repetition of drums







"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"

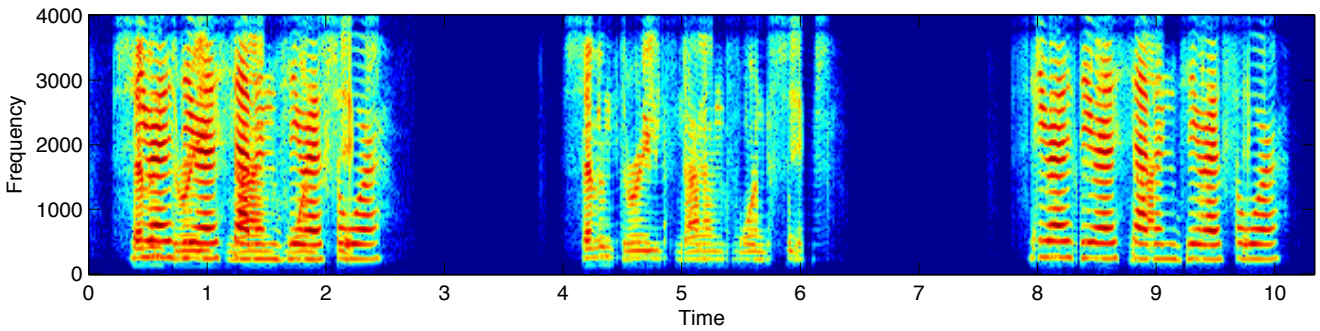
t5_bwam5s_m5_bbilzp_6p1.wav

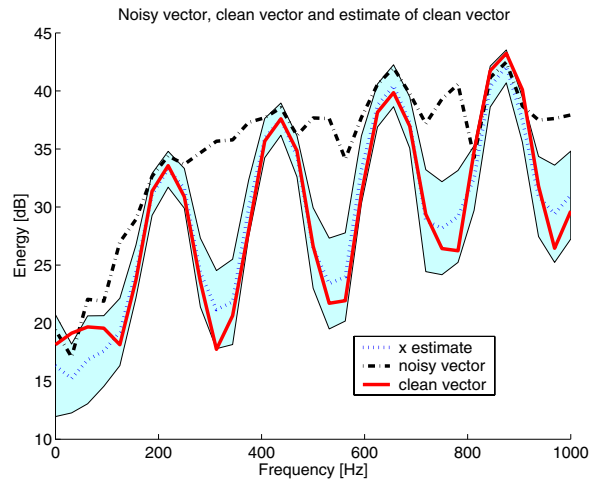
- Results to be presented at Interspeech'06
  - http://www.dcs.shef.ac.uk/~martin/SpeechSeparationChallenge.htm
- See also "Statistical And Perceptual Audition" workshop
  - http://www.sapa2006.org/

IBM's "Superhuman" Separation

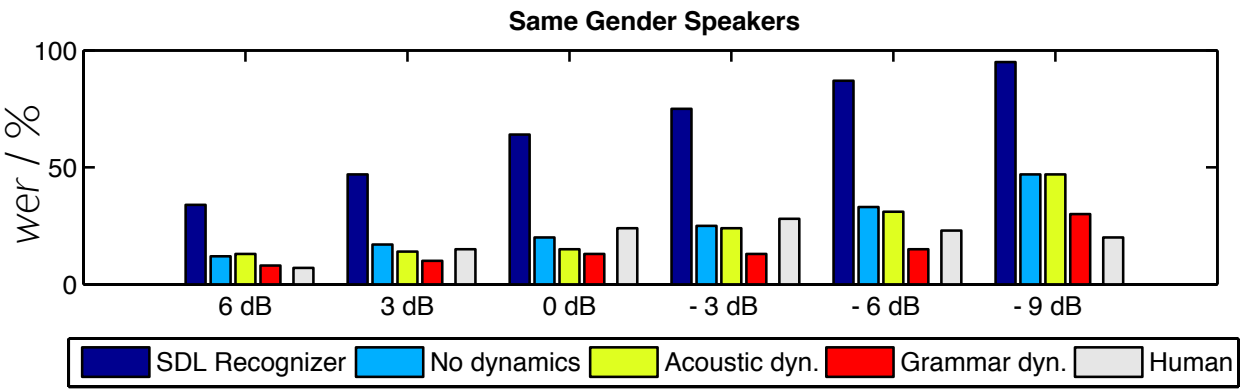
Kristjansson et a
Interspeech'06

- Optimal inference on Mixed Spectra
  - model each speaker (512 mix GMM)





