About this Non-Negative Business

Paris Smaragdis
paris@illinois.edu

University of Illinois at Urbana-Champaign & Adobe Research

10 years ago to the day …

2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics October 19-22, 2003, New Paltz, NY

DINNER 6:00PM-7:20PM

SESSION N: RESYNTHESIS AND CROSS-SYNTHESIS 7:20PM-8:20PM

7:20pm **Rejection Phenomena in Inter-Signal Voice Transplantations**
Werner Verhelst and Henk Brouckxon, *Vrije Universiteit Brussel, Brussels, Belgium*

7:40pm **Discrimination of Sustained Musical Instrument Sounds Resynthesized With Randomly Altered Harmonic Amplitudes**
Andrew B. Horner, *Hong Kong University of Science and Technology, Kowloon, Hong Kong*
James W. Beauchamp, *University of Illinois at Urbana-Champaign, Urbana, IL, USA*

8:00pm **Time-Scale Modification of Music Using a Subband Approach Based on the Bark Scale**
David Dorran, *Dublin Institute of Technology, Dublin, Ireland*
Robert Lawlor, *National University of Ireland, Maynooth, Ireland*

SESSION O: MUSIC SIGNAL PROCESSING - MUSIC TRANSCRIPTION 8:20PM-9:20PM

8:20pm **Non-Negative Matrix Factorization for Polyphonic Music Transcription**
Paris Smaragdis, *Mitsubishi Electric Research Lab, Cambridge, MA, USA*
Judith C. Brown, *Wellesley College, Wellesley, MA, USA*

8:40pm **Generative Model Based Polyphonic Music Transcription**
Ali Taylan Cemgil and Bert Kappen, *University of Nijmegen, The Netherlands*
David Barber, *Edinburgh University, UK*

What is this talk about?

- What are all these "non-negative" papers?
- What is special about this approach?
- What can we do with it?
  - And why should we bother?

Traditional signal processing

- Axiom 1: “Thou shall love the Gaussian”
  - Why? It makes the math easy

- Gave rise to least squares models:

$$y(t) = x(t) + n(t)$$

What we get

What we want

Gaussian noise

A misunderstood model

- Abusing the noise model

$$y(t) = x(t) + n(t)$$

Target sound

"Other sounds"

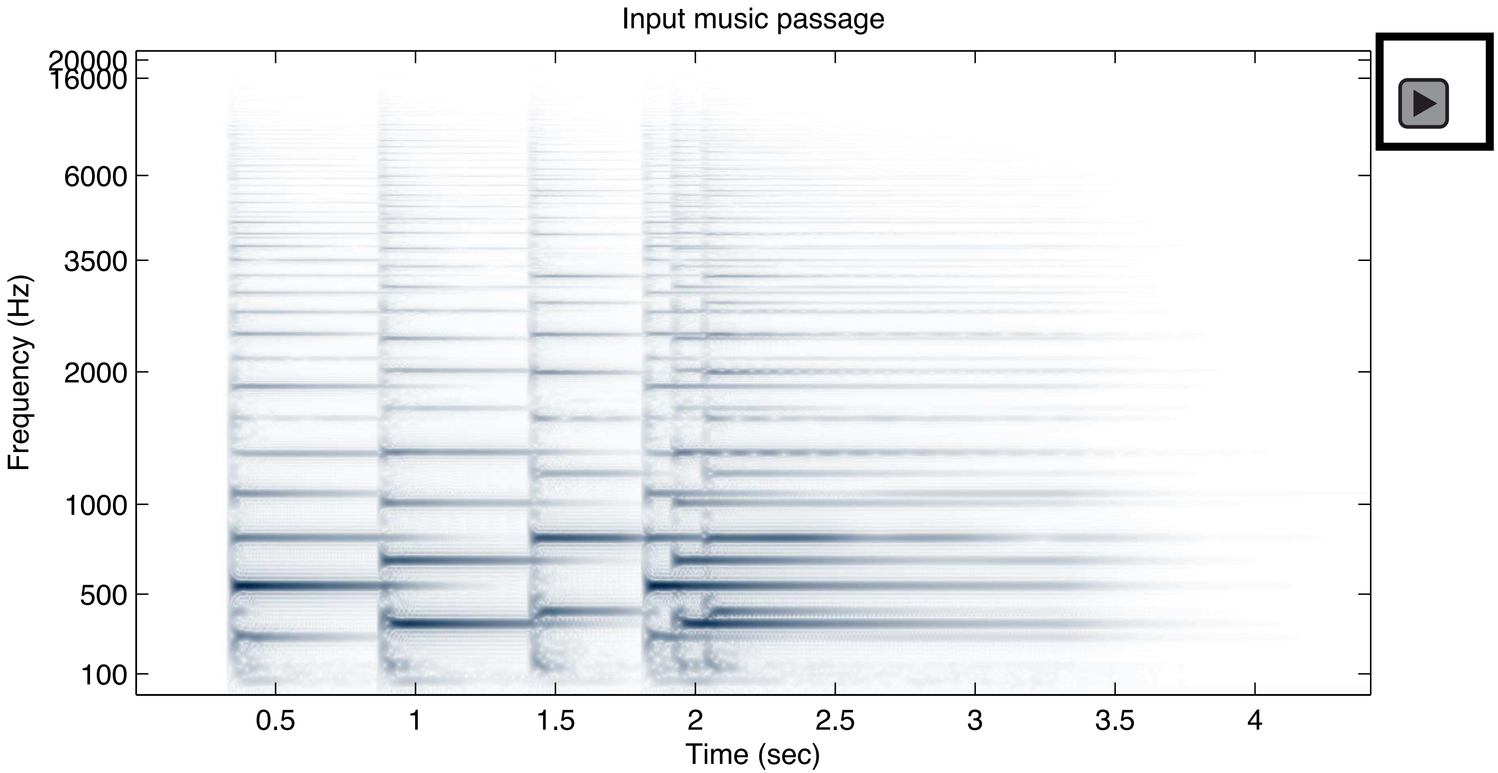
- Other sounds are not Gaussian noise!
  - In fact neither is your target sound

And the impending revolution

- mid-90's: The ICA community
  - Sources are not really Gaussian
- mid-2000's: Compressive Sensing
  - Data is sparse in the right domain
- mid-2000's: Non-Negative Models
  - We only care about positive-valued quantities

Picking a meaningful domain

- Waveforms are not that intuitive, we instead use spectrograms to examine audio signals

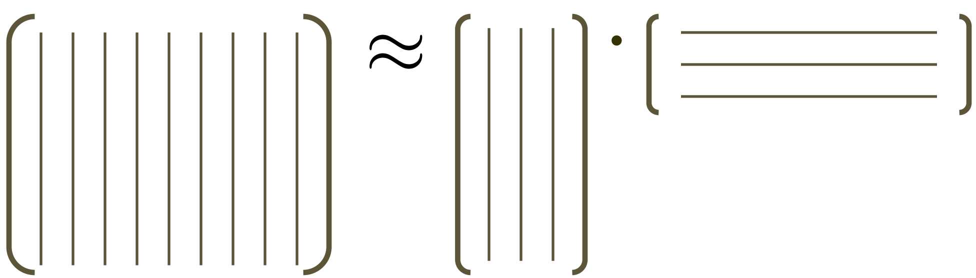


Decomposing spectrograms

- What are the building blocks of spectrograms?
  - Standard question in machine learning

- The low-rank matrix factorization:

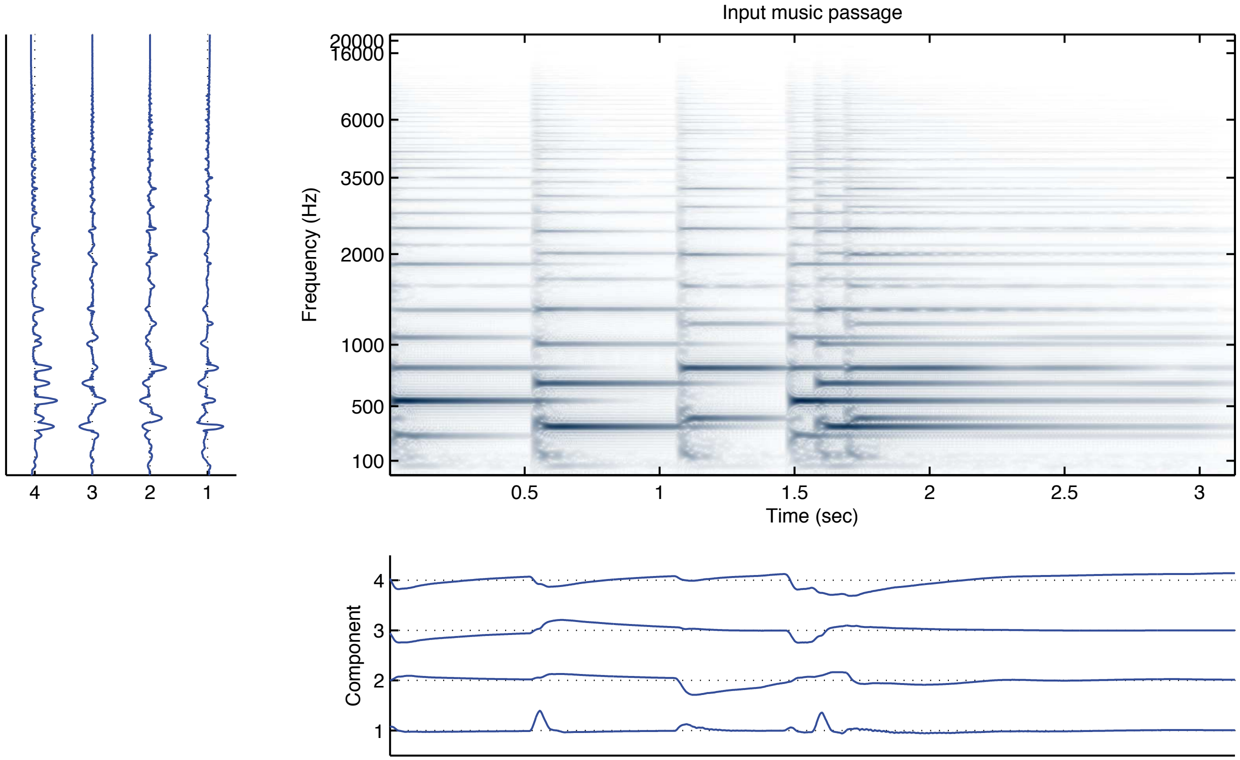
$$\mathbf{X} \approx \mathbf{W} \cdot \mathbf{H}$$



The usual suspect

- Principal Component Analysis: $\mathbf{X} = \mathbf{W} \cdot \mathbf{H}$

orthonormal *decorrelated*



Why is this result meaningless?

