LabROSA Research Overview

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/

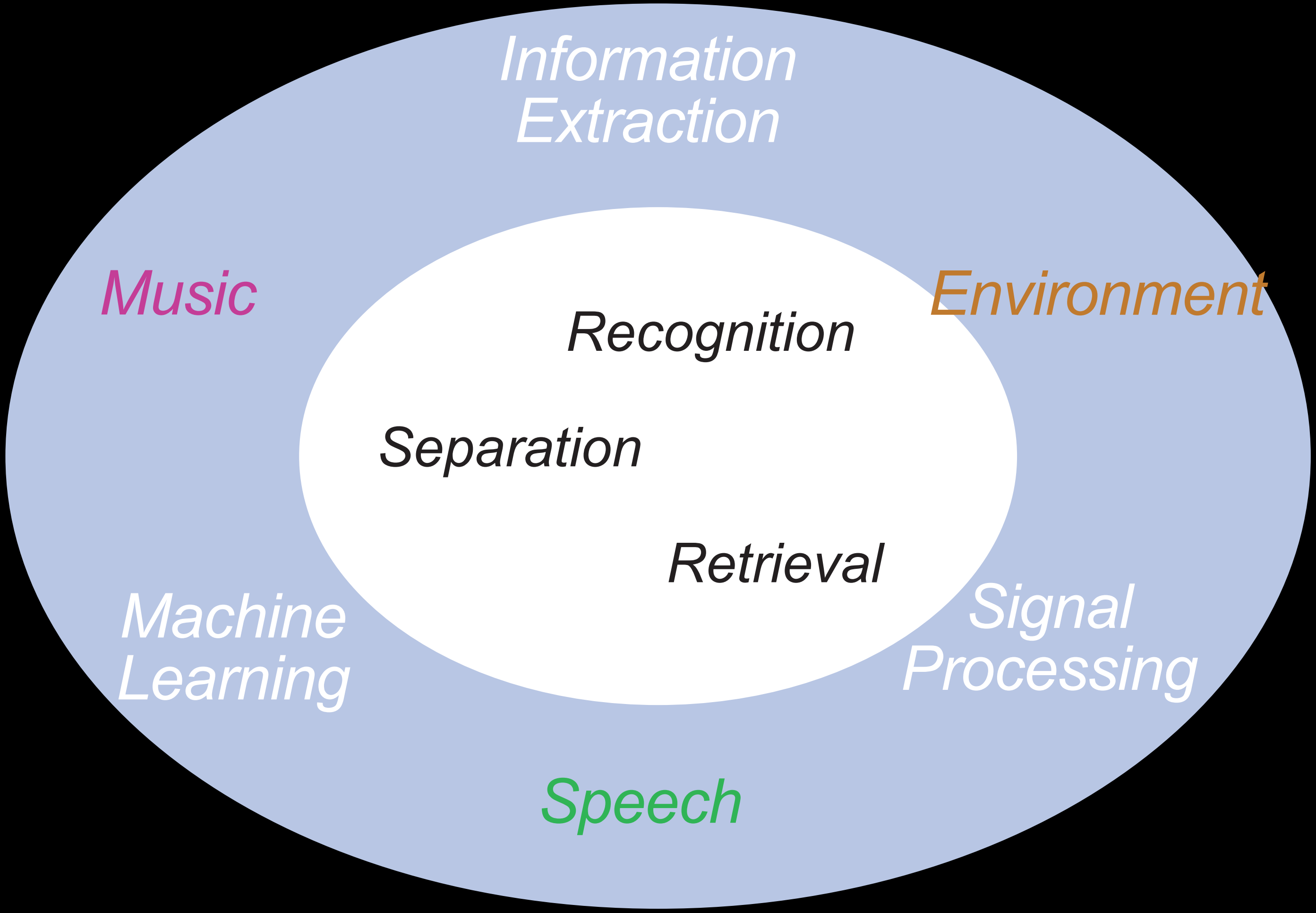
1. Music
2. Environmental sound
3. Speech Enhancement