- Applied to Speech Separation Challenge:



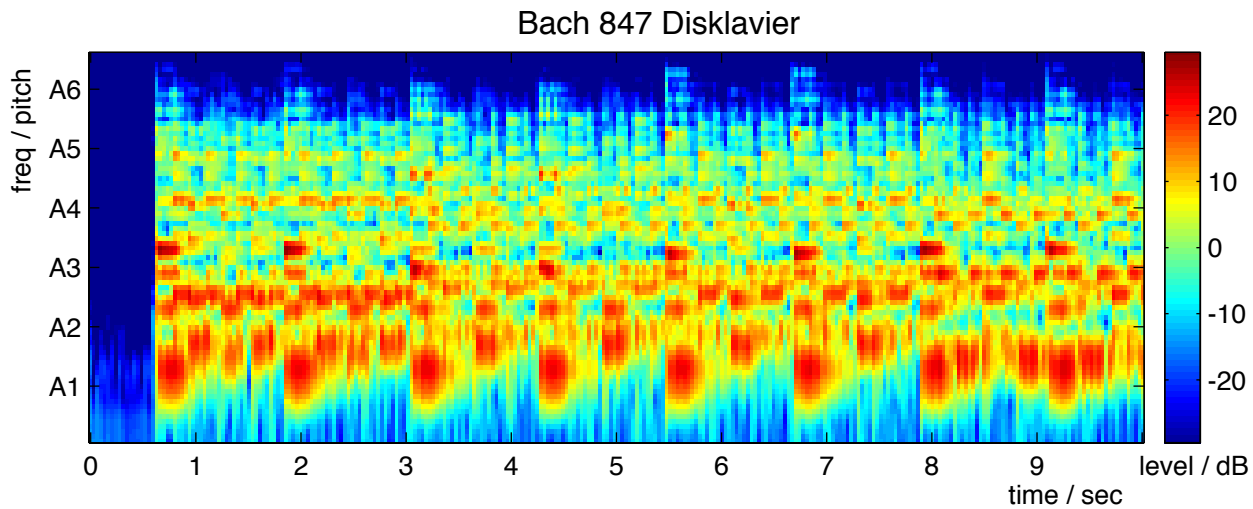
  - Infer speakers and gain
  - Reconstruct speech
  - Recognize as normal...
- Use grammar constraints

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

Columbia University
In the City of New York

Transcription as Separation

- Transcribe piano recordings by classification
  - train SVM detectors for every piano note
  - 88 separate detectors, independent smoothing
- Trained on player piano recordings



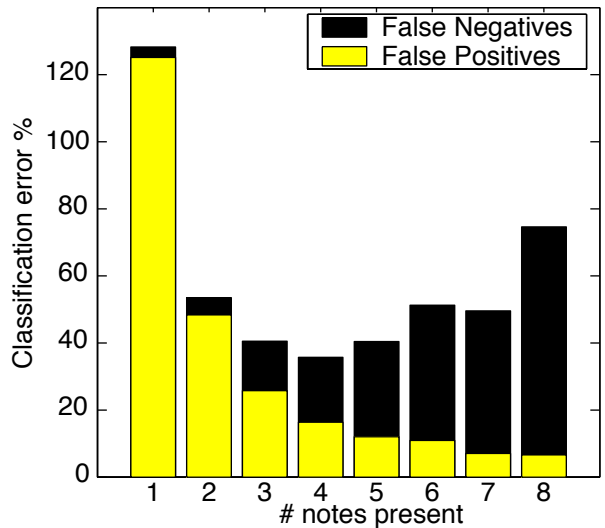
- Sse transcription to resynthesize...

Piano Transcription Results

- Significant improvement from classifier:
  - frame-level accuracy results:

Algorithm	Errs	False Pos	False Neg	d'
SVM	43.3%	27.9%	15.4%	3.44
Klapuri&Ryynänen	66.6%	28.1%	38.5%	2.71
Marolt	84.6%	36.5%	48.1%	2.35

  - Breakdown by frame type:



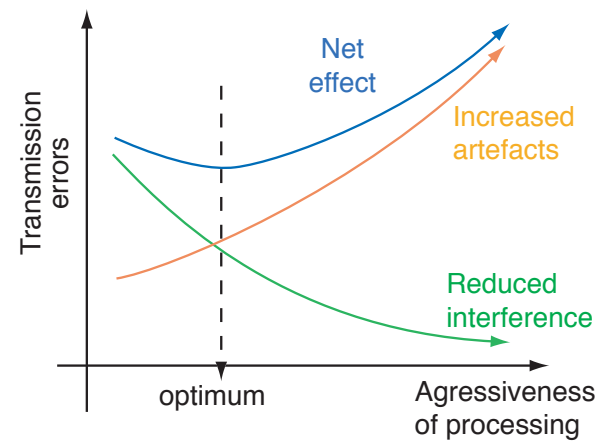
  - http://labrosa.ee.columbia.edu/projects/melody/

Outline

1. The ASA Problem
2. Human ASA
3. Machine Source Separation
4. Systems & Examples
5. **Concluding Remarks**
   - Evaluation

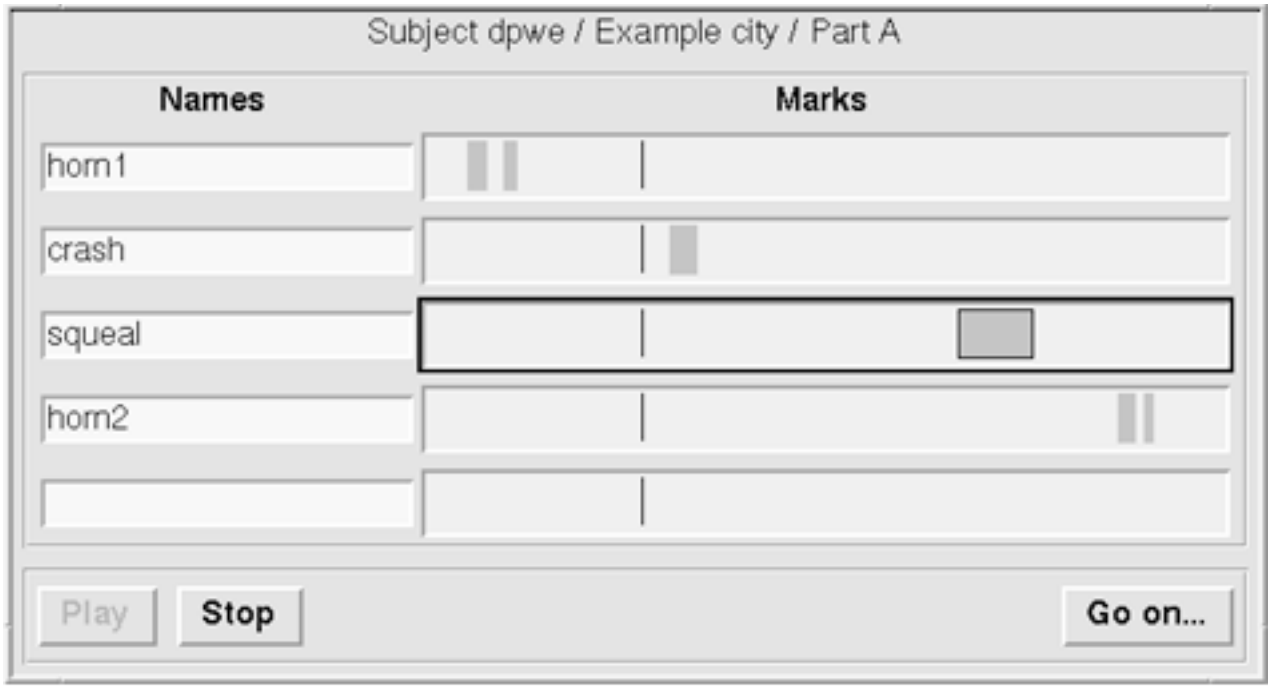
Evaluation

- 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



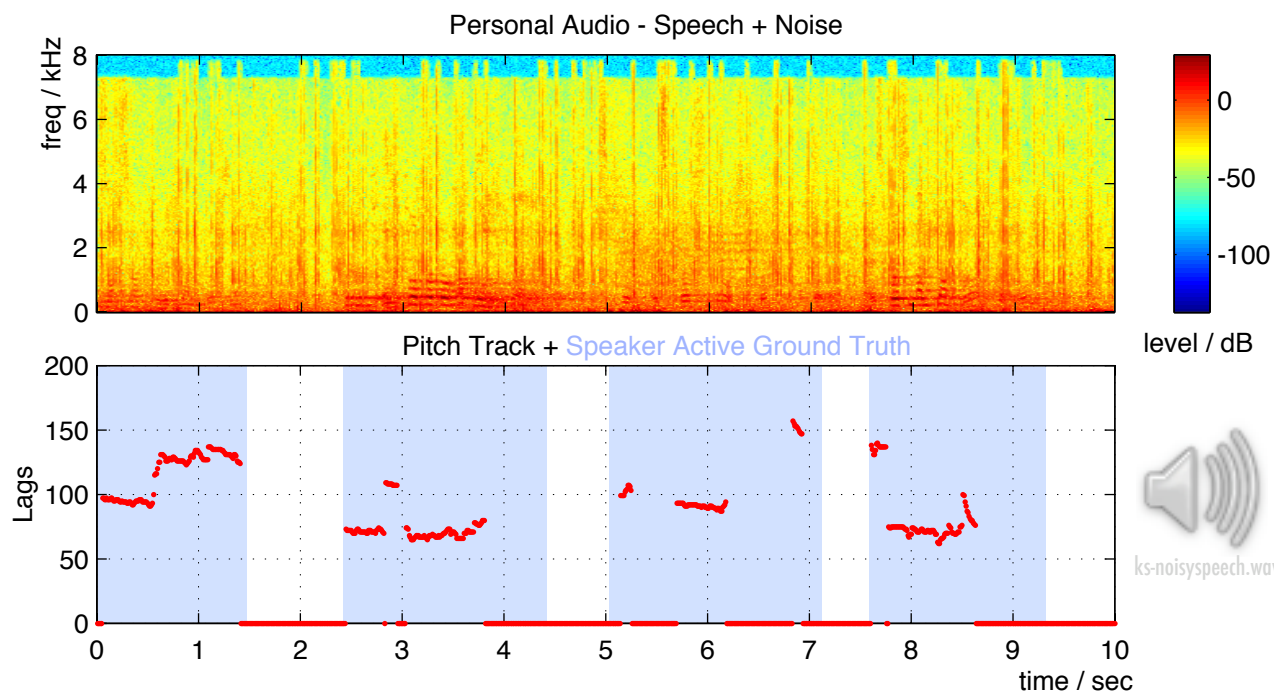
Evaluating Scene Analysis

- Need to establish ground truth
  - subjective sources in real sound mixtures?



More Realistic Evaluation

- Real-world speech tasks
  - crowded environments
  - applications: communication, command/control, transcription
- Metric
  - human intelligibility?
  - 'diarization' annotation (not transcription)

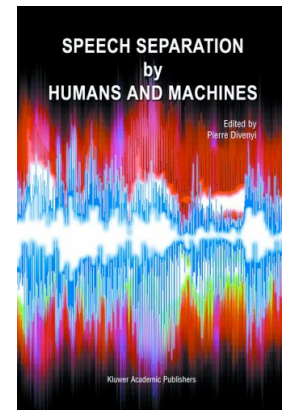


Summary & Conclusions

- Listeners do well separating sound mixtures
  - using signal cues (location, periodicity)
  - 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
- Separation feasible in certain domains
  - describing source properties is easier

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.