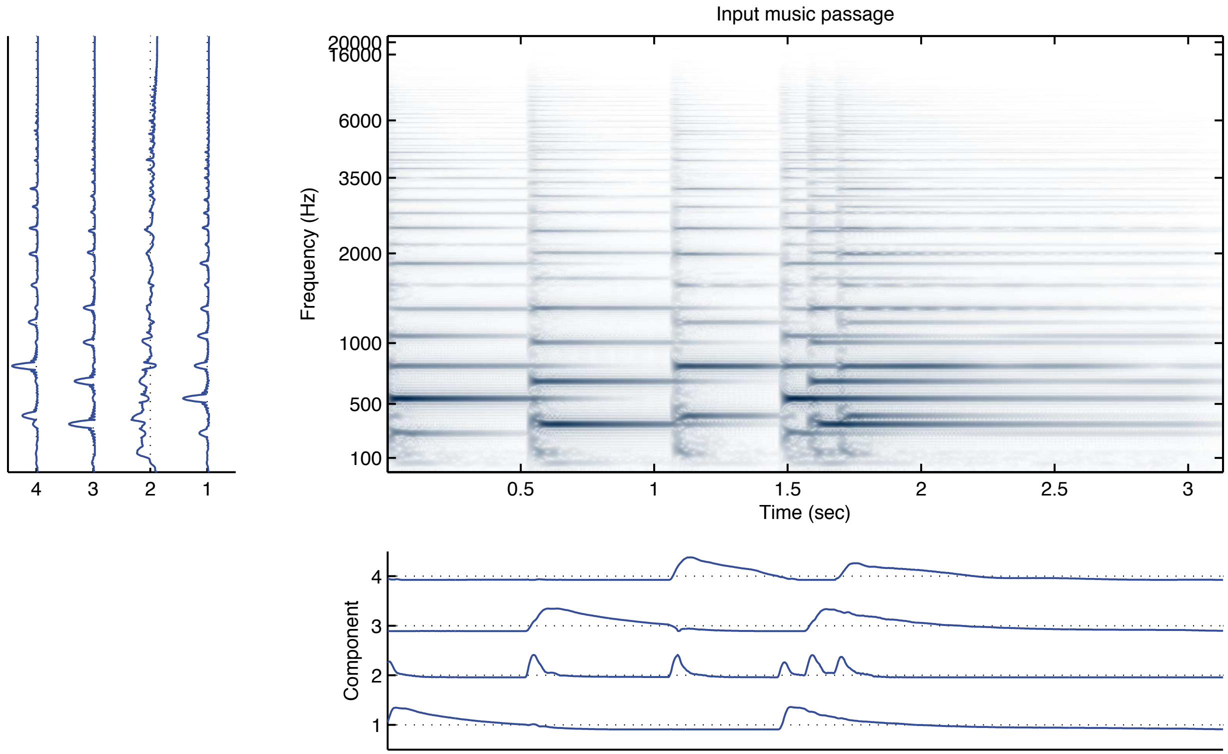
- This least-squares/Gaussian model is counter-intuitive for sound
  - Makes use of cross-cancellation

- We perceive scenes additively
  - We need an additive decomposition!

Non-Negative Matrix Factorization

- All factors are positive-valued: $\mathbf{X} \approx \mathbf{W} \cdot \mathbf{H}$
  - Resulting reconstruction is additive

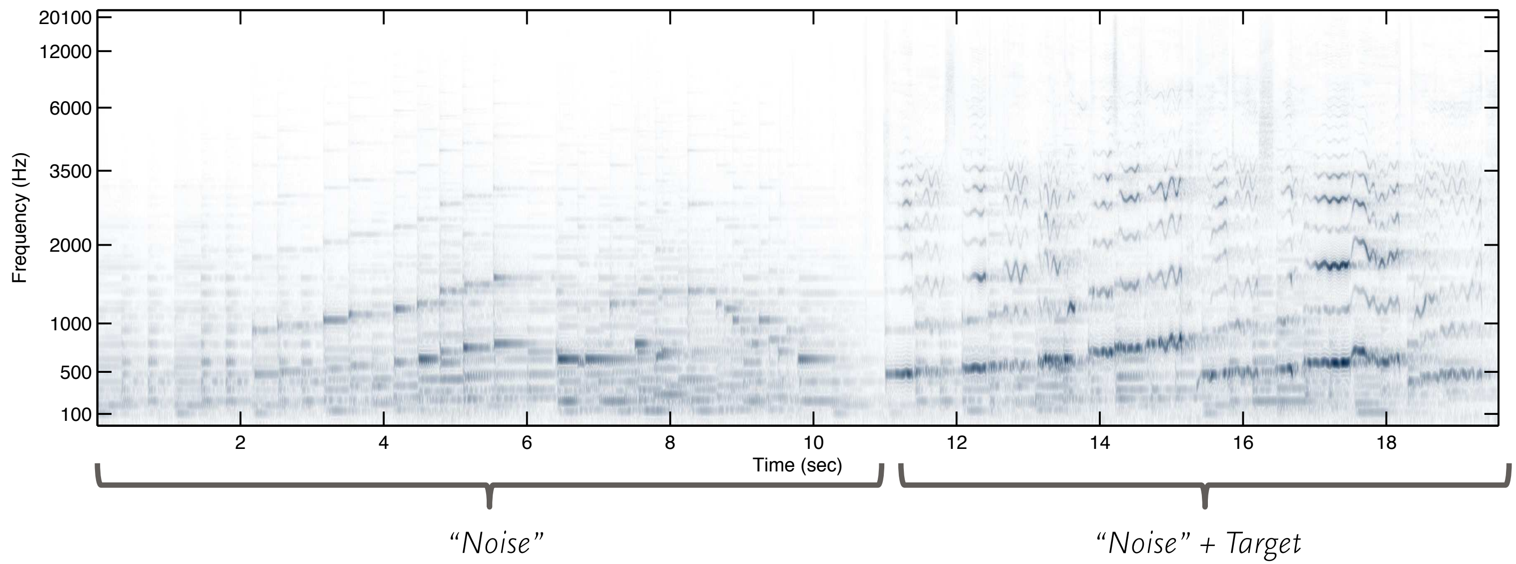
Non-negative



Why is this a better model?

- 1) It allows us to intuitively model sounds
  - All quantities mean something
- 2) The model parameters are additive
  - This also means we are invariant to mixtures
- We can easily redefine previous work
  - And reap the benefits!

Wiener filtering / Spectral subtraction

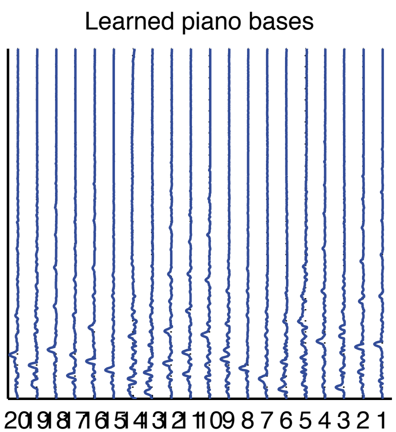


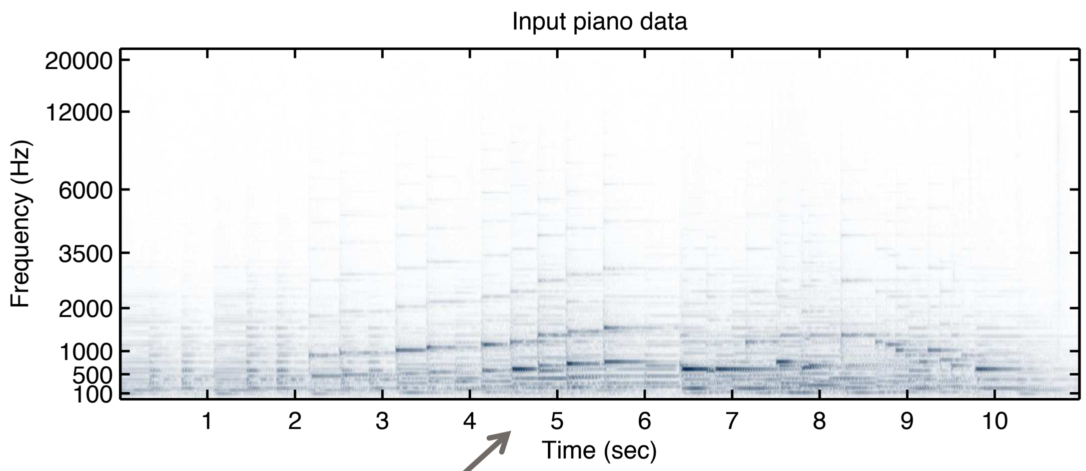
- Learn "noise" spectrum, and filter/subtract
  - And it doesn't work with complex noises ...
  - Extra complications due to negative values

The non-negative version

- Learning a sound model
  - An additive dictionary instead of a spectrum

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

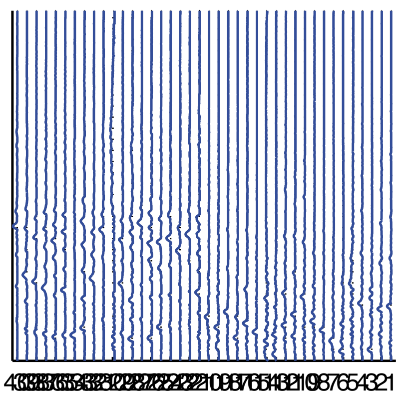




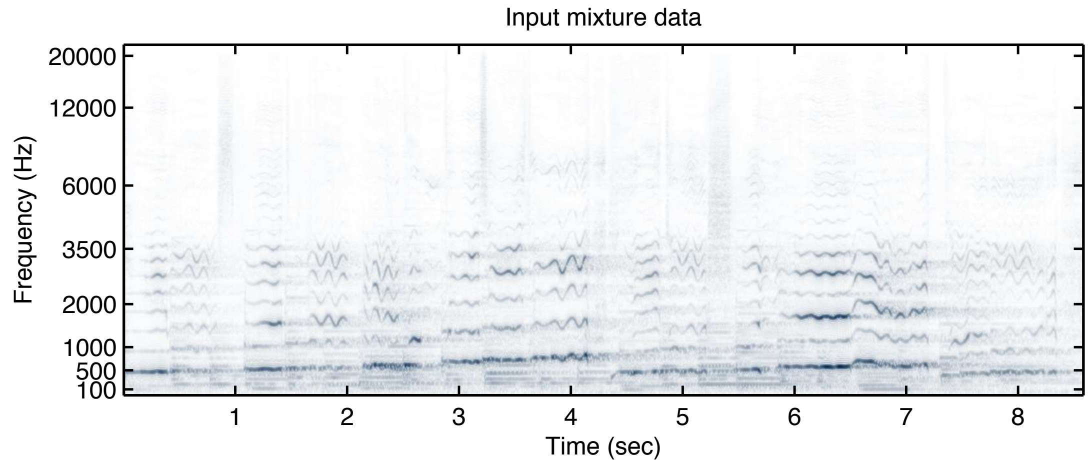
Linear combinations of these, explain these

Denoising

- Explain a mixture with the existing model
  - Add new elements to explain the rest of the signal



Extra bases Known model



- Still the same model $\mathbf{X} \approx \begin{bmatrix} \mathbf{W}_u & \mathbf{W}_k \end{bmatrix} \cdot \begin{bmatrix} \mathbf{H}_u \\ \mathbf{H}_k \end{bmatrix}$

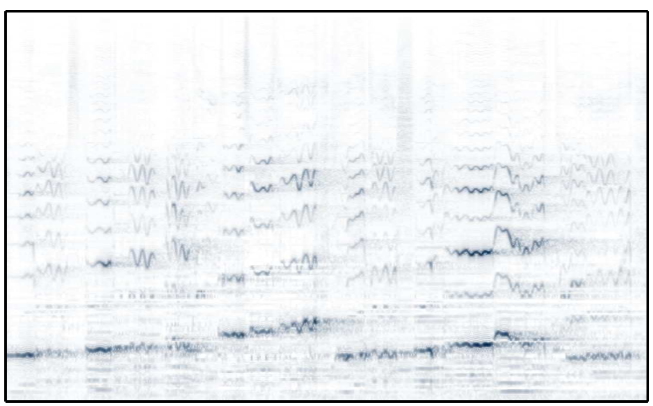
Known
Estimated

Reconstruction

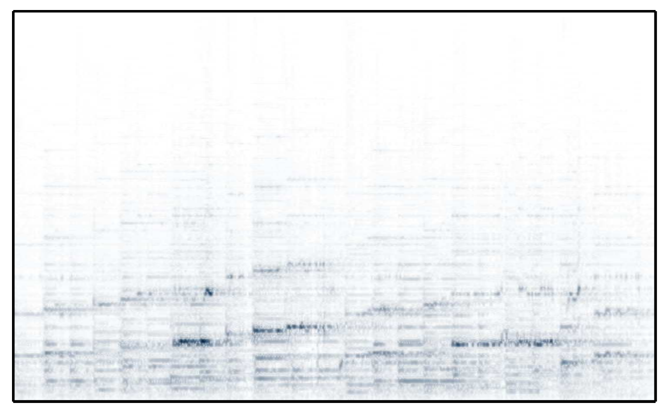
- Parts-wise reconstruction:

$$\mathbf{X} = \mathbf{X}_u + \mathbf{X}_k \approx \underbrace{\mathbf{W}_u \cdot \mathbf{H}_u}_{\textit{Spectrogram of unknown target}} + \underbrace{\mathbf{W}_k \cdot \mathbf{H}_k}_{\textit{Spectrogram of known "noise"}}$$

Extracted target



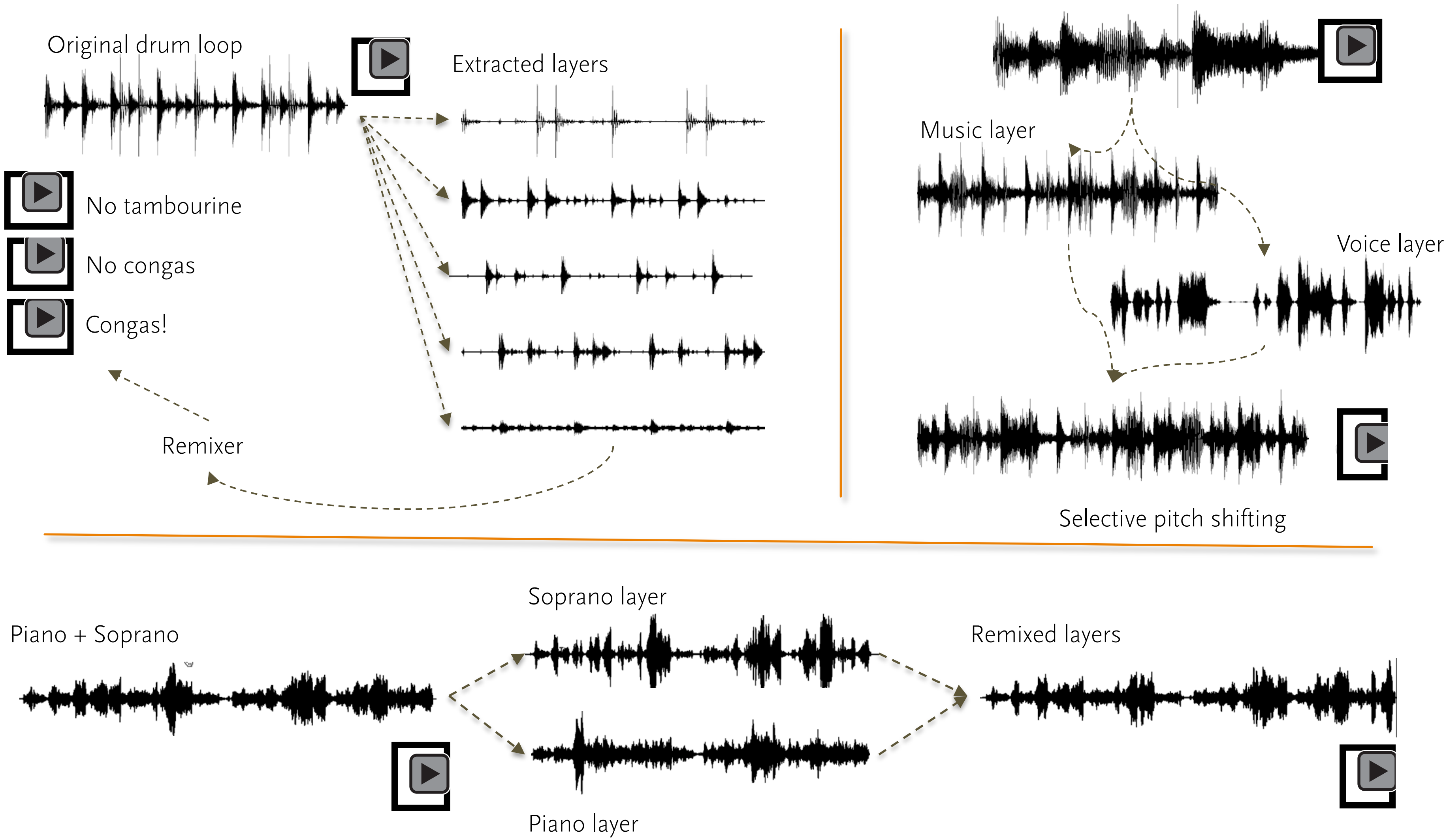
Extracted "noise"



Why bother?

- Better statistical fit for the data
  - Results in better sounding outputs
- Flexible learning of "noise" model
  - No need to simply temporally segment
    - Spatial guidance, user guidance, TF guidance, ...
- Demo time!

Layer editing options



So what?

- We can resolve mixtures well
  - But what's the use of that?
  - My mantra: "Separation is useless"

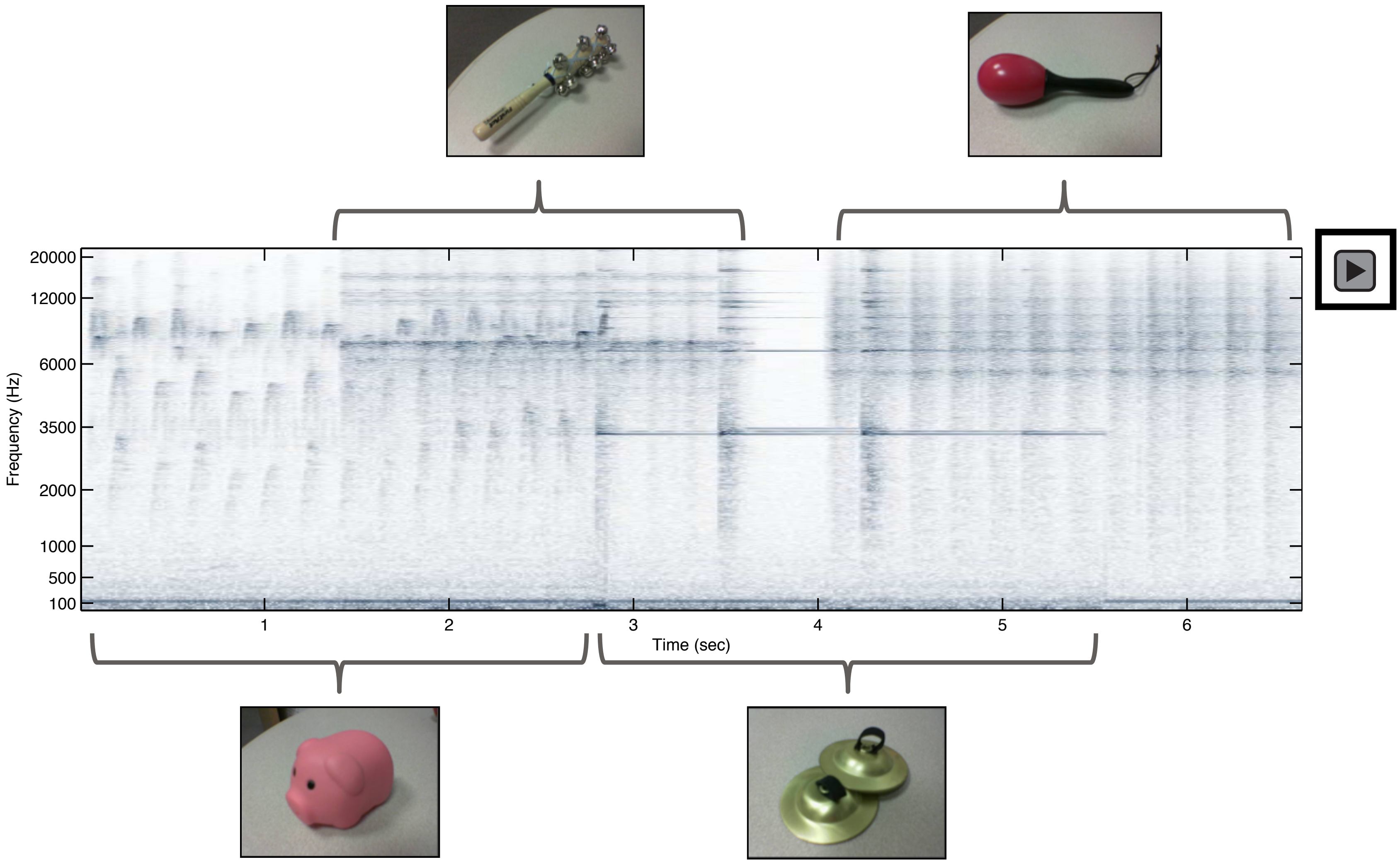
- What matters is the additivity of the model
  - Allows us to not care about mixing

$$\mathbf{H}_{x(t)+y(t)} \approx \mathbf{H}_{x(t)} + \mathbf{H}_{y(t)}$$

Sound classification/detection

- Machine learning approaches are a poor fit
  - Can’t use winner-takes-all classification
- The real question: How active is each class?
  - Not whether it exists

A challenging example

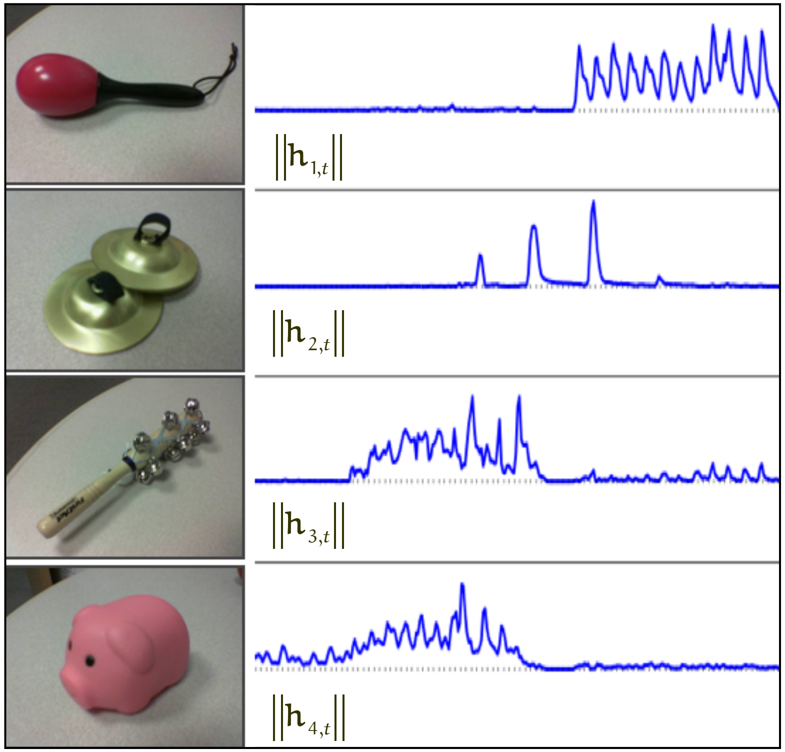


The non-negative treatment

- Decompose as:

$$\underline{\mathbf{x}_t} = \begin{bmatrix} \underline{\mathbf{W}_1 \quad \mathbf{W}_2 \quad \mathbf{W}_3 \quad \mathbf{W}_4} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{h}_{1,t} \\ \mathbf{h}_{2,t} \\ \mathbf{h}_{3,t} \\ \underline{\mathbf{h}_{4,t}} \end{bmatrix}$$

Known
Estimated



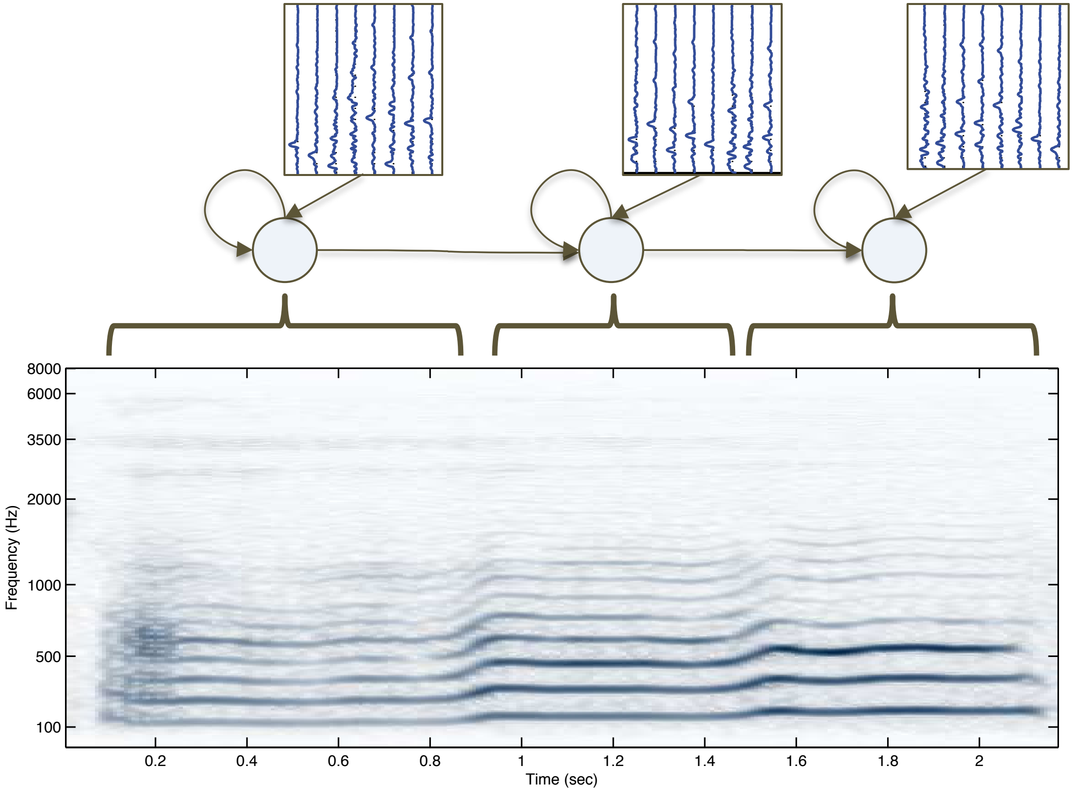
- Energies in $\mathbf{h}$ express presence of each sound

"Additive" sound recognition

- We can now find simultaneous sound classes

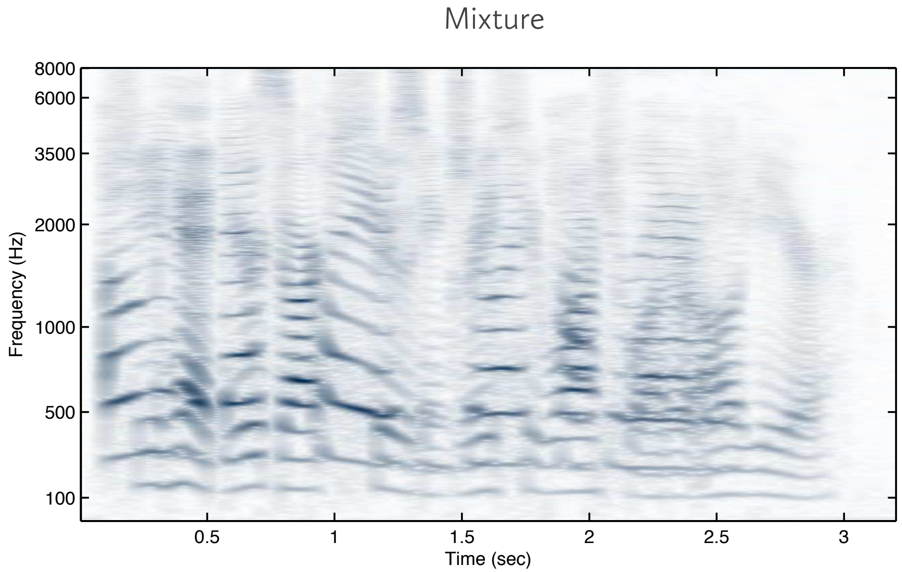
Adding the temporal dimension

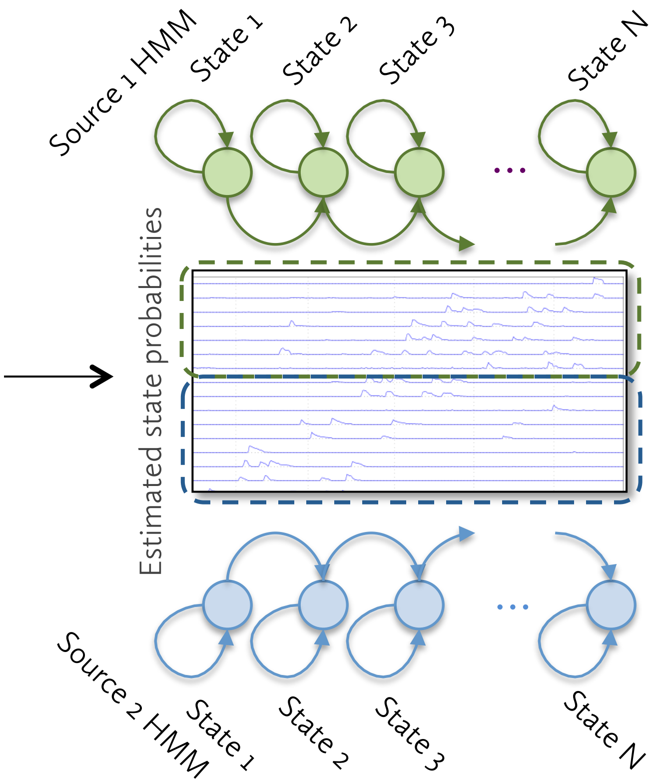
- To be serious we should use Markov models
  - The non-negative HMM:



Advantages over GMM HMMs

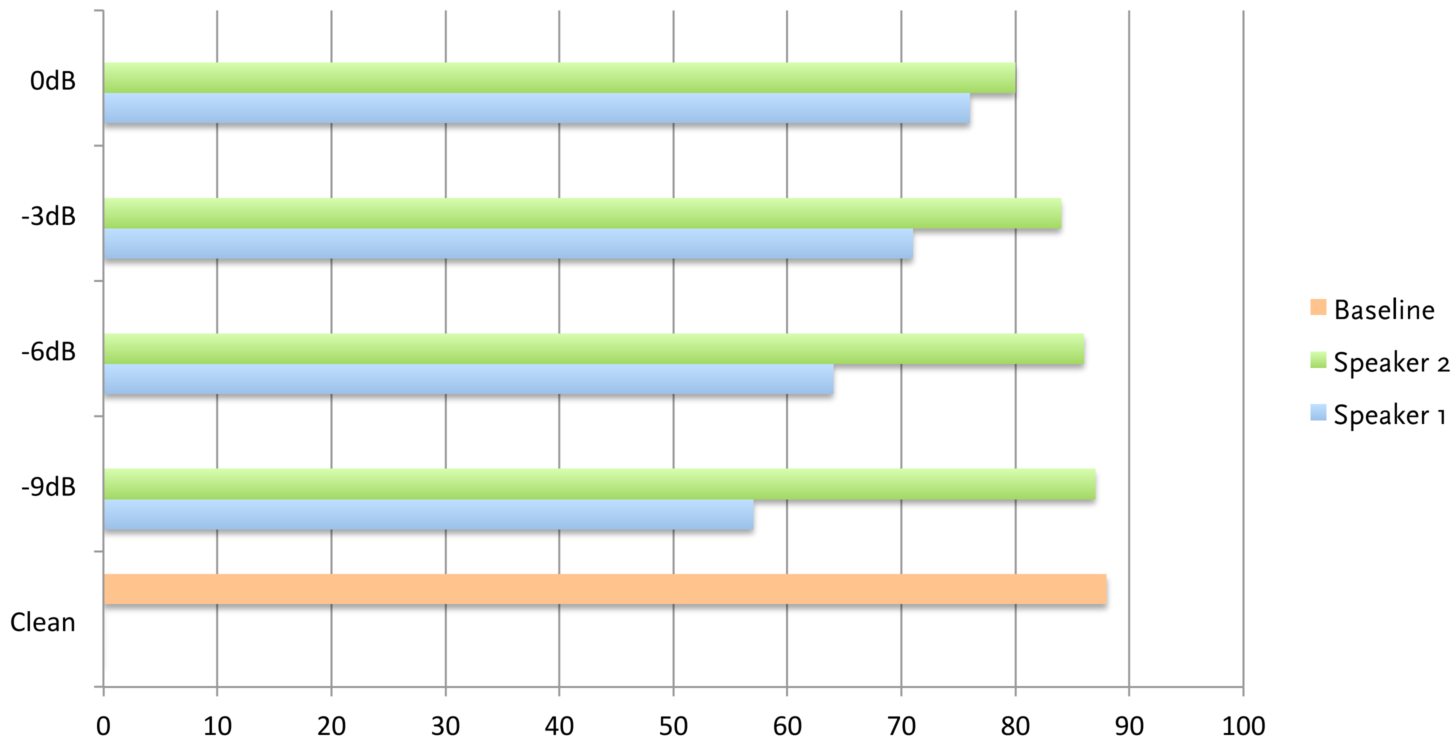
- No need for factorial models
  - Sum of models = model of sum of sounds





Speaker separation challenge

- WER doesn't drop drastically with maskers

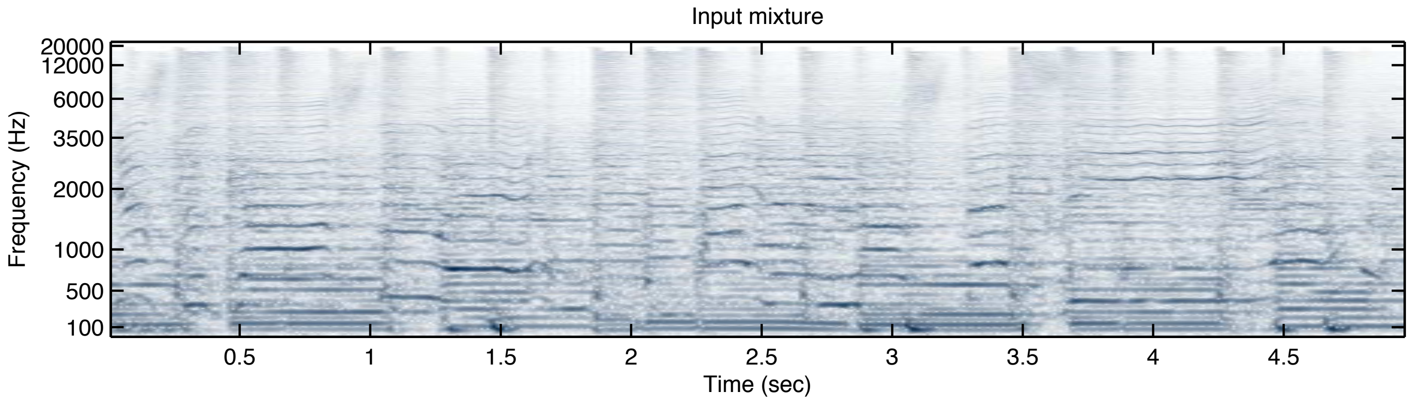


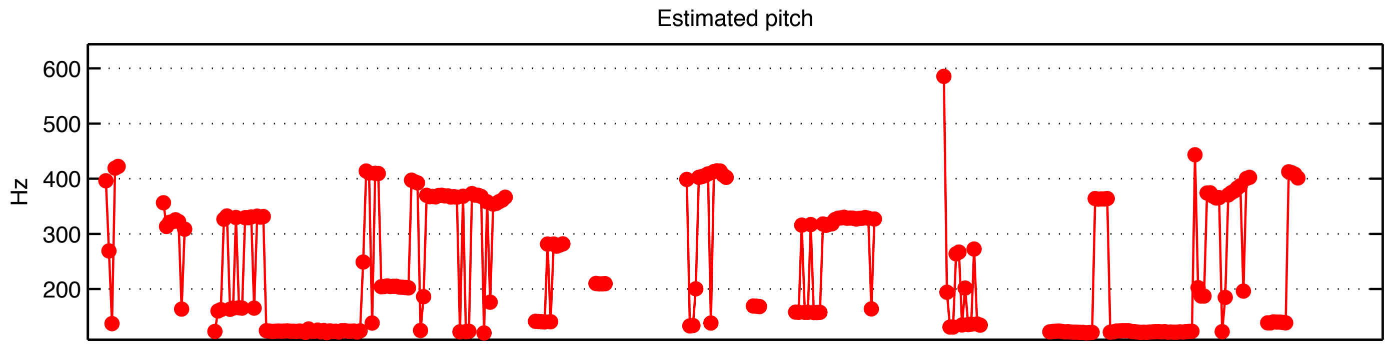
Parameter estimation in mixtures

- Estimate parameters of only one sound in mix
  - Usually hard due to mixing
- Associate components with parameter
  - Learn on tagged data
- Explain new input with model
  - Use component / parameter association

Example: Pitch tracking

- Works fine on clean sounds
  - Fails miserably on dense mixtures …





The non-negative pitch tracker

- Learn model from tagged data:

$$\mathbf{x}_t \to p_t$$

$$\mathbf{x}_t \approx \mathbf{W} \cdot \mathbf{h}_t$$

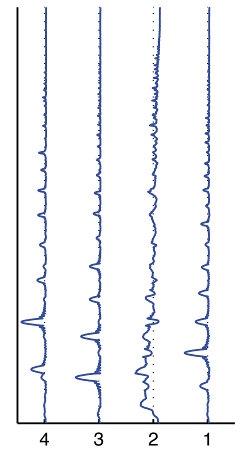
- Associate components & pitch:

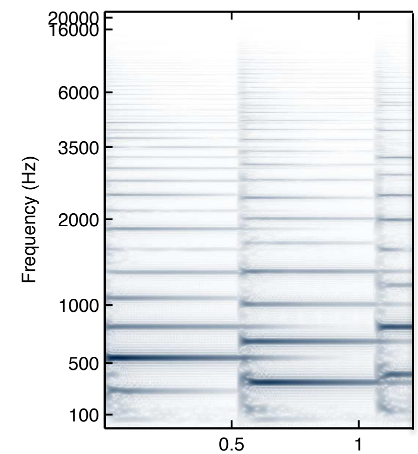
$$P\left(\mathbf{W}_i \to p_t\right) \propto \mathbf{h}_{i,t} \Big/ \sum_i \mathbf{h}_{i,t}$$

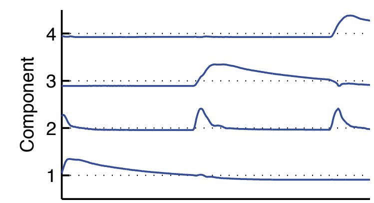
- Associate pitch to new inputs:

$$\mathbf{y}_t \approx \mathbf{W} \cdot \mathbf{h}_t$$

$$P\left(y_t \to p_t\right) \propto \sum_i h_{i,t} p_i \Big/ \sum_i h_{i,t}$$

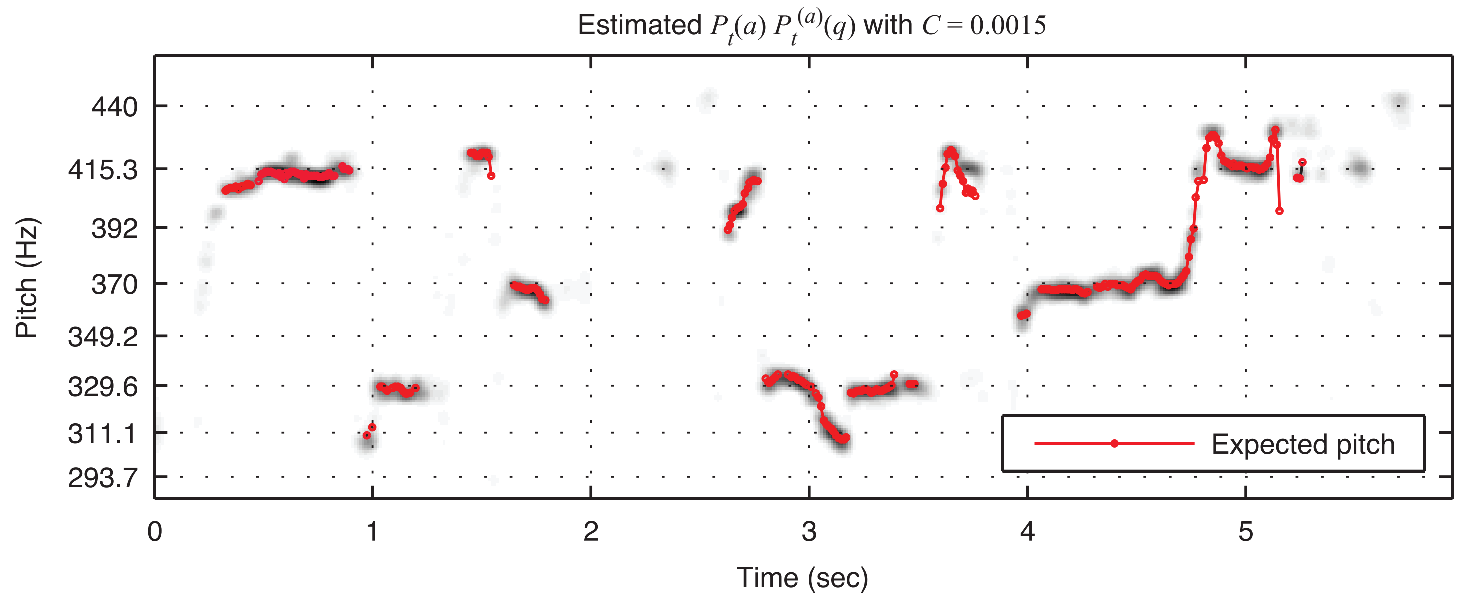






Result

- Sharp pitch probabilities on mixture



Training data

Mixture

- And also works for phonemes, sound class, loudness, and other parameters

And I could go on and on …

- Echo-cancellation, dereverberation, multi-modal processing, missing data, convolutive models, tensor versions, …

- Rich literature on non-negative models
  - Lots of WASPAA/ICASSP papers

So what is coming up next?

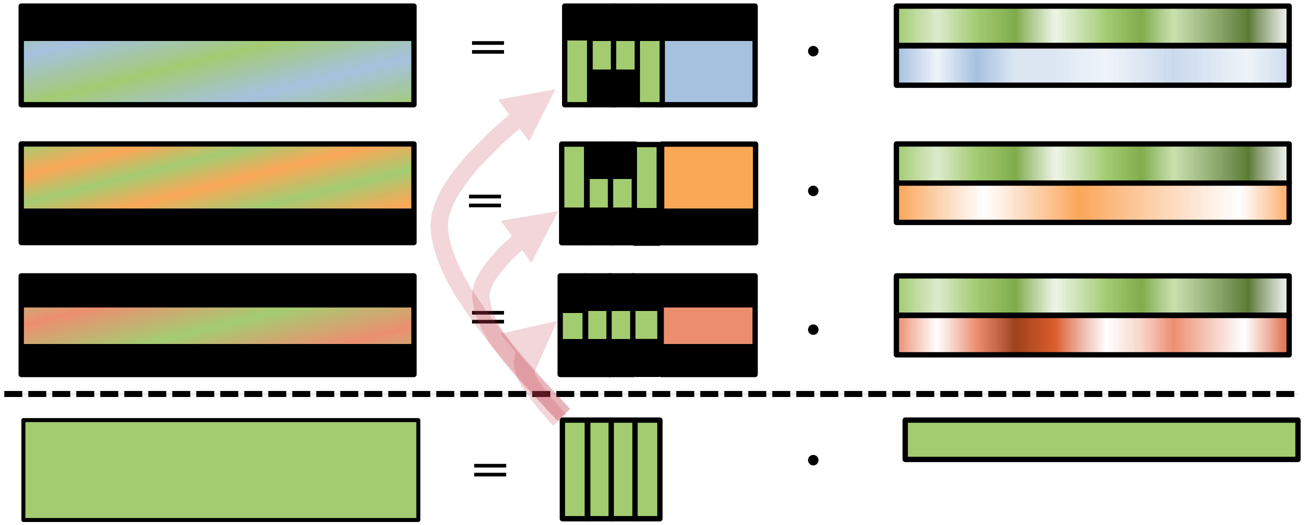
- Theory:
  - Problem definition, parameter estimation, convergence properties, variations and generative models, dynamical systems, …
- Practical directions:
  - Multi-channel data formulations
  - Alternative TF front-ends
  - Efficient formulations for big data

Rethinking the array

- We can re-conceptualize beamforming
  - Example case: Lots of cell phones in concert
    - All recordings will be bad and non-synced

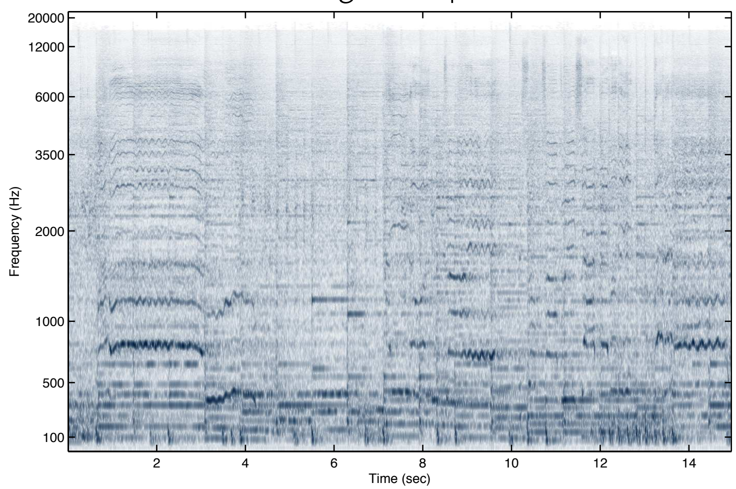
A non-negative take

- Joint component analysis
  - Common components are of interest
  - Non-common components are noise
  - Optional priors from reference recordings

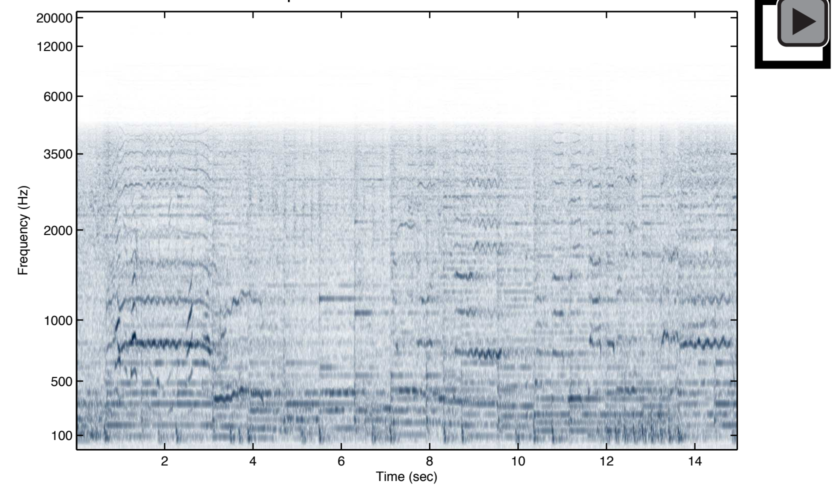


Example case

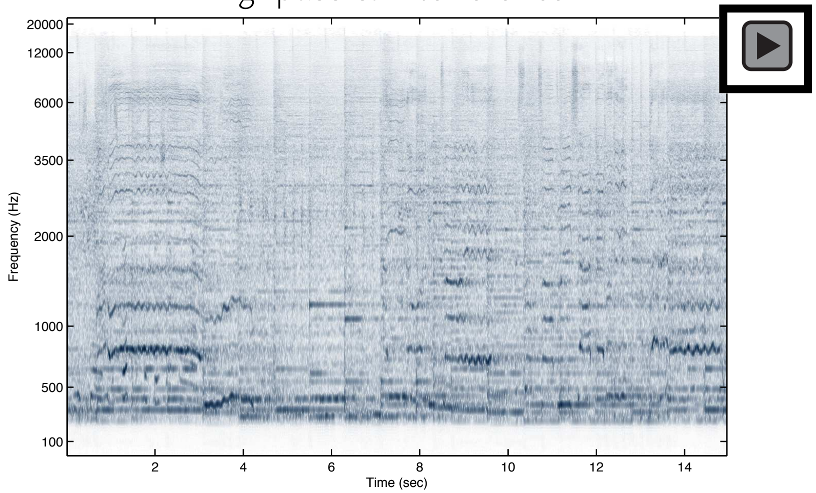
Original input



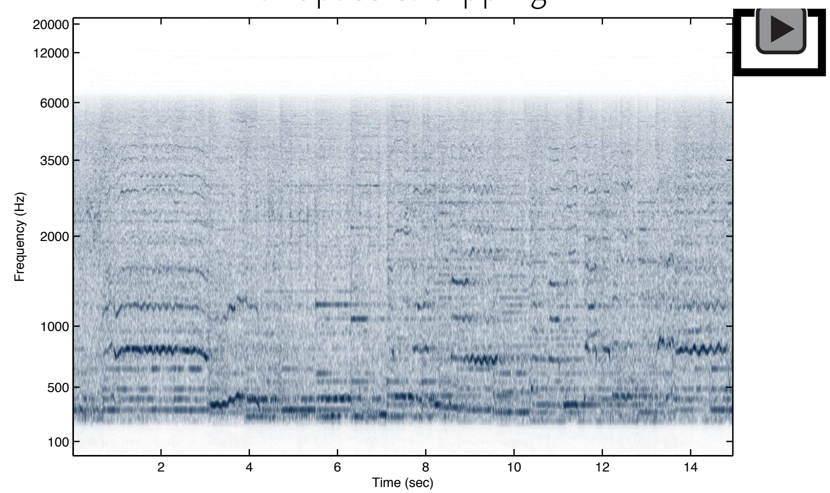
Lowpass & interference



Highpass & interference

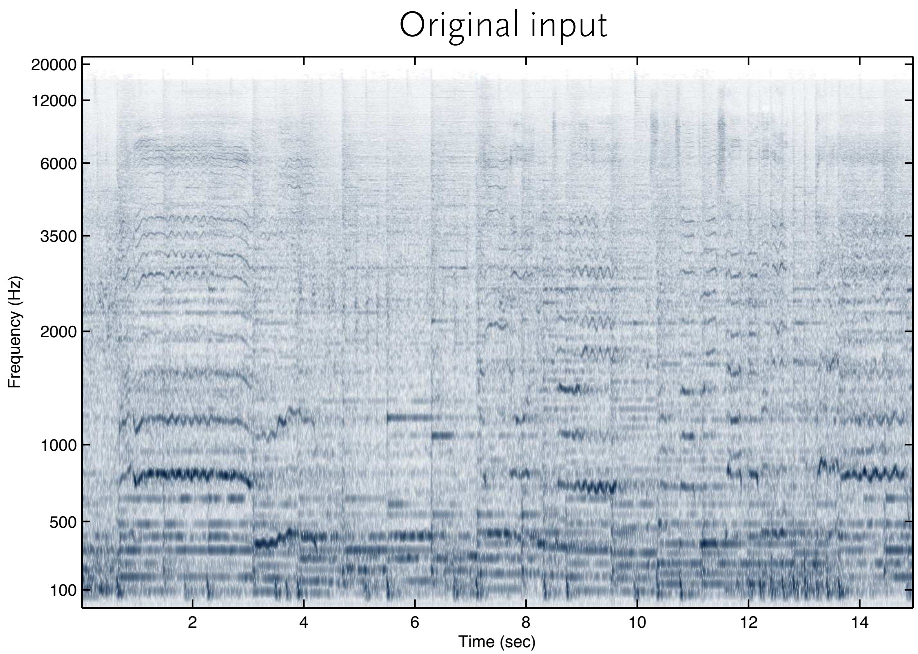


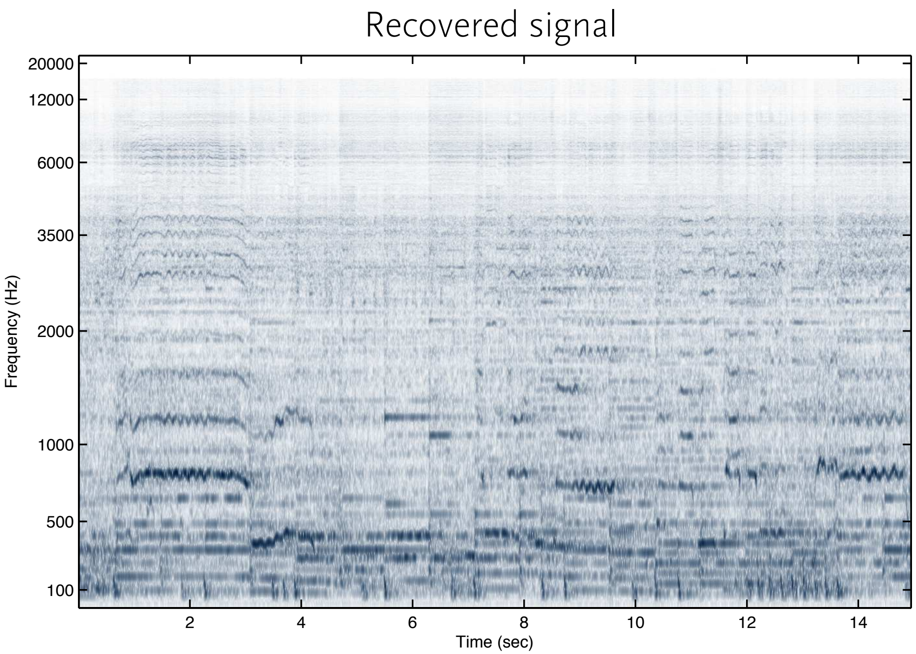
Bandpass & clipping



Recovered signal

- Recovery of full bandwidth
  - Suppression of uncommon elements
  - Not sensitive to non-linearities/synchronization



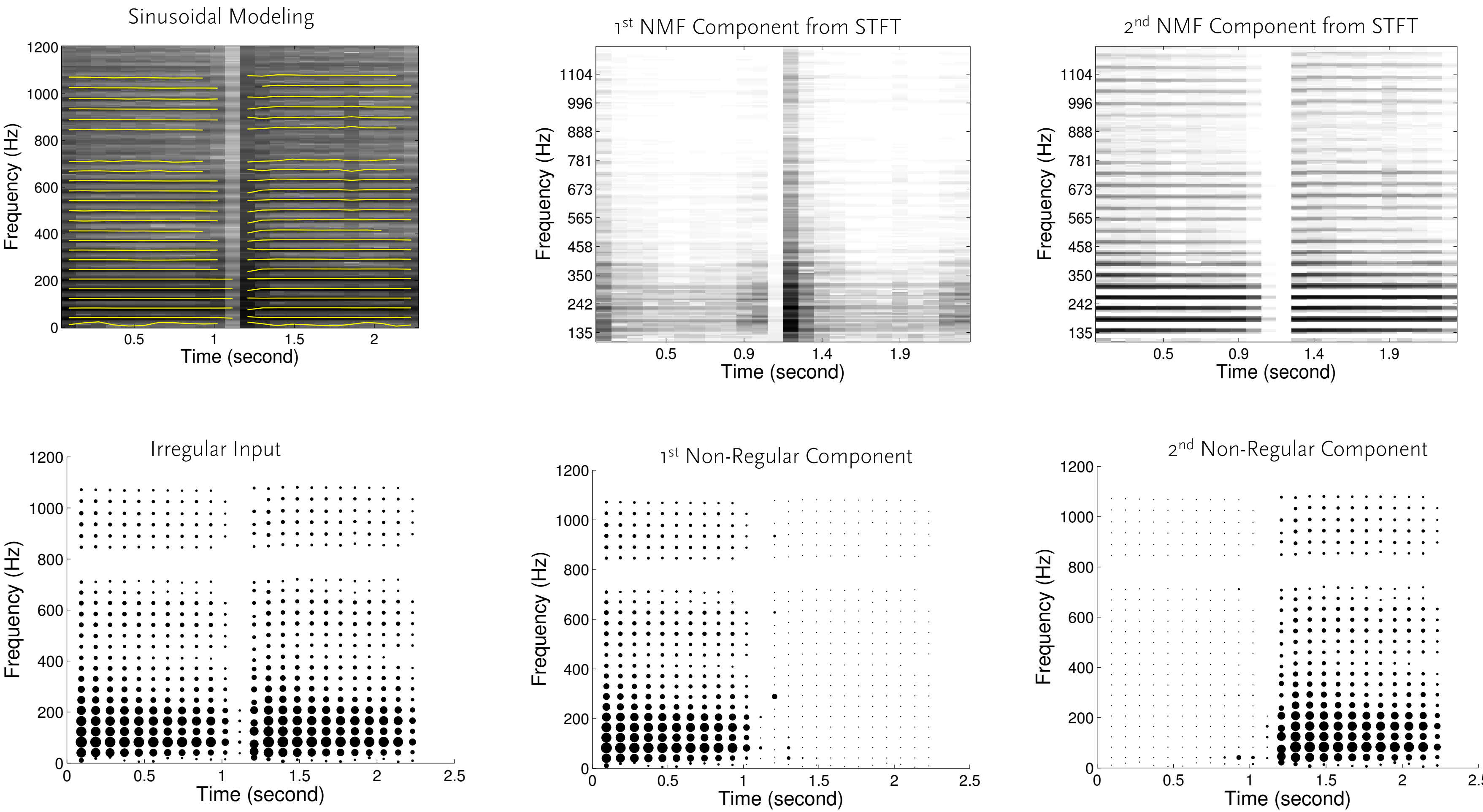


Alternative TF front ends

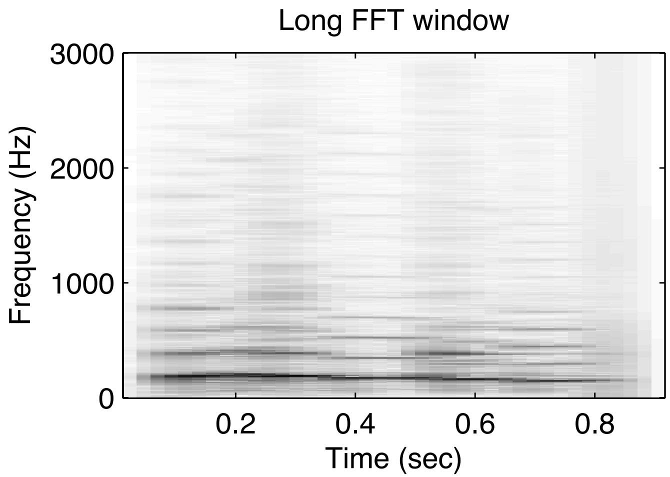
- The STFT has poor frequency resolution
  - We can do better with other transforms
    - Constant-Q, reassigned spectra, sinusoidal models, ...

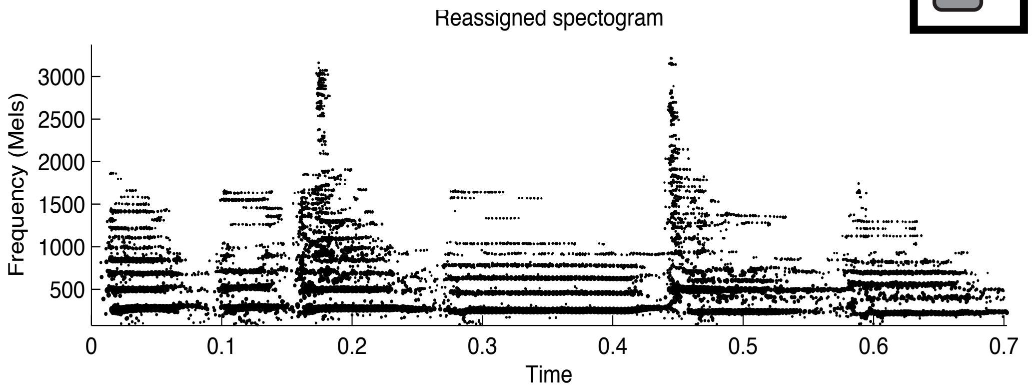
- But that data is not in a matrix format!!
  - Reformulate NMF as a function approximation
  - Allows us to use arbitrary TF representations

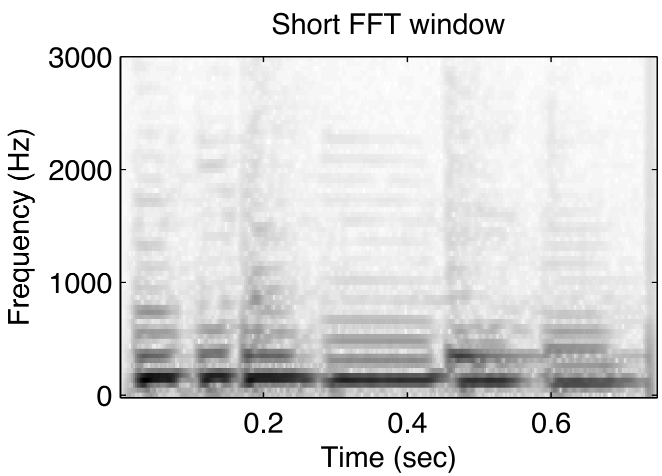
Sinusoidal model example

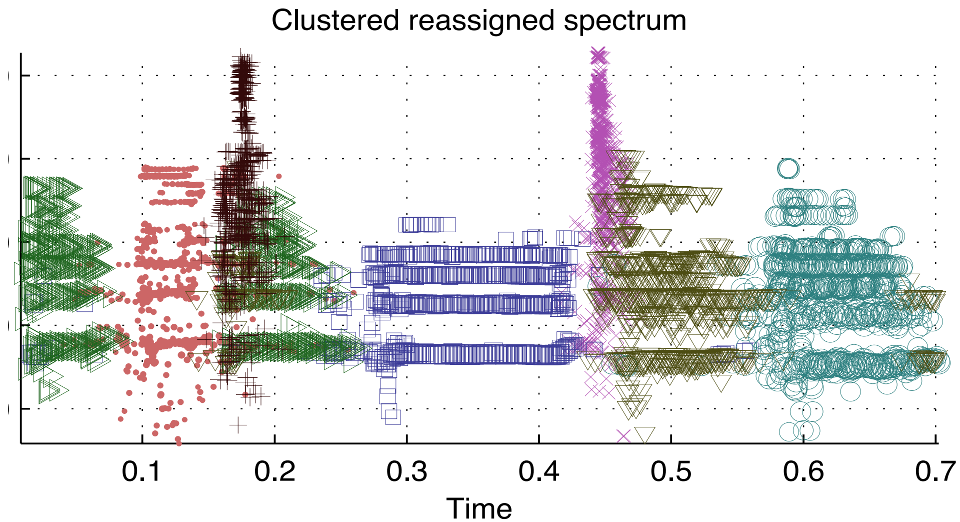


Reassigned spectra example



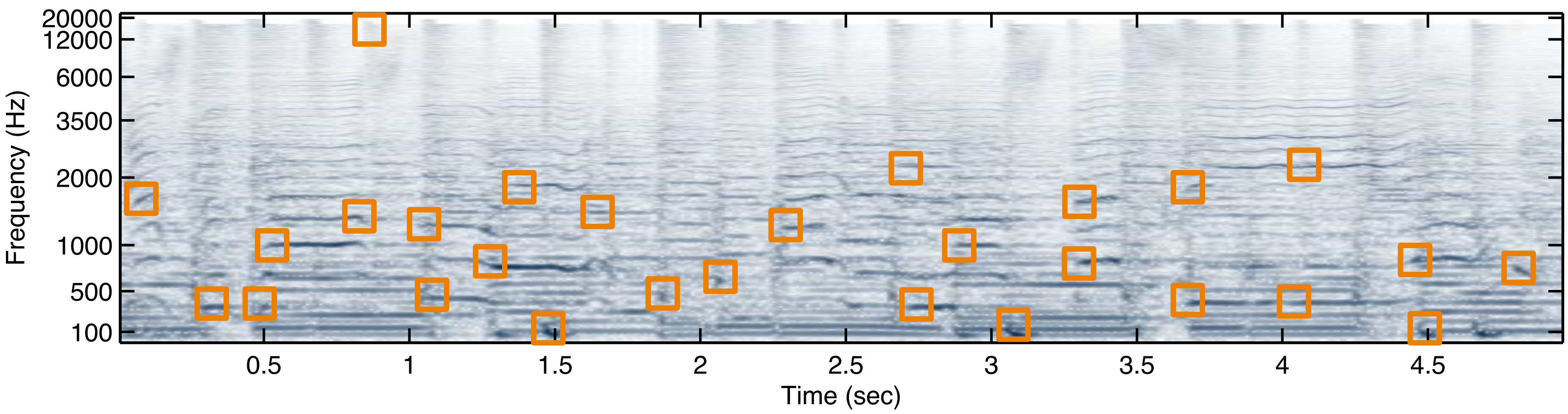






NMF for big data

- How do we analyze huge recordings?
  - Operate on landmark space instead



To conclude

- The wild west is in non-negative models
  - Can they be the new Gaussian?
- A more perceptual take on analysis
  - Still on unclear math ground though
- Thanks!
  - And many thanks to Nick Bryan, Minje Kim, Gautham Mysore, Madhu Shashanka