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

Columbia University
in the city of New York

LabROSA

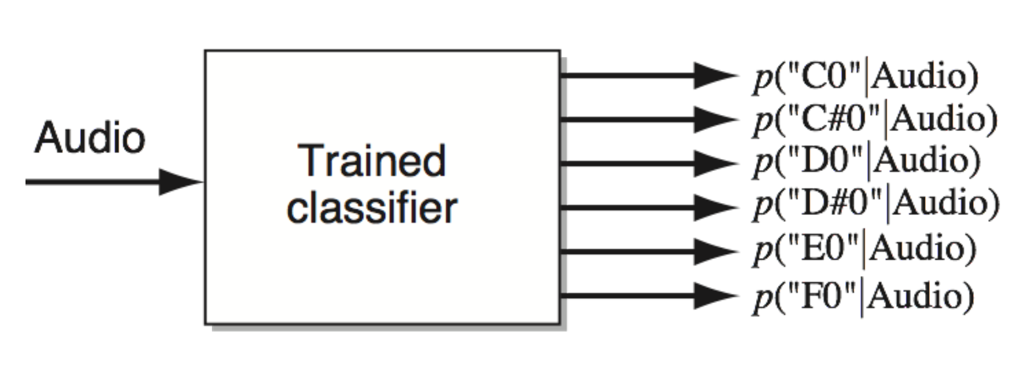
- Getting information from sound



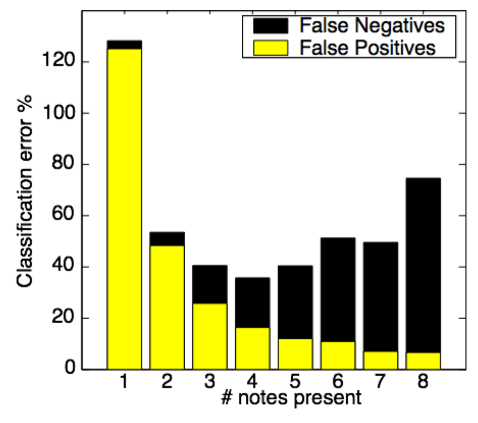
1. Music Audio Analysis

Poliner & Ellis '06

- Trained classifiers for low-level information
  - notes, chords, beats, section boundaries
- E.g. Polyphonic transcription


Audio
Trained classifier
$p(\text{"C0"}|\text{Audio})$
$p(\text{"C\#0"}|\text{Audio})$
$p(\text{"D0"}|\text{Audio})$
$p(\text{"D\#0"}|\text{Audio})$
$p(\text{"E0"}|\text{Audio})$
$p(\text{"F0"}|\text{Audio})$

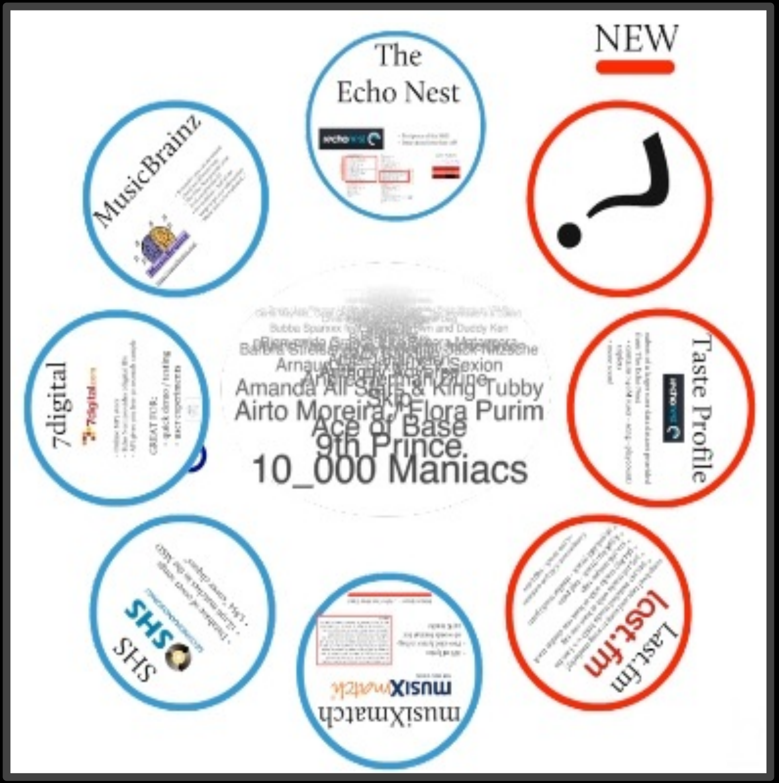
  - feature agnostic
  - needs training data



Million Song Dataset

Bertin-Mahieux
McFee

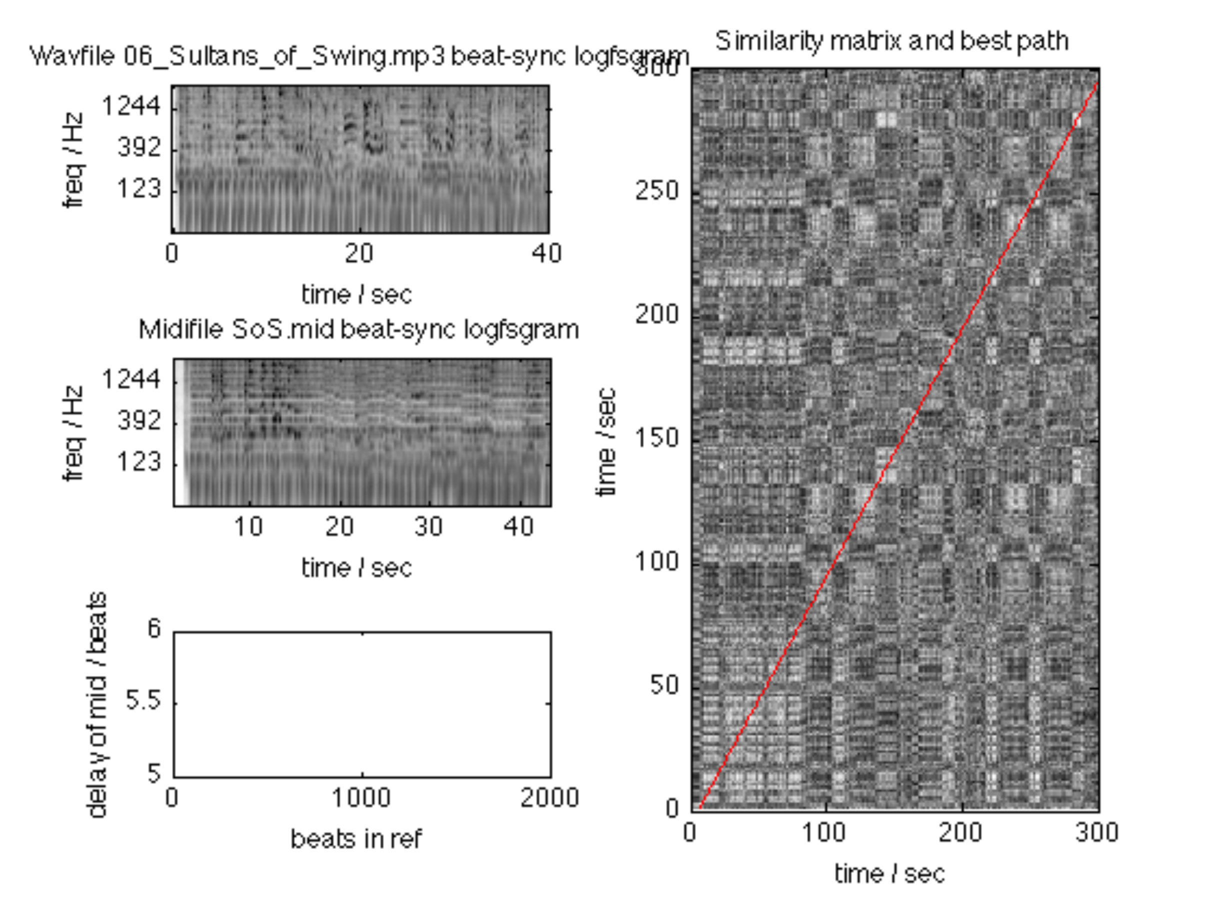
- **Industrial-scale database for music information research**
- **Many facets:**
  - Echo Nest audio features + metadata
  - Echo Nest "taste profile" user-song-listen count
  - Second Hand Song covers
  - musiXmatch lyric BoW
  - last.fm tags
- **Now with audio?**
  - resolving artist / album / track / duration against what.cd



MIDI-to-MSD

Raffel

- Aligned MIDI to Audio is a nice transcription

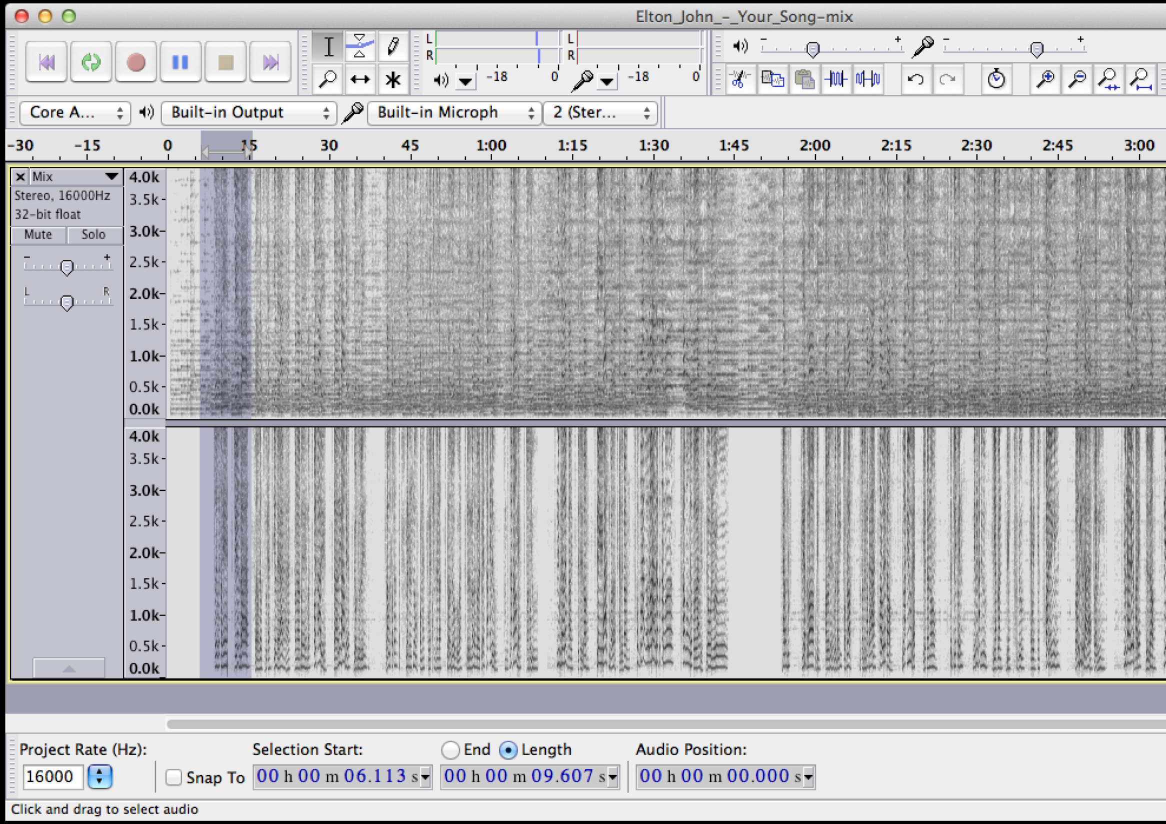


- Can we find matches in large databases?

Singing ASR

McVicar

- Speech recognition adapted to singing
  - needs aligned data
- Extensive work to line up scraped "acapellas" and full mix
  - including jumps!



Stereo, 16000Hz
32-bit float

Project Rate (Hz): 16000

☐ Snap To

Selection Start: 00 h 00 m 06.113 s

○ End ● Length: 00 h 00 m 09.607 s

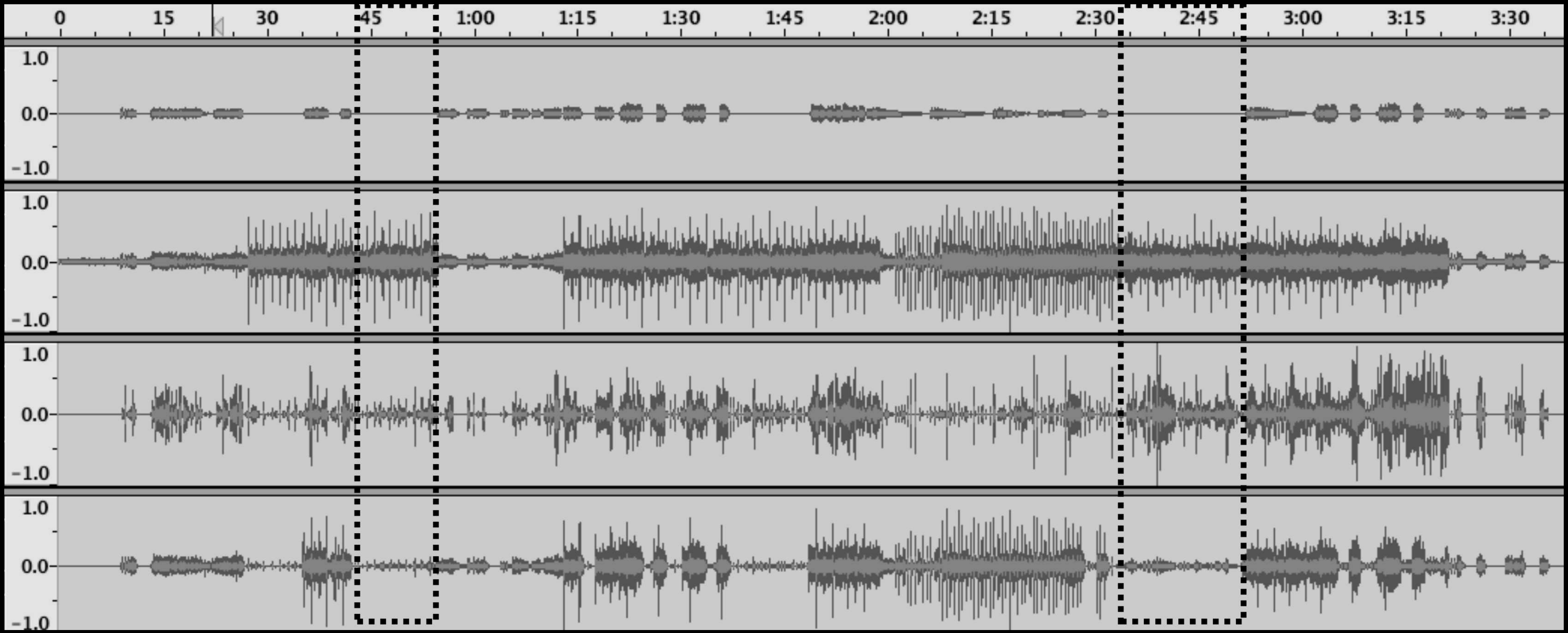
Audio Position: 00 h 00 m 00.000 s

Click and drag to select audio

Block Structure RPCA

Papadopoulos

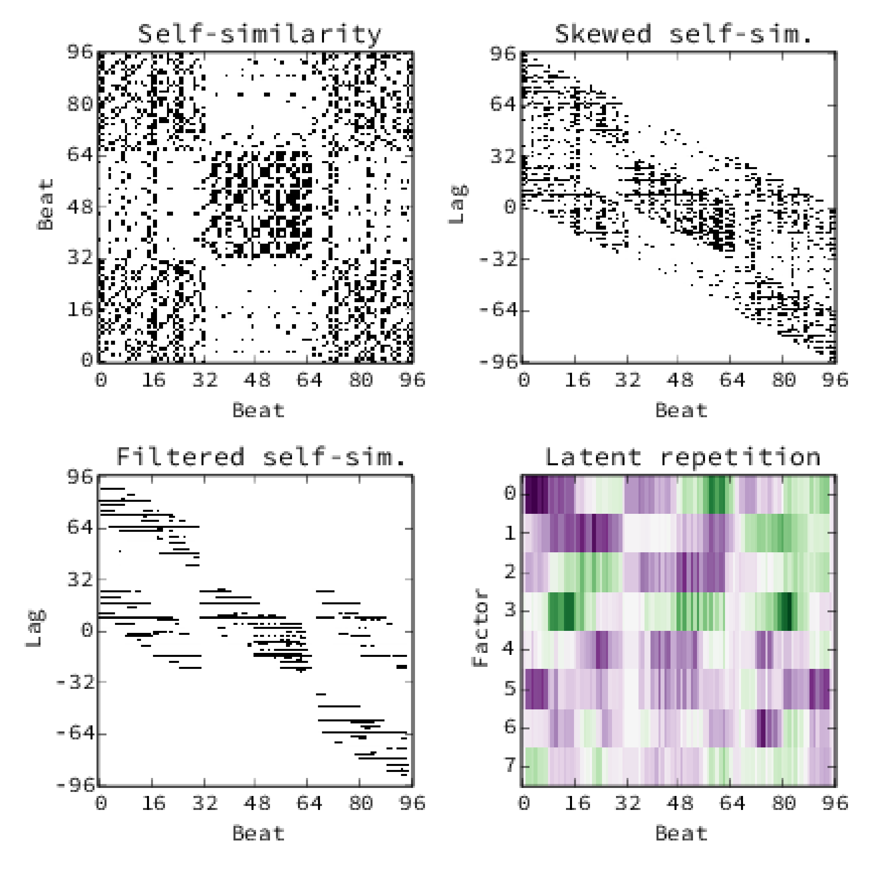
- RPCA separates vocals and background based on low rank optimization
  - single trade-off parameter
  - adjust based on higher-level musical features?



Ordinal LDA Segmentation

McFee

- Low-rank decomposition of skewed self-similarity to identify repeats
- Learned weighting of multiple factors to segment
  - Linear Discriminant Analysis between adjacent segments



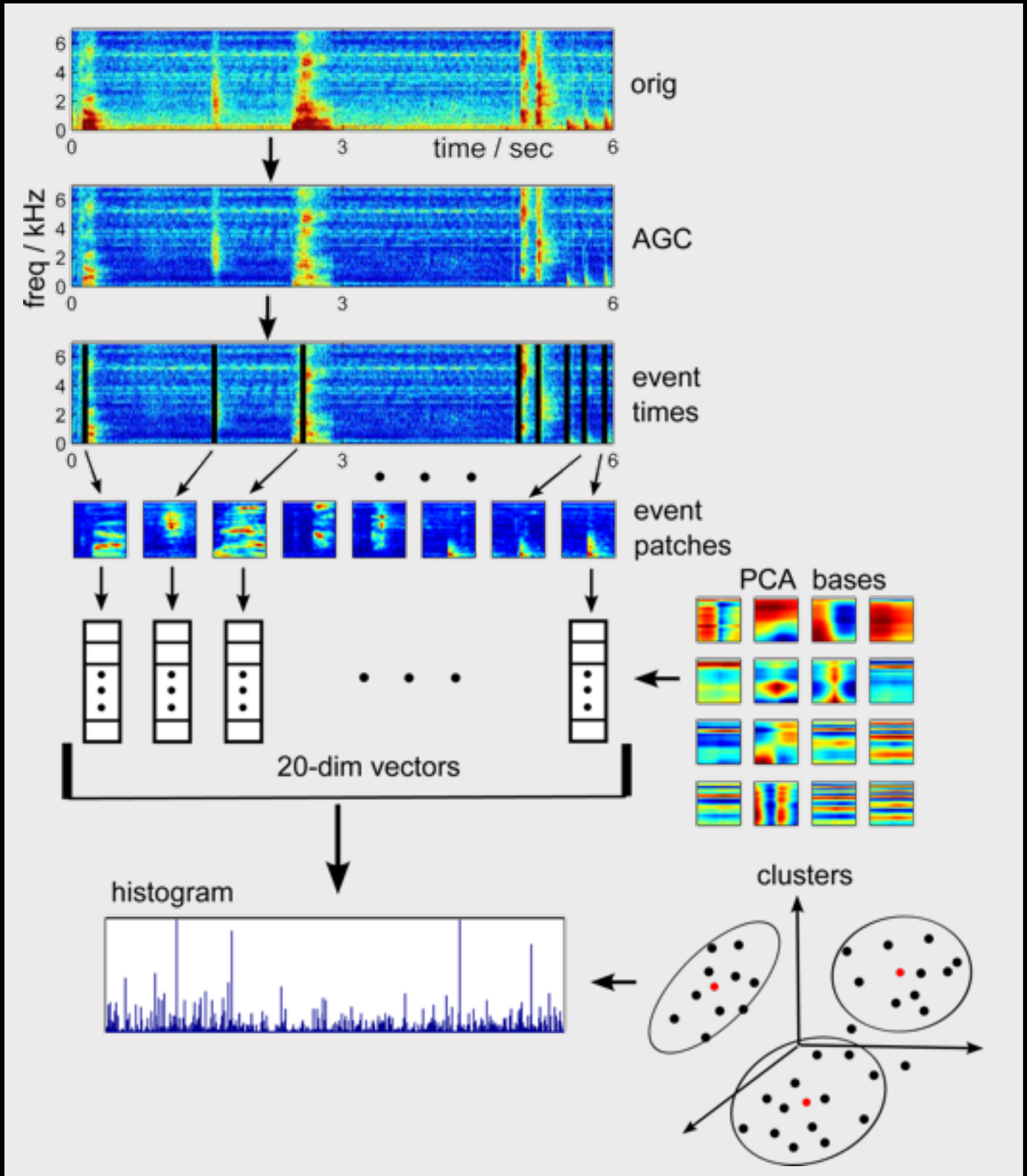
2. Environmental Sound

- Extracting useful information from soundtracks
- e.g. TRECVID Multimedia Event Detection (MED)
  - "Making a Sandwich", "Getting a Vehicle Unstuck"
  - 100 examples, find matches in 100k videos
  - manual annotations for ~10 h

E009 Getting a Vehicle Unstuