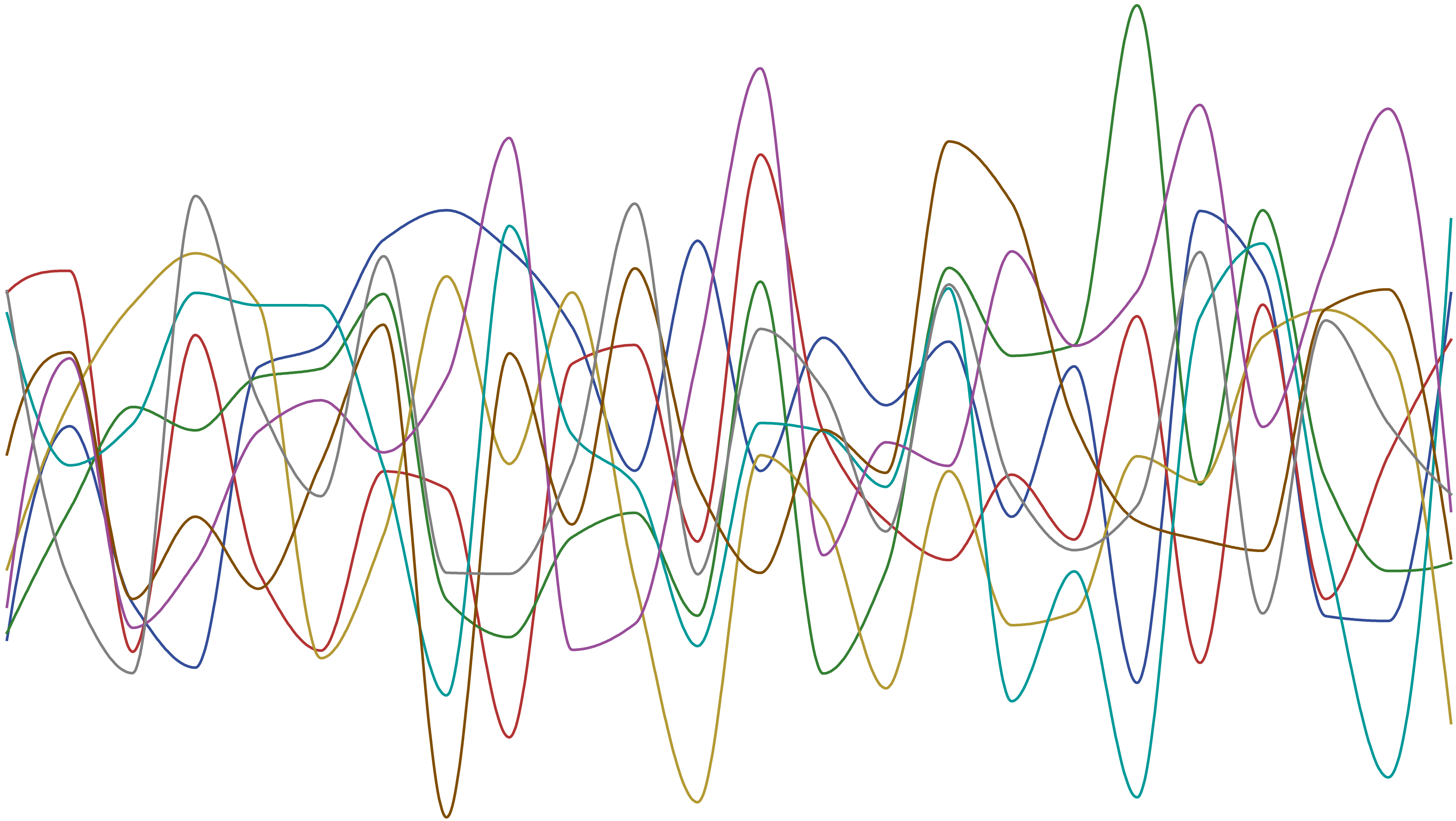
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



Fun with audio mixtures

(or why DSP is actually useful)

Paris Smaragdis

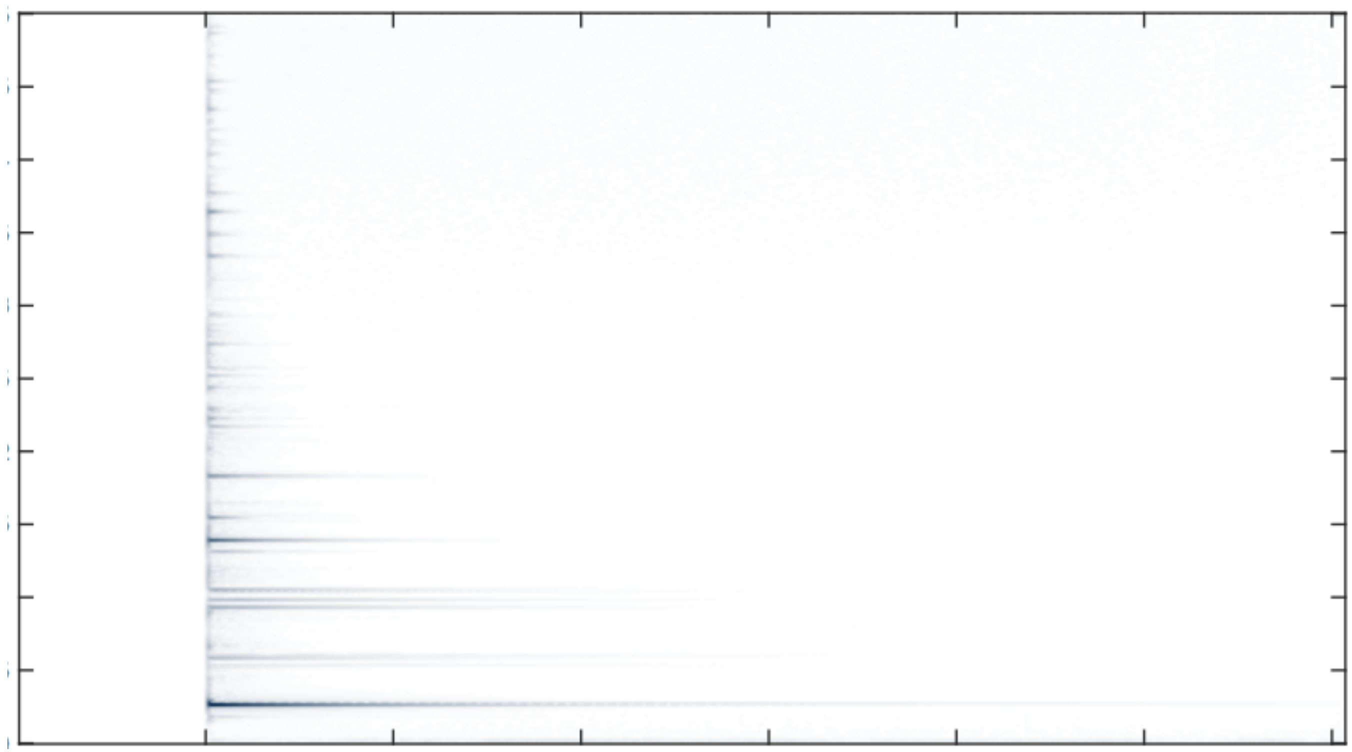
paris@illinois.edu

What fun?

- Audio mixture problems
  - Scene analysis
  - Separation
  - Denoising
  - Manipulation
  - Classification and detection
  - …

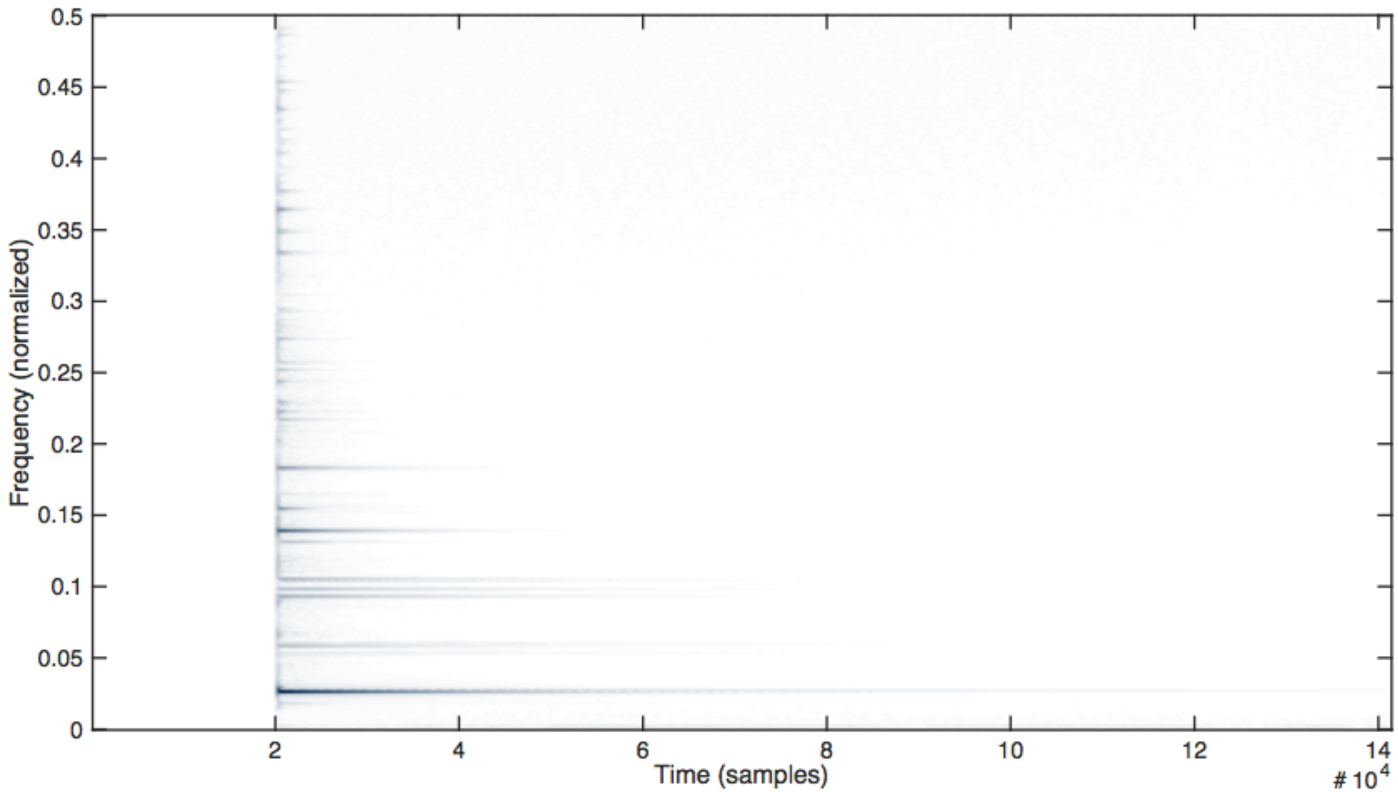
A starter

- What's this?



A spectrogram

- Representing energy in time/frequency



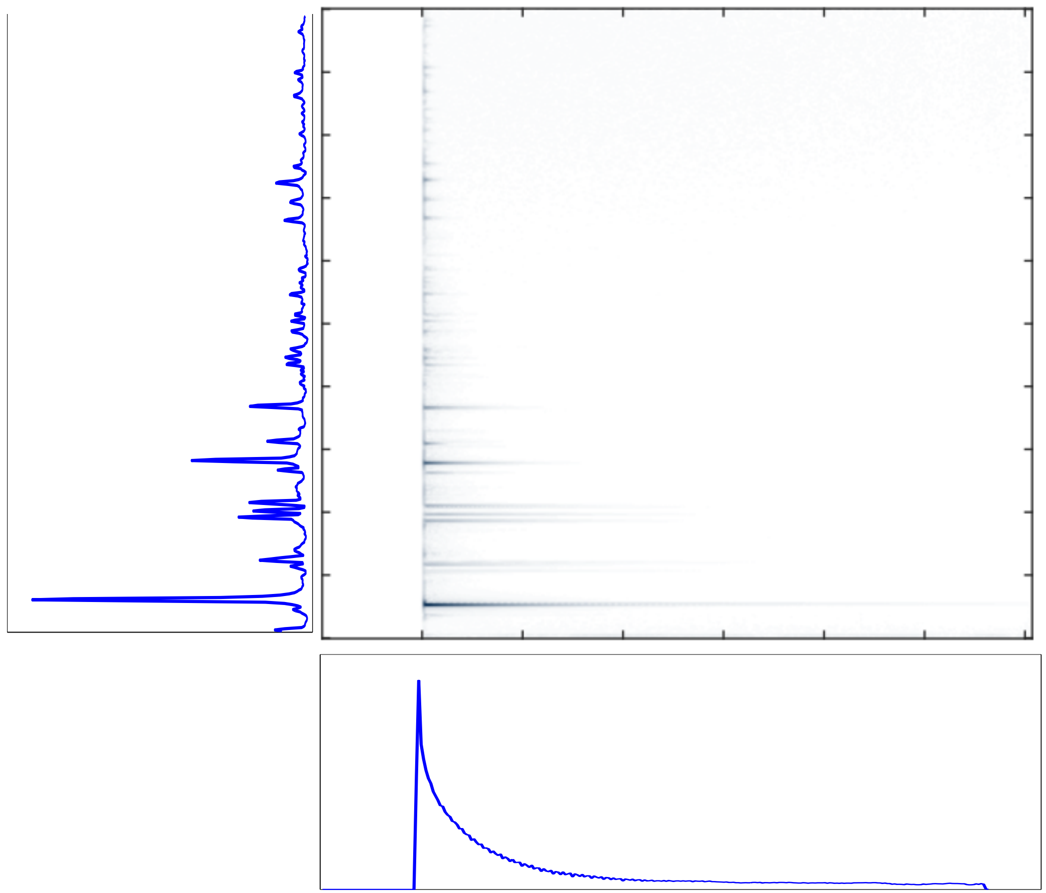
Factorizing a spectrogram

- We can approximate the 2D magnitude spectrogram as a product of two 1D functions
  - A "frequency function" and a "time function"

$$\underbrace{S(\omega,\tau)}_{\textit{Spectrogram}} = \underbrace{F(\omega)}_{\textit{Frequency function}}\underbrace{A(\tau)}_{\textit{Time function}}$$

Estimating the factors

- We integrate over each dimension



Nothing new here

- Frequency factor == Power spectrum
  - Energy distribution across frequency

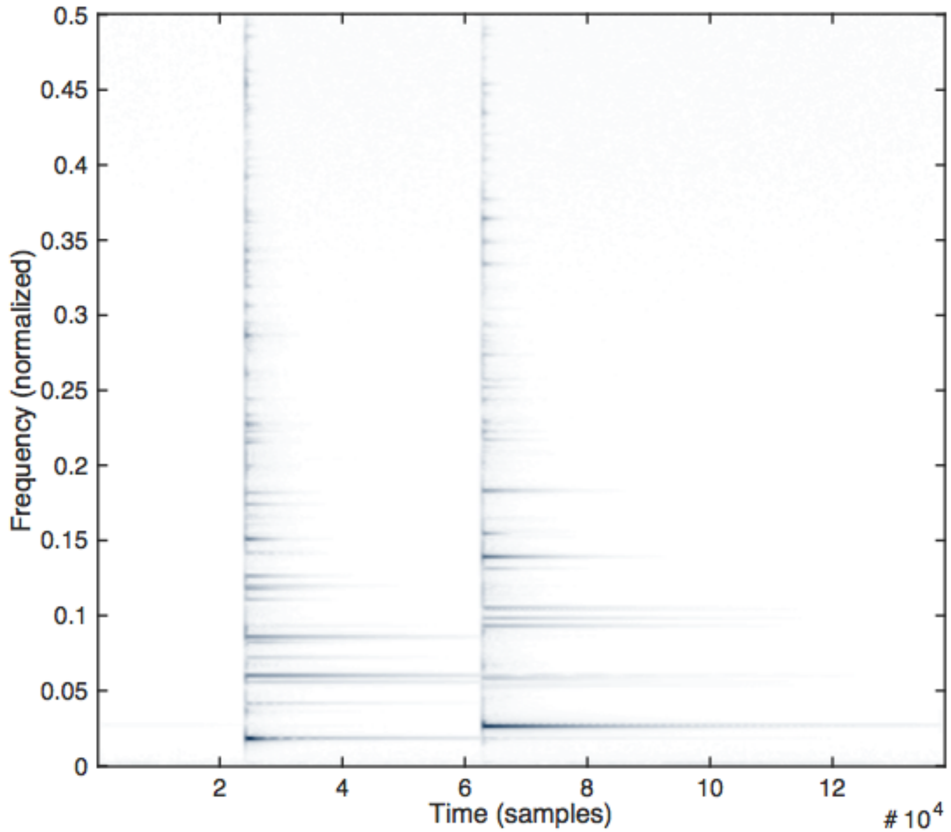
$$F(\omega) = \int S(\omega, \tau) d\tau$$

- Time factor == Time envelope
  - Energy distribution across time

$$A(\tau) = \int S(\omega, \tau) d\omega$$

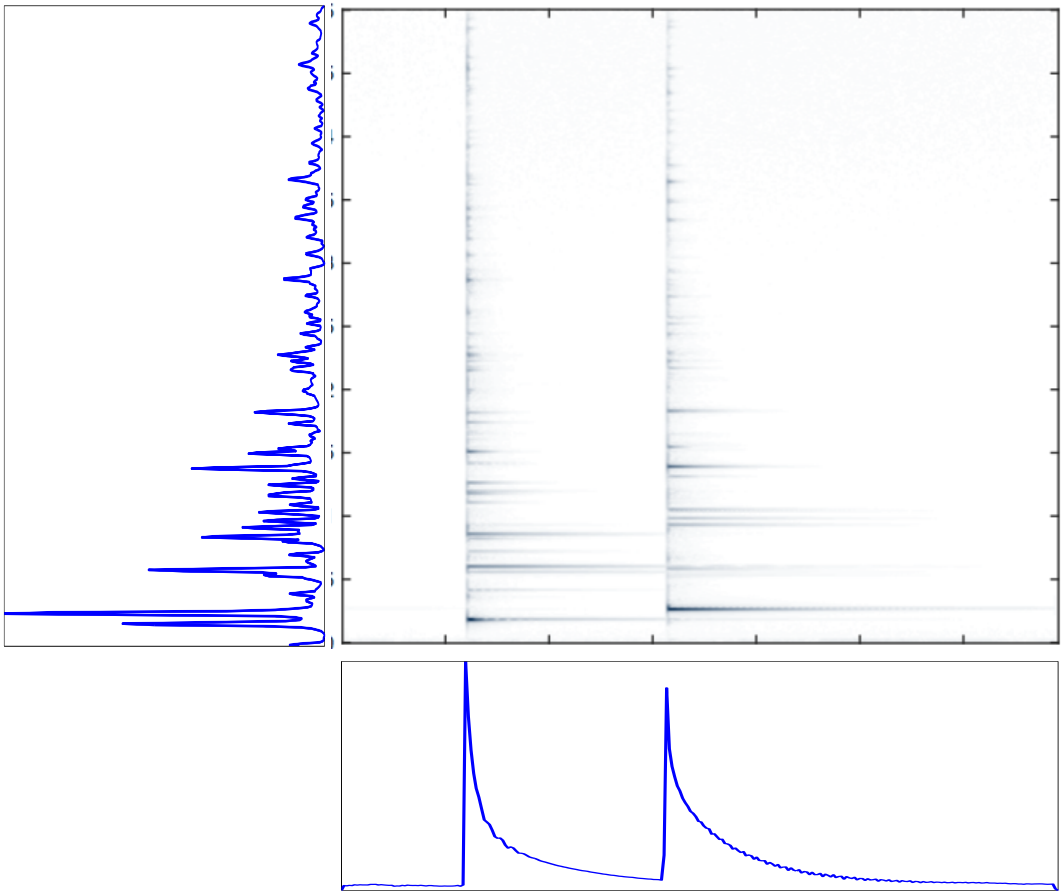
Another example

- What now?



Not as useful this time

- The factors are "mixed"



What if?

- Could we get two sets of factors?
  - One for each sound

- Initial factorization model:

$$S(\omega,\tau) = F(\omega)A(\tau)$$

- A "multiple factors" model:

$$S(\omega,\tau) = F_1(\omega)A_1(\tau) + F_2(\omega)A_2(\tau)$$

Learning this model

- Simple two-stage process
  - 1. "Assign blame"

$$\gamma_i(\omega,\tau) = \frac{F_i(\omega)A_i(\tau)}{F_1(\omega)A_1(\tau)+F_2(\omega)A_2(\tau)}$$

  - 2. Re-estimate

$$F_i(\omega) = \int \gamma_i(\omega,\tau)S(\omega,\tau)d\tau$$

$$A_i(\tau) = \int \gamma_i(\omega,\tau)S(\omega,\tau)d\omega$$

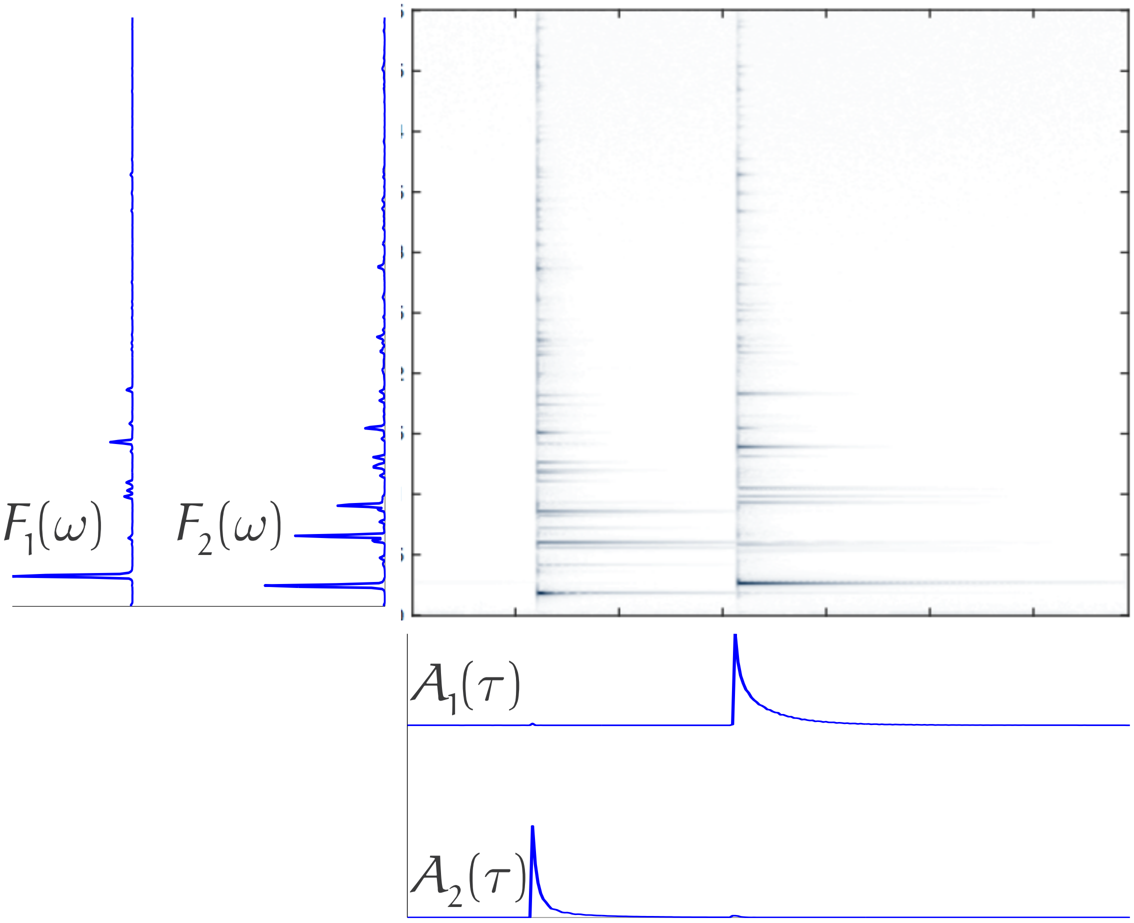
  - Repeat …

MATLAB code:

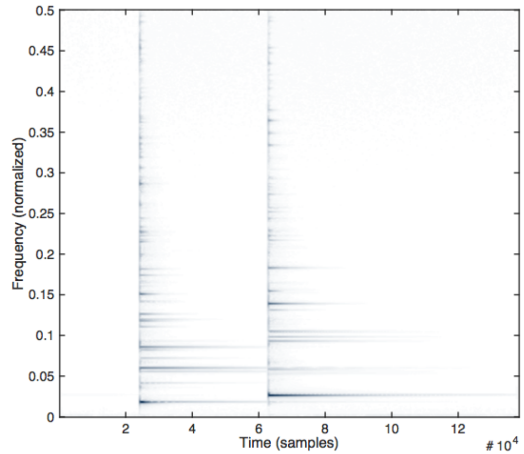
```
for i = 1:100
  g = s ./ (f *a );
  f = f .* (s *g');
  a = a .* (f’*g );
end
```

What does this do?

- More useful representation!



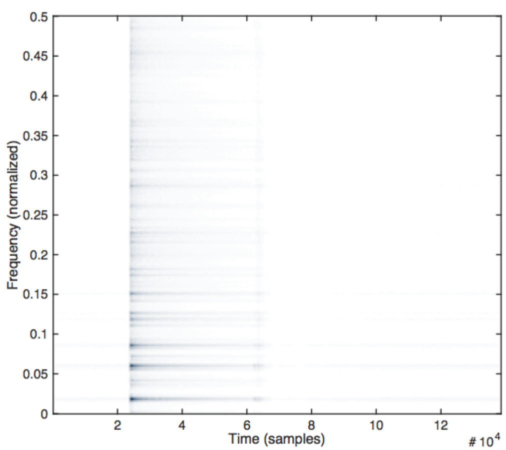
Factors learn sources



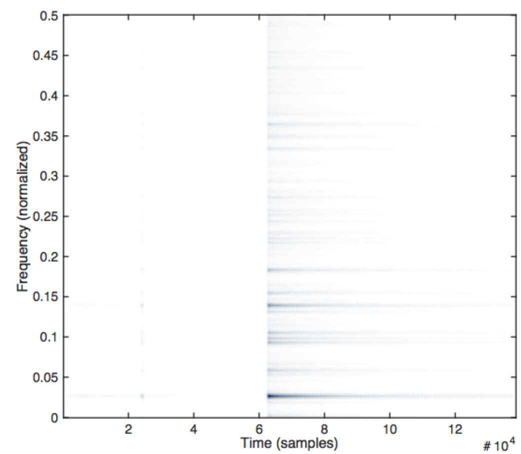
Two bells

$F_1(\omega)A_1(\tau)$

$F_2(\omega)A_2(\tau)$



One bell



The other bell

Going back to time

- Original input had amplitude *and* phase

$$S(\omega,\tau) = \left|\mathrm{STFT}\; x(t)\right|$$

$$\Phi(\omega,\tau) = \angle\, \mathrm{STFT}\; x(t)$$

- To get each source assume:

$$\left|\mathrm{STFT}\; y_i(t)\right| = F_i(\omega) A_i(\tau)$$

$$\angle\, \mathrm{STFT}\; y_i(t) = \Phi(\omega,\tau)$$

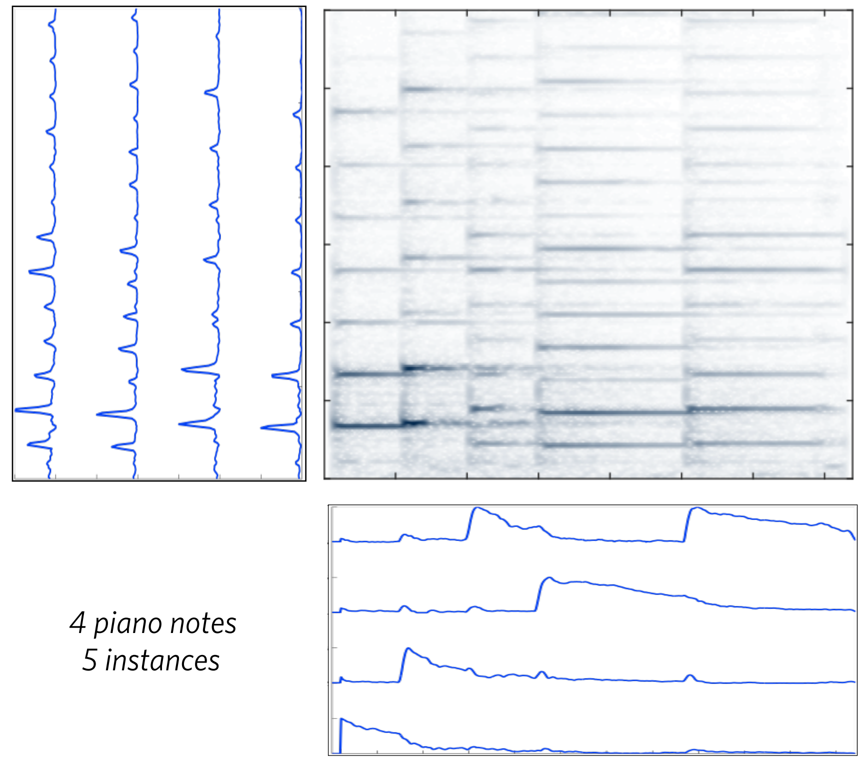
- So now we can extract each bell:  

Let's try some music

- Use more factors
  - Four in this case

- What do they correspond to?

A bit of Bach



4 piano notes
5 instances

General algorithm

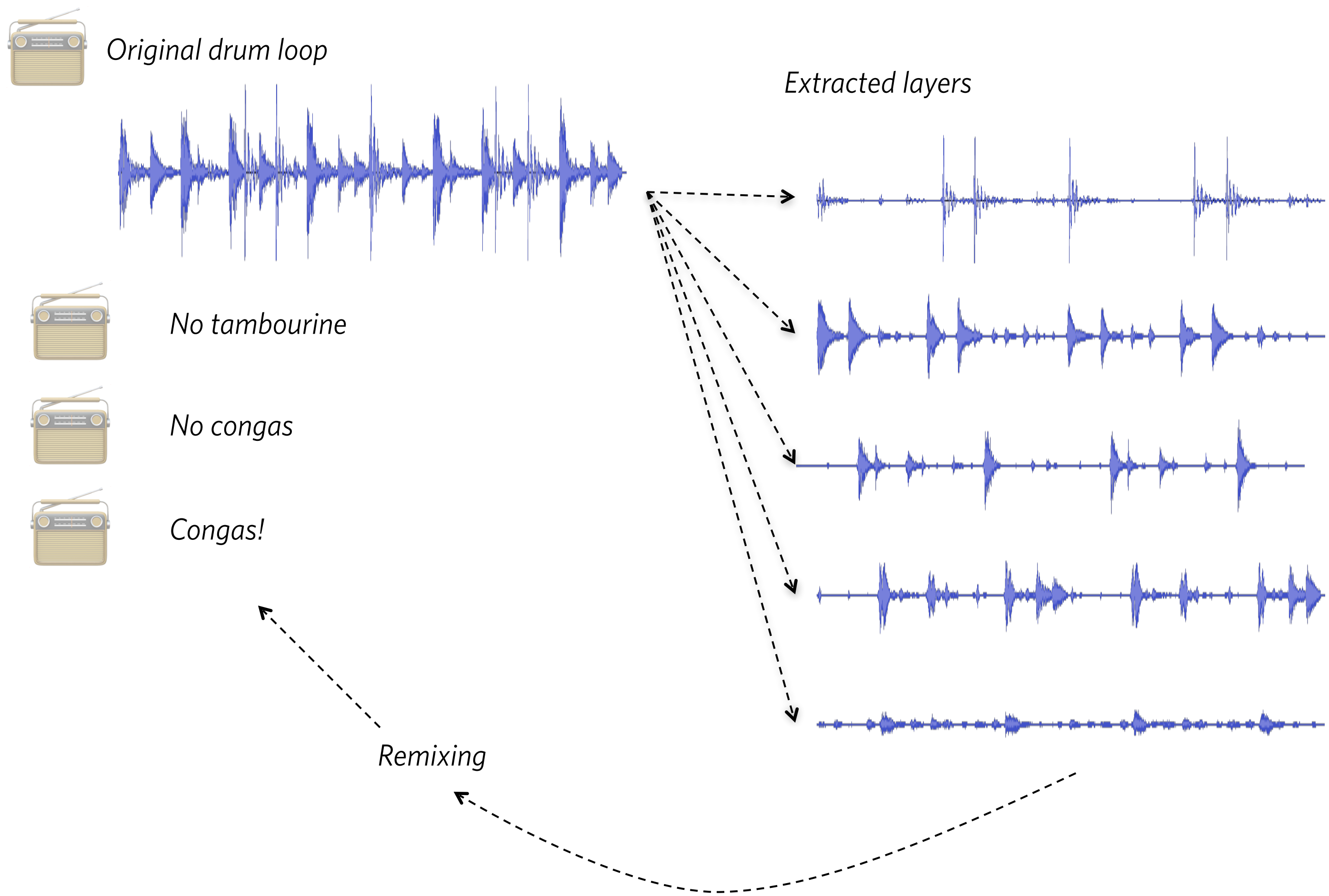
- If you like linear algebra:
  - Non-negative Matrix Factorization

$$\mathbf{F} = \mathbf{W} \cdot \mathbf{H}$$

- If you like probability:
  - Probabilistic Latent Component Analysis

$$P(\omega, \tau) = \sum_z P(z) P(\omega \mid z) P(\tau \mid z)$$

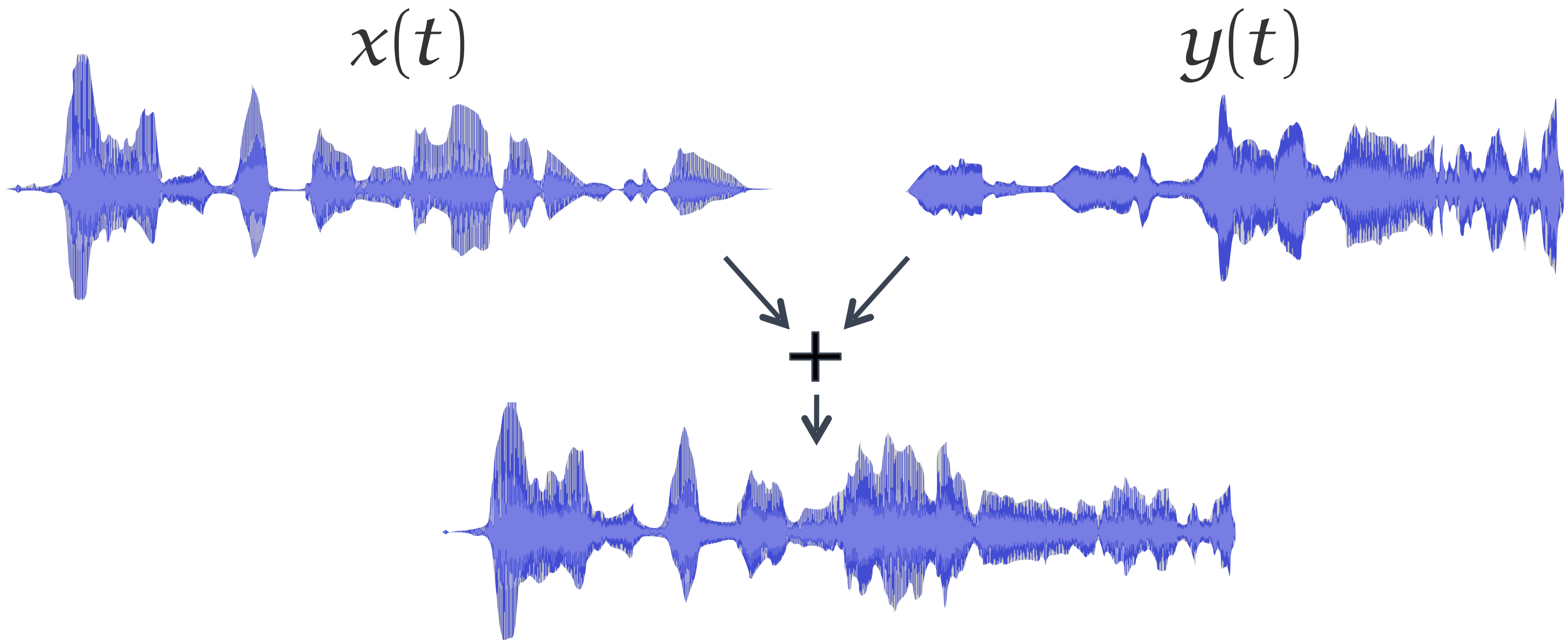
Audio remixing



But …

- We can only deal with simple sounds
  - Why?
- Let’s look at real-world source separation

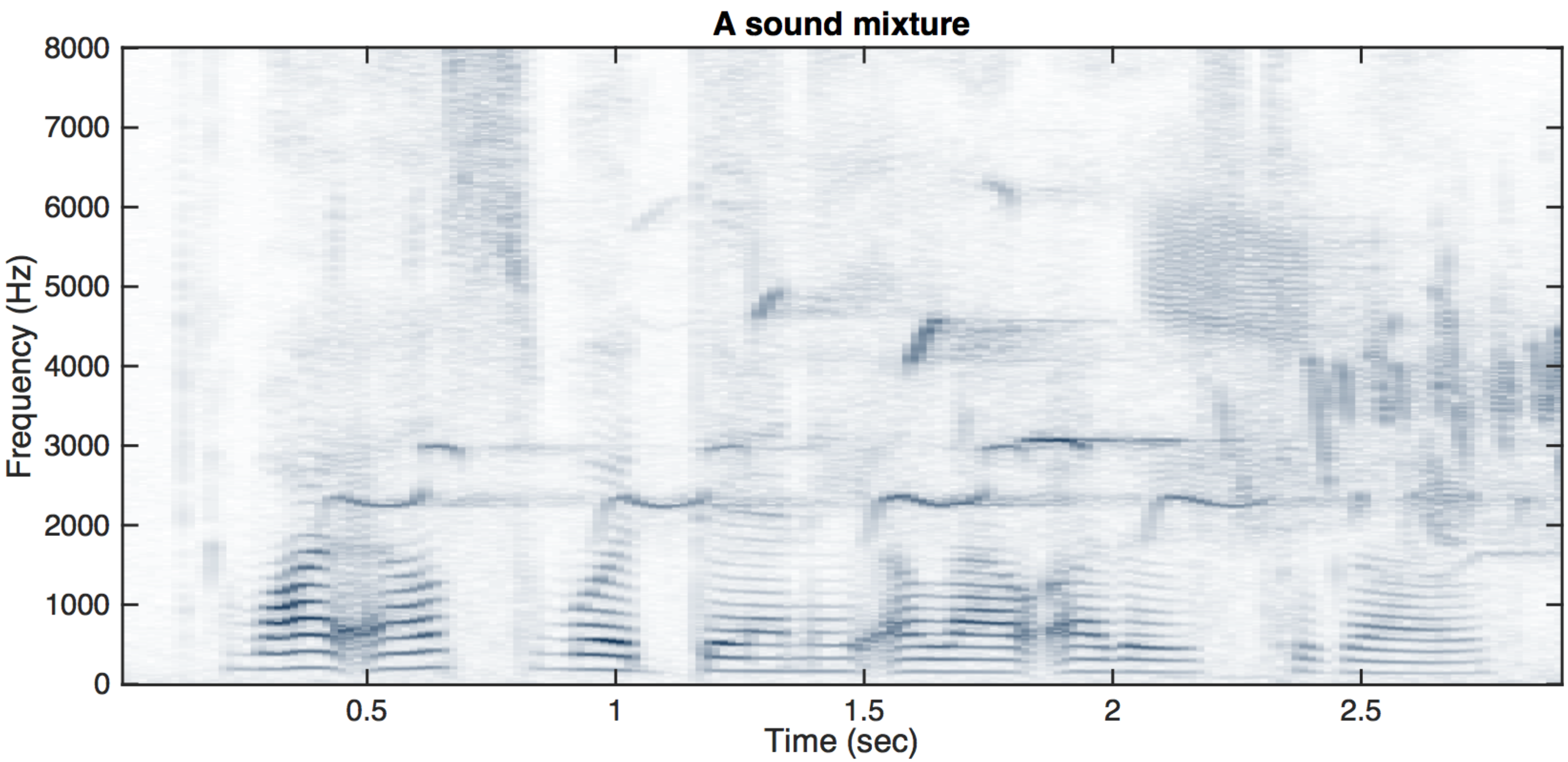
Defining the problem



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

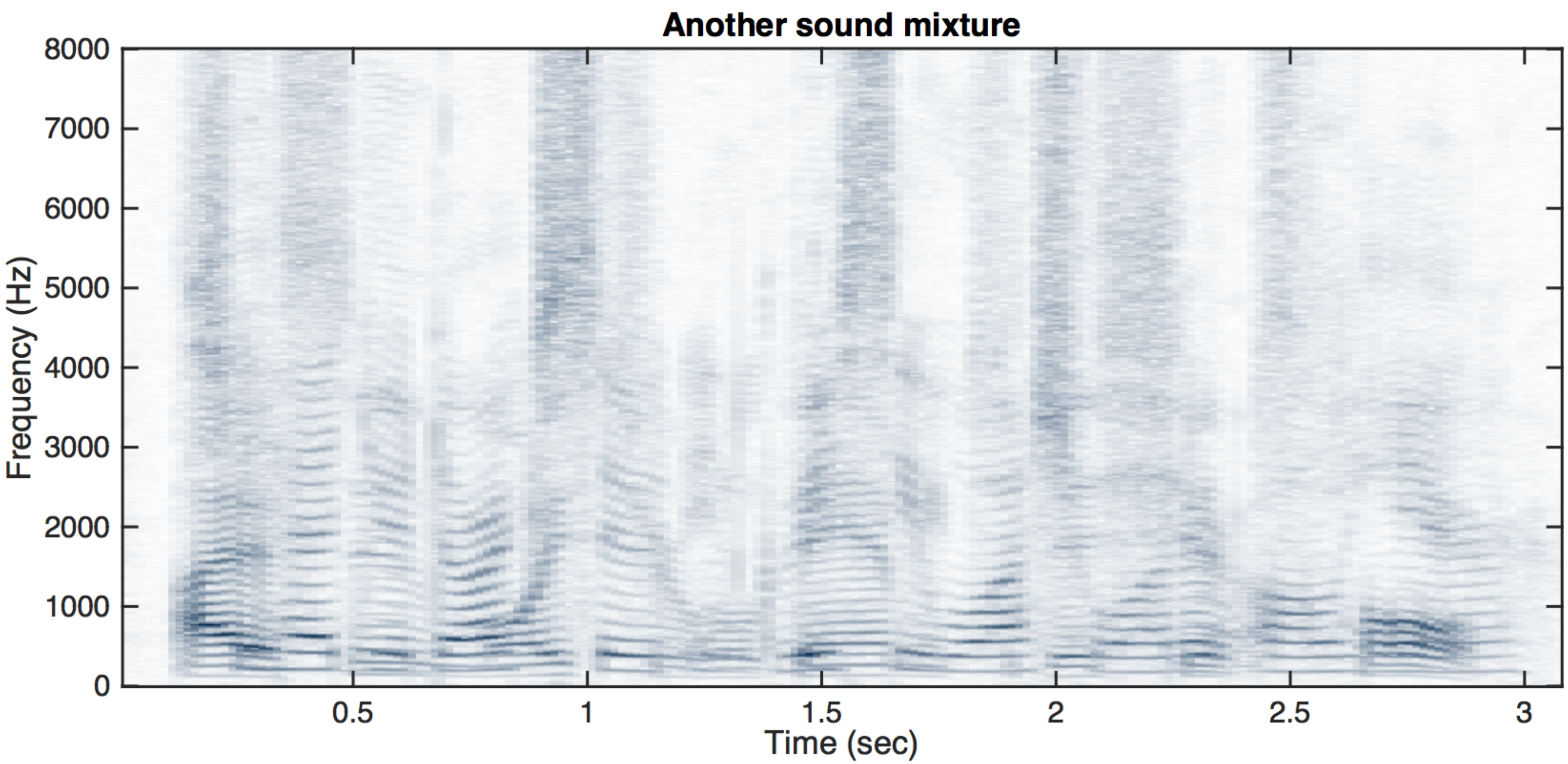
- An ill-defined problem!
  - “Single-channel source separation”

The name of the game



- Finding signal priors to perform separation
  - School a: Perceptually-minded approaches
  - School b: Statistical approaches

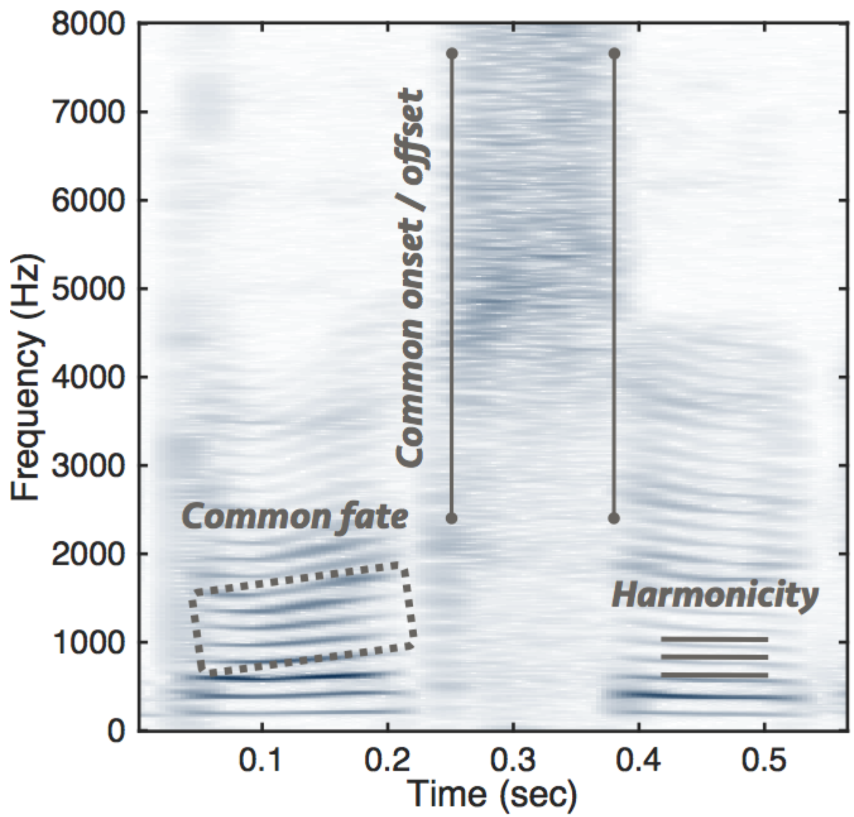
The name of the game



- Finding signal priors to perform separation
  - School a: Perceptually-minded approaches
  - School b: Statistical approaches

Perceptual approaches

- “Computational Auditory Scene Analysis”
  - Driven by psychoacoustic experiments



Brown

Ellis

Wang

Some (general) statistical approaches

- Approaches with general source assumptions
  - Lee and Jang 
    - ICA dictionaries of time waveforms
  - Reyes, Jojic and Ellis 
    - Graphical model on TF distributions
  - Lagrange, et al. 
    - Normalized cuts
  - Bach and Jordan 
    - Spectral clustering for perceptual grouping
- Things aren't great …

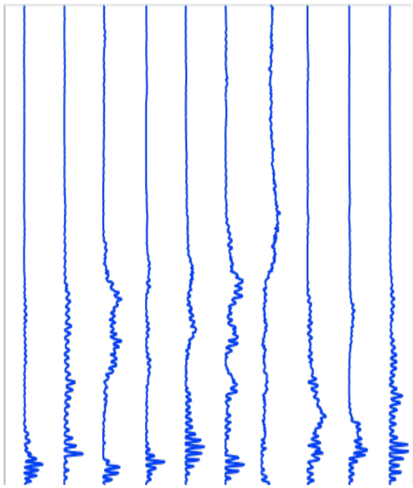
Forgetting unsupervised methods

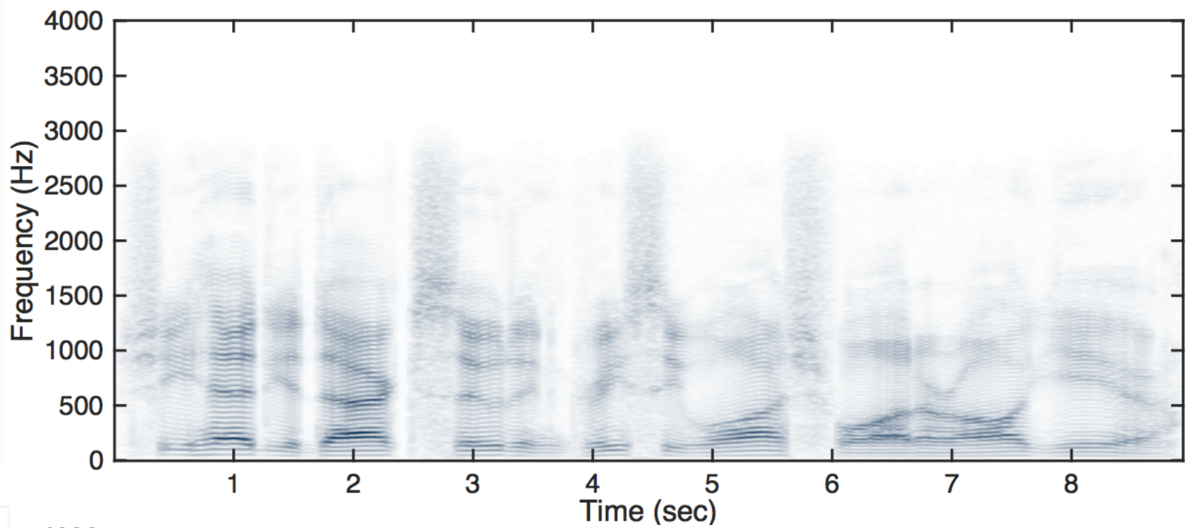
- It is hard to define source structure
  - We should learn it instead

- Supervised source separation
  - Use training data to hints what you want

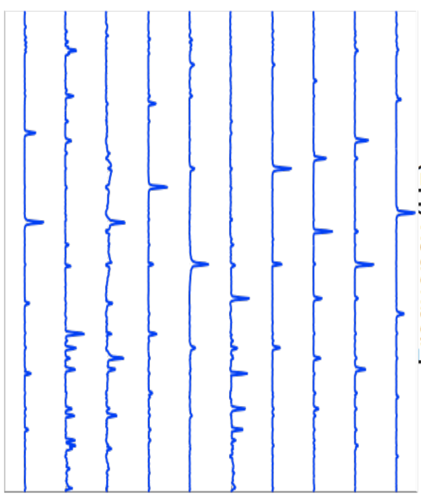
Learning factor dictionaries

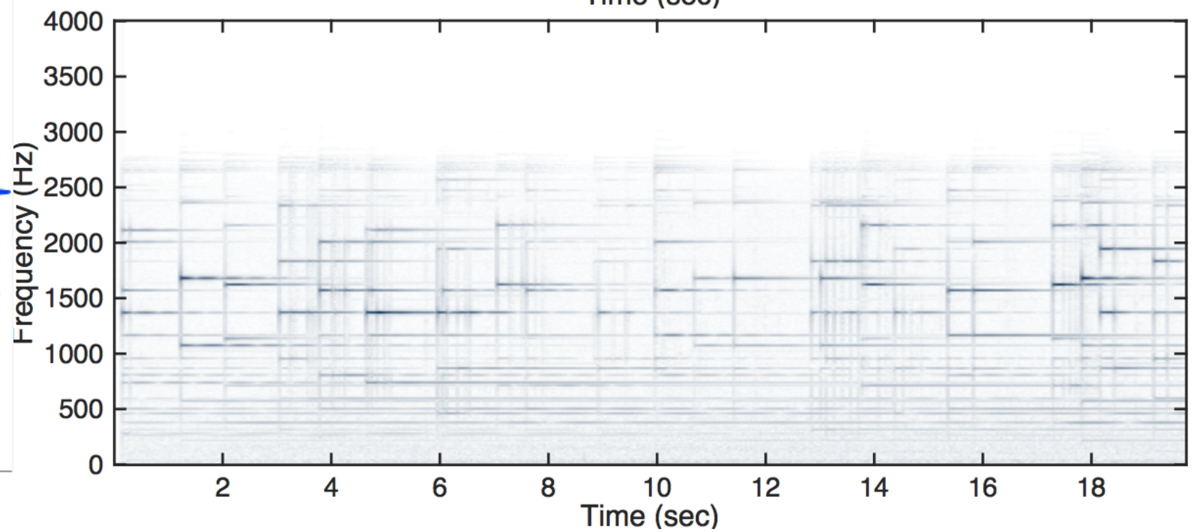
Speech example





Chime example





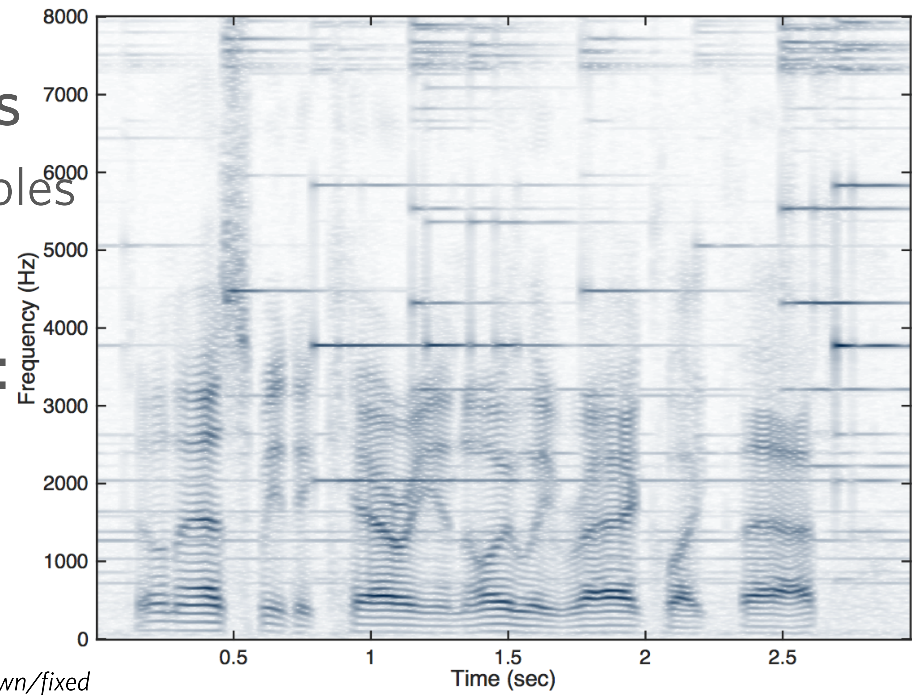
Mixtures of sounds

- Assumption: Sound mixtures are composed by frequency factors of the included sounds
  - Which we can learn from examples
- To separate estimate their proportions from the mixture:

$$\underline{S(\omega,\tau)} = \sum_i \underline{F(\omega,i)}\,\underline{A(\tau,i)}$$

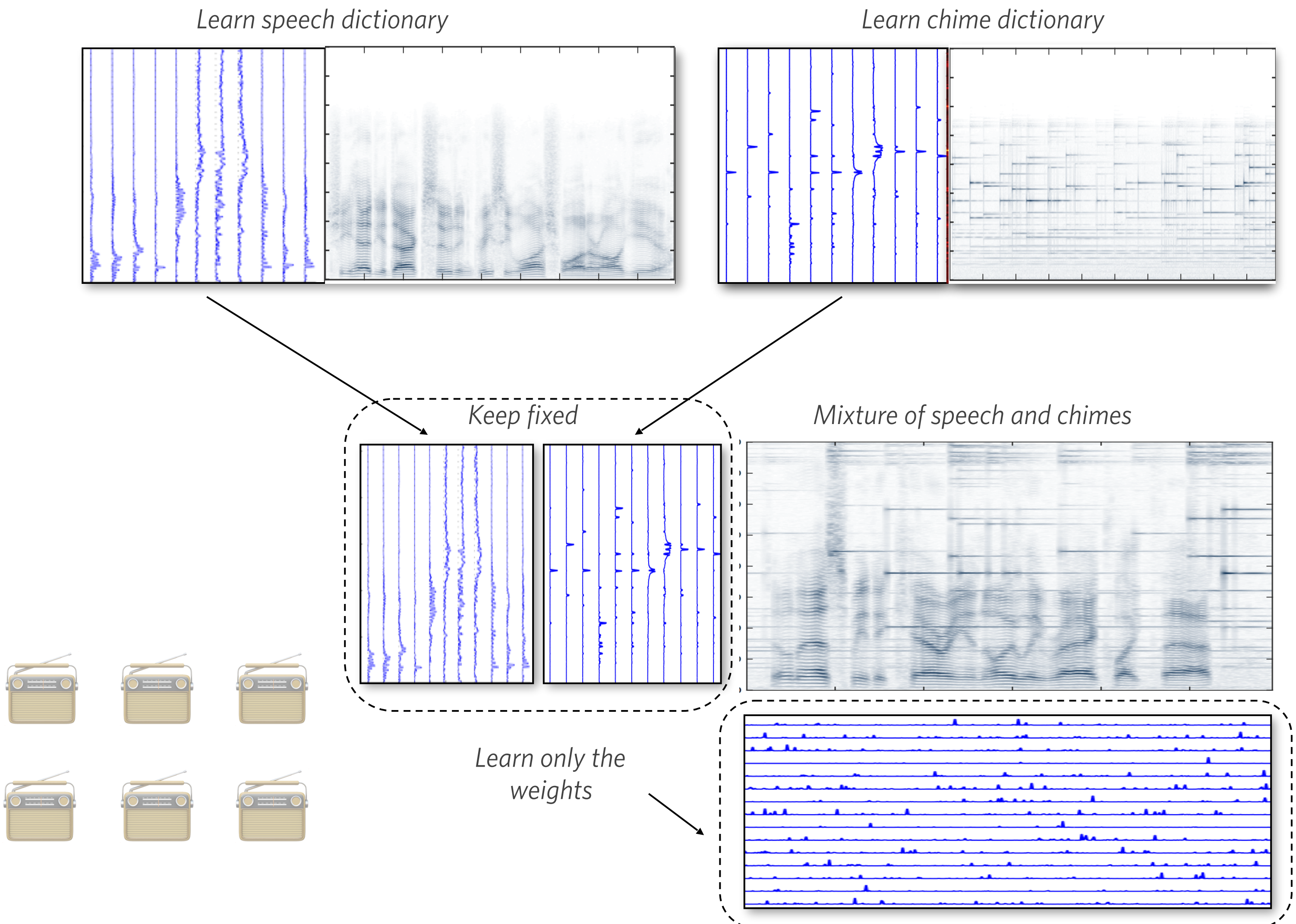
$$F(\omega,i) = \begin{cases} F_{chimes}(\omega,i) \\ F_{speech}(\omega,i) \end{cases}$$

Known/fixed
Estimated



Speech and Chimes

Huh?



Speech and Chimes

Speech and Speech

What if we don't have models?

- We usually don't know the frequency dictionary of all sounds in a mixture
  - We might know only some
- Complementary learning:

$$S(\omega,\tau) = \sum_i F(\omega,i)A(\tau,i)$$

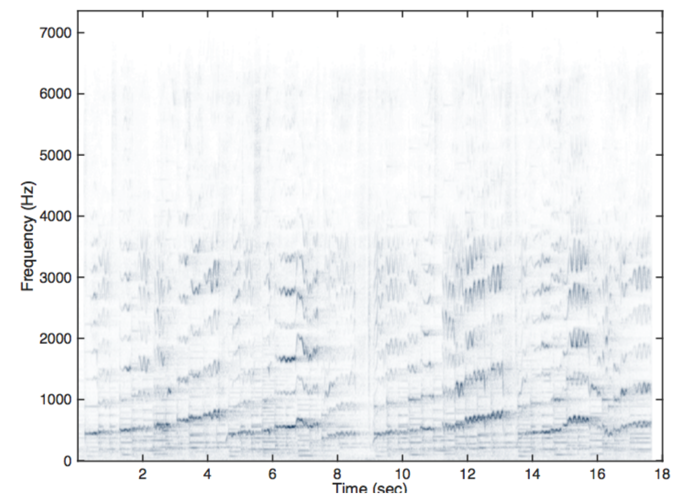
Known/fixed
Estimated

$$F(\omega,i) = \begin{cases} F_{known}(\omega,i) \\ F_{unknown}(\omega,i) \end{cases}$$

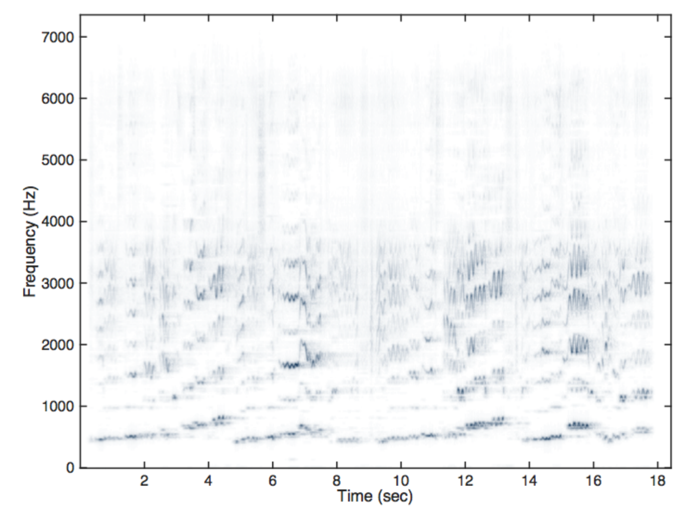
Soprano + Piano 

Soprano 

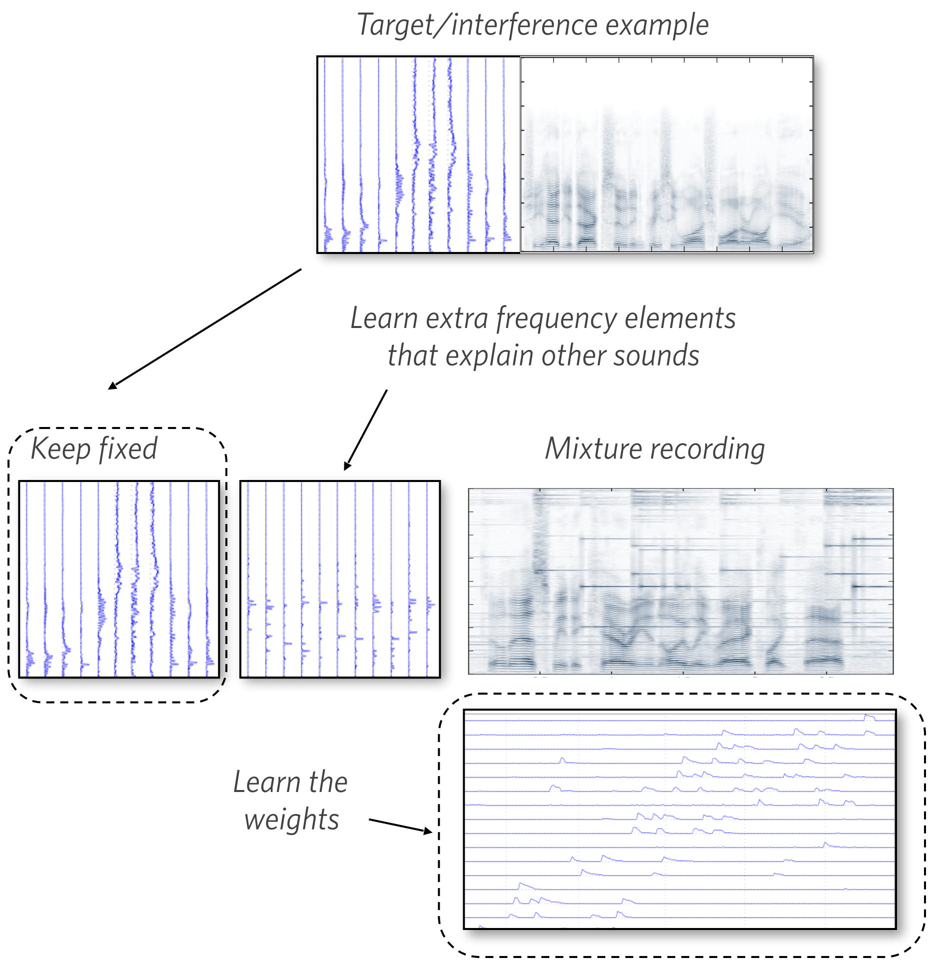
Soprano + Piano



Extracted soprano



Huh again?


Target/interference example
Learn extra frequency elements
that explain other sounds
Keep fixed
Mixture recording
Learn the
weights

Use in denoising

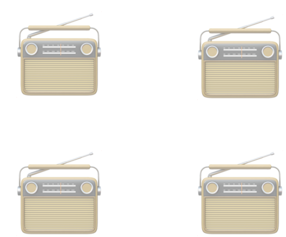
- Two cases
  - Have noise model, extract target
    - E.g. noise is street ambience
  - Have target, remove noise
    - E.g. cell phone knows your voice

- Denoising, Separation, Karaoke, …

Speech & the beauty of mechanics

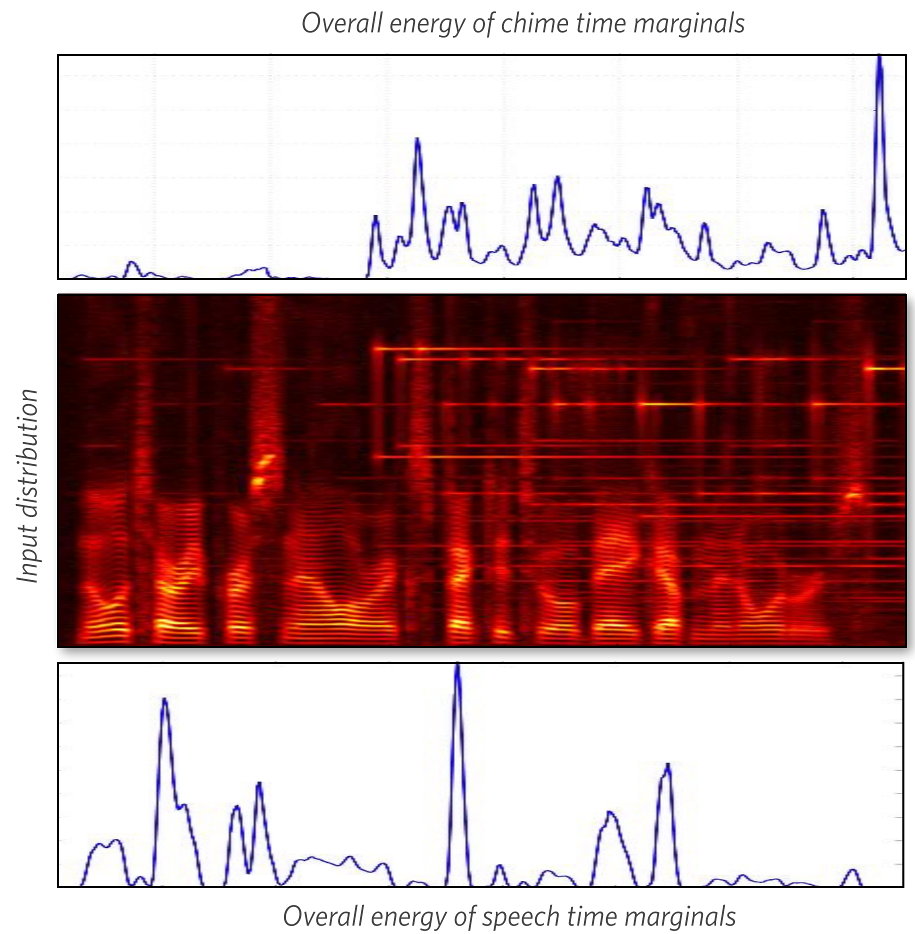
Wideband noise

Noise removal example



Presence of factors

- Time factors indicate amount of presence of a sound in a mixture
- This is very useful for analyzing scenes where we care to find sounds



Sound recognition in mixtures

- Sound classification is very poorly defined in machine learning
  - Uses winner-takes-all approach
  - But we have simultaneous sounds

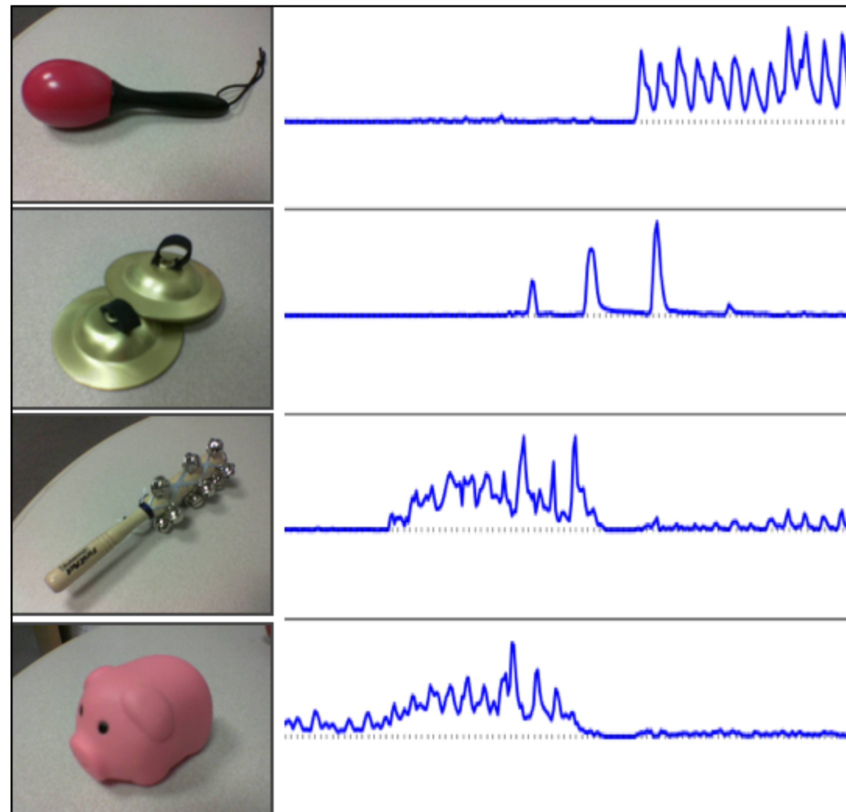
- Mixed sound recognition model

$$\underline{S(\omega,\tau)} = \sum_i \underline{\begin{bmatrix} P_{shaker}(\omega,i) \\ P_{cymbals}(\omega,i) \\ P_{jingles}(\omega,i) \\ P_{pig}(\omega,i) \end{bmatrix}} \cdot \underline{\begin{bmatrix} P_{shaker}(\tau,i) \\ P_{cymbals}(\tau,i) \\ P_{jingles}(\tau,i) \\ P_{pig}(\tau,i) \end{bmatrix}}$$

Known/fixed

Estimated

Estimated source likelihoods

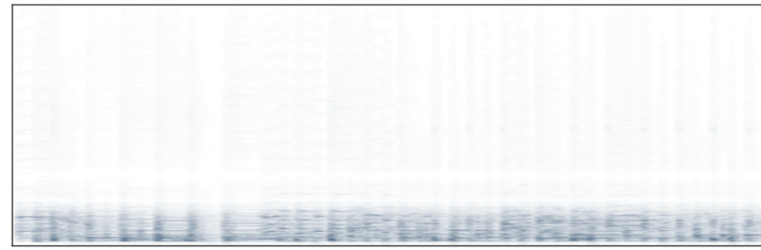


Audio superresolution

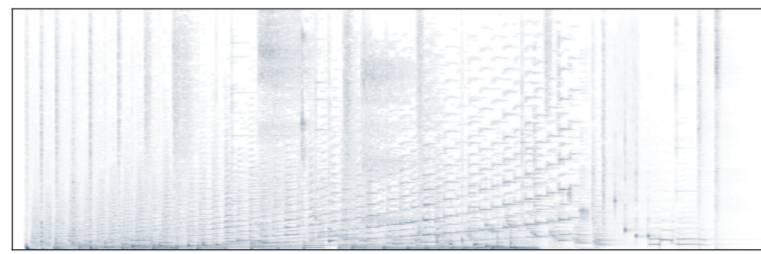
- Problem: We are missing some frequency bands
  - And lost frequencies are gone
    - forever …

- Upsampling by example
  - Fit sound class frequency factors to bandlimited data
  - Reconstruct from learned factors using full frequency range

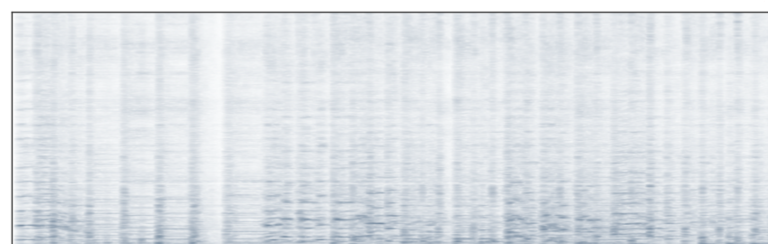
Bandlimited recording



Example sounds

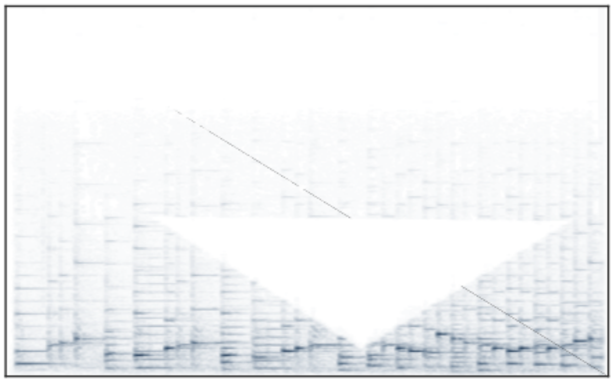


Expanded recording

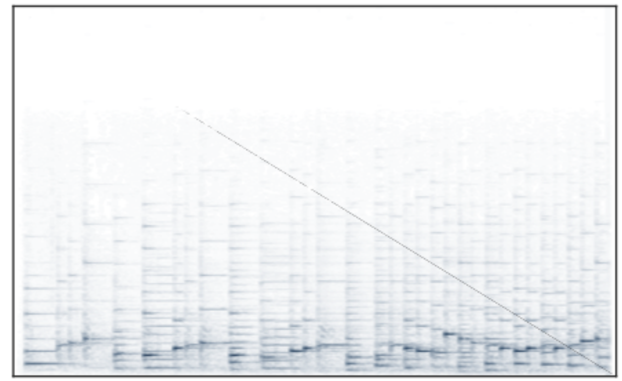


Ditto for missing data

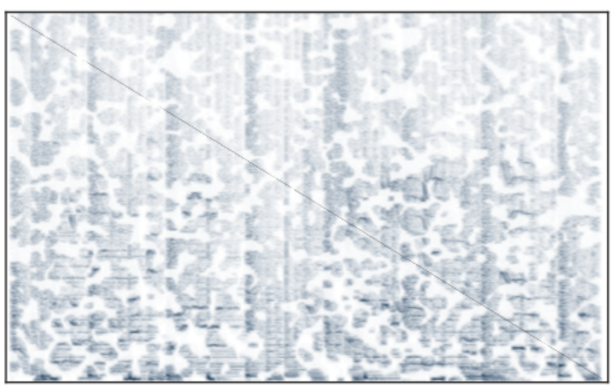
Large gap missing data input



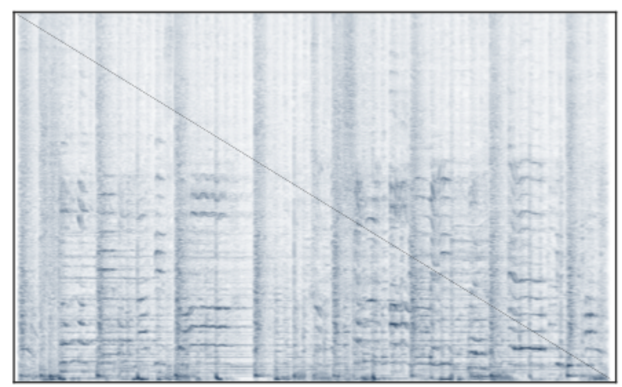
Recovered output



Random mask missing data input

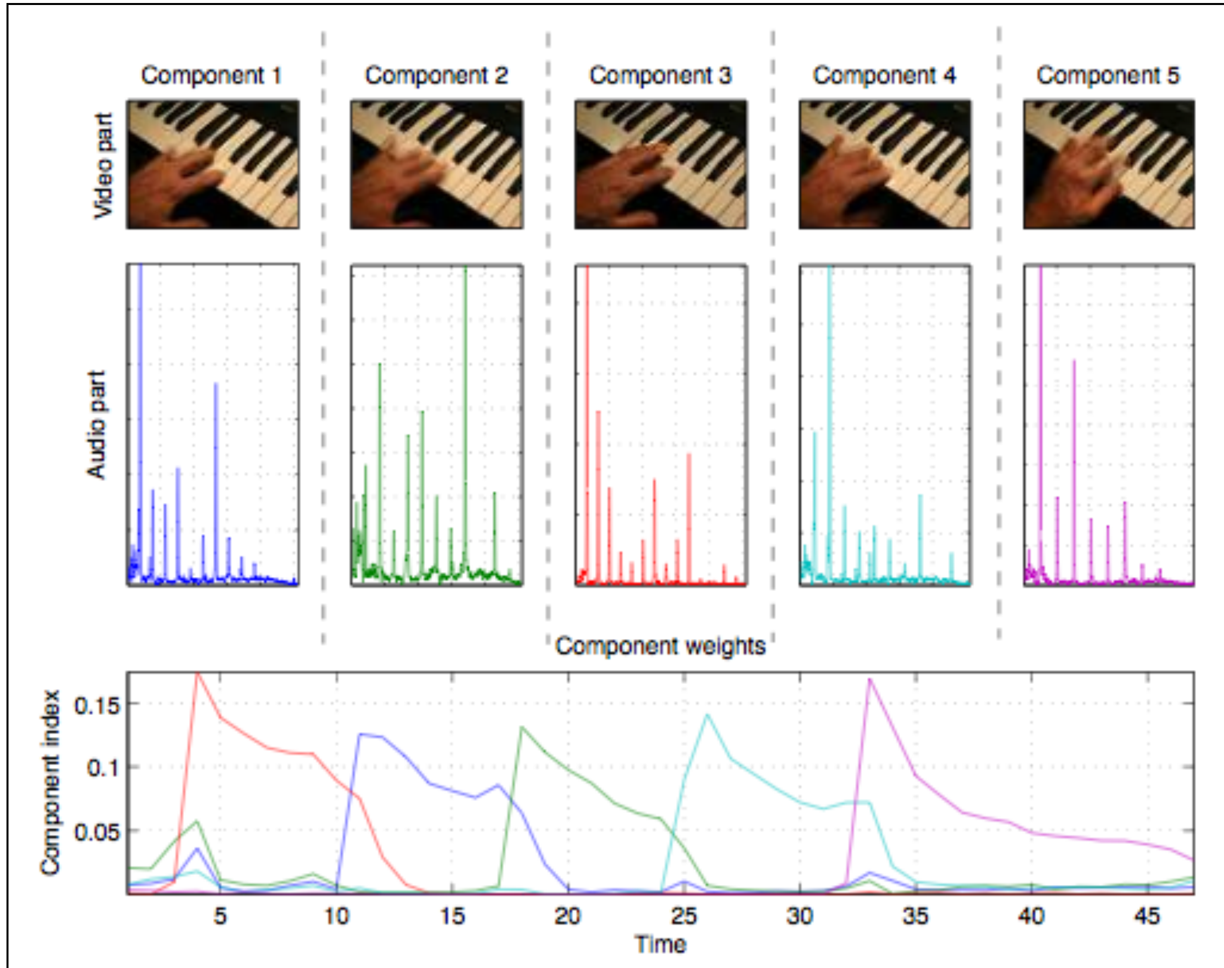


Recovered output



Audio-visual example

- Input with correlated images and pictures



All that's neat, but why?

- So far we focused on simple problems
  - Good for writing papers, but not for real-life

- Let's use this technology in the real-world
  - Sensing using audio
  - Being invariant to mixture issues

Video Content Analysis

- Audio is a strong cue for detecting various events in video
- Classify sounds to perform semantic analysis on video
  - Specific subclasses for type of broadcast (e.g. for news we use male and female speech, for sports use cheering, etc)
- Build in high-end Mitsubishi PVRs, TV sets and "HDTV cell phones"

Hard computer vision problems

Was there a goal?

Sad or funny clip?

Real-time movie sound parsing

Traffic Monitoring

- “Interesting” event discovery
  - Very hard/demanding visual task
  - Easier on the audio domain
- Reliable performance
  - ~5% false negatives,
  - ~4% false positives
- Hopefully we saved a life

  (or an insurance company …)

Examples of actual detected “interesting” parts

Normal crash

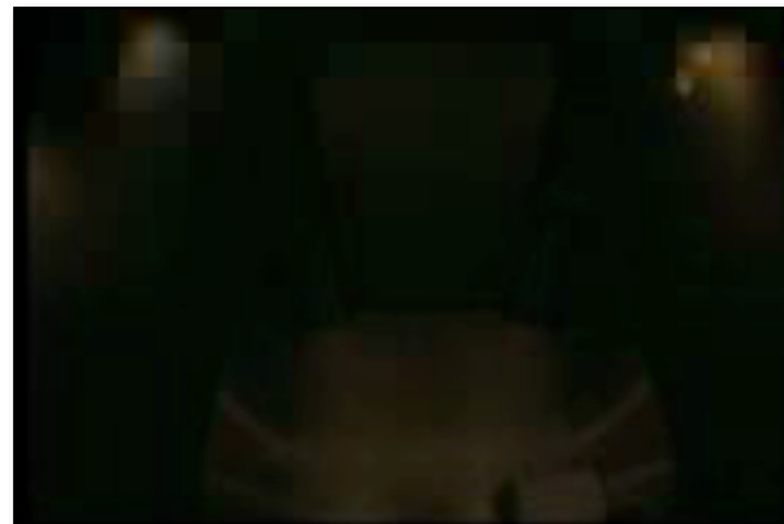
Hard-to-see crash

Near crash

Notable (?) event

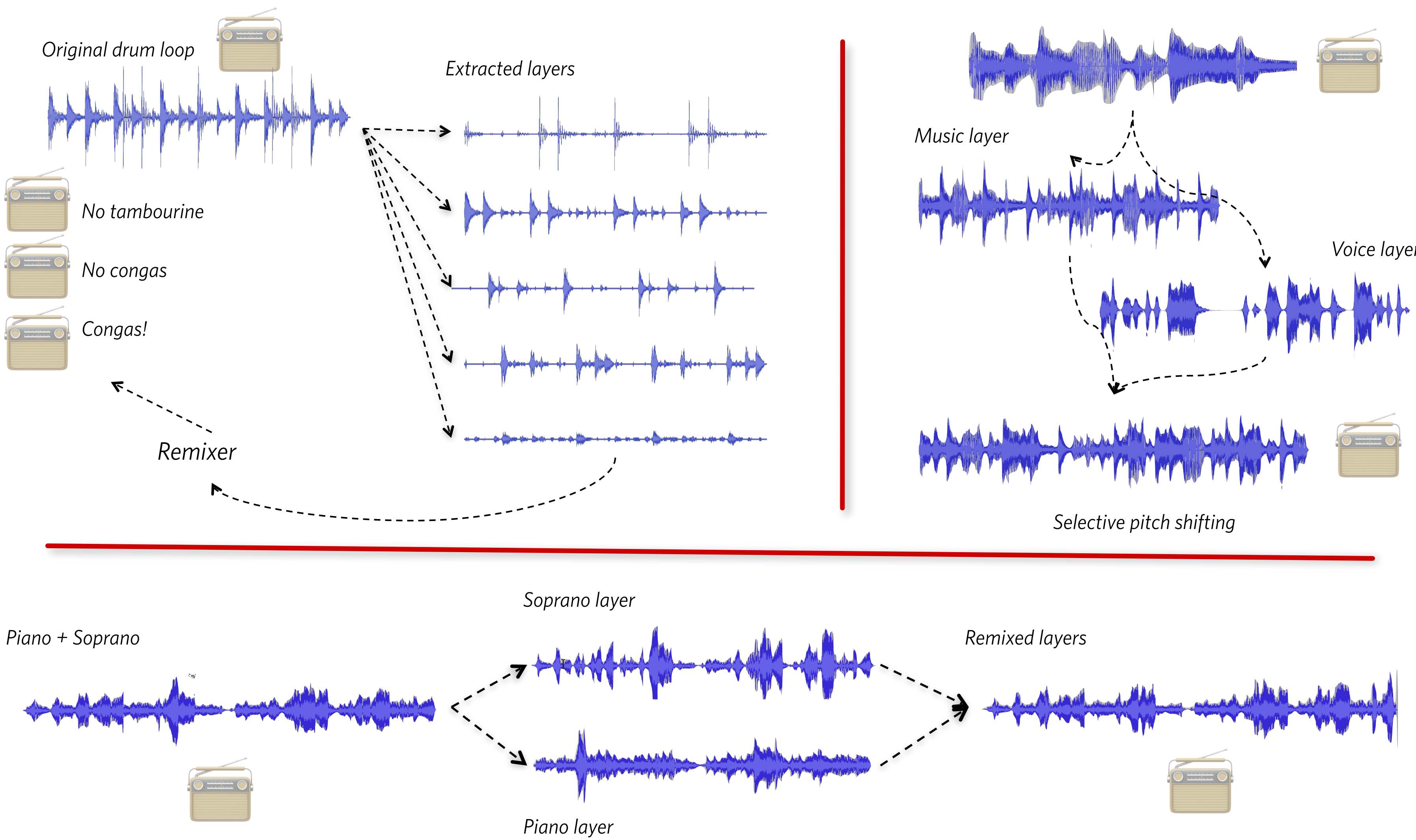
Security Surveillance

- Detect sounds in elevators
  - Normal speech, excited speech, footsteps, thumps, door open & close, screams, etc.
- When detecting suspicious sounds we can raise an alert
  - 96% accuracy in elevator test recordings with actors



Elevators are a dark environment with poor visual analysis prospects

Audio layer editing

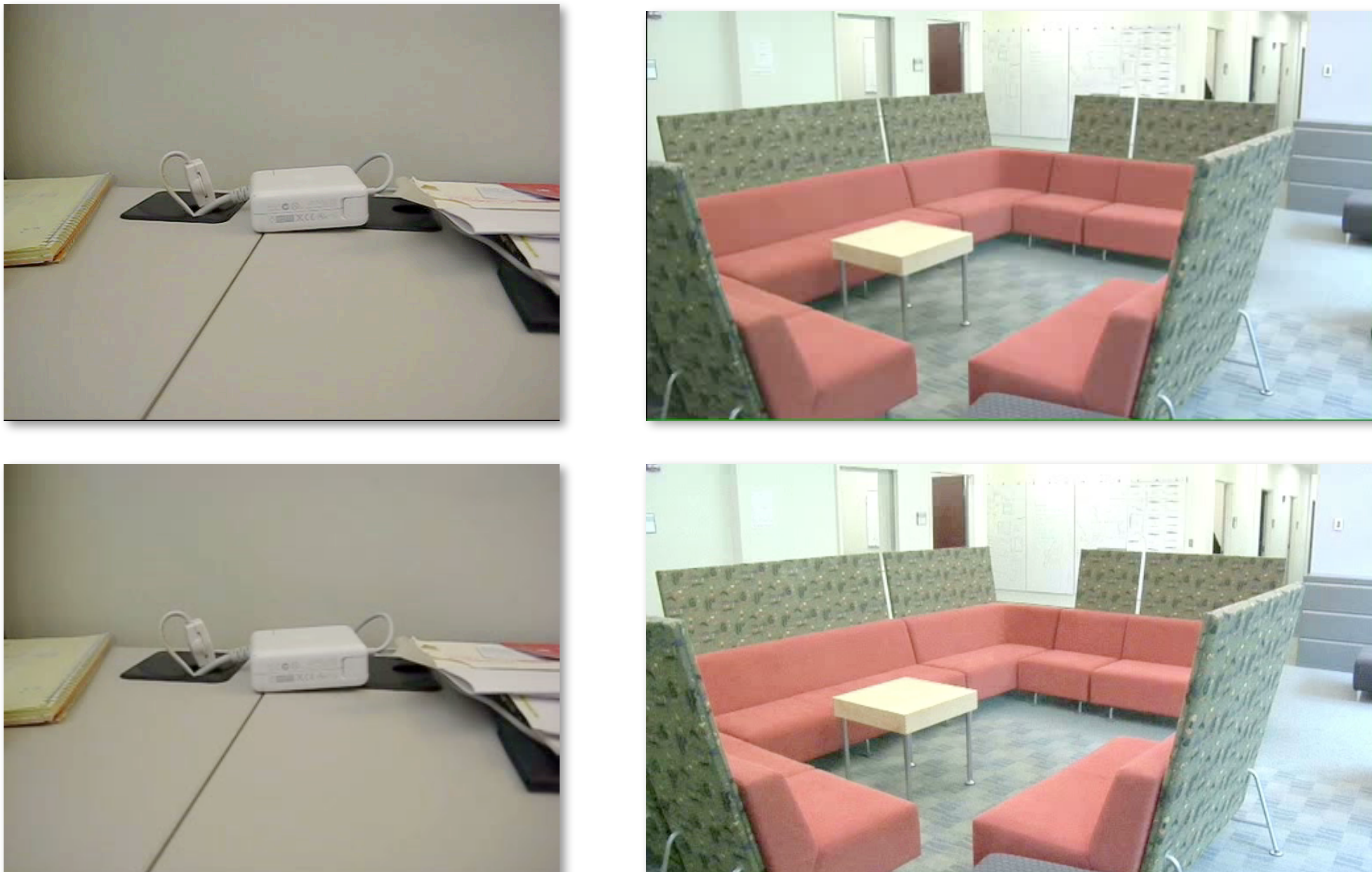


User-guided sound selection

Audio/visual editing

Input sequences

Output sequences



Many more things to do ...

- Mixtures are everywhere in DSP
  - Medical data, mechanical readings, ...
- I only talked about audio here
  - Lots of work to do in the field of mixtures
  - Lots of cool things coming out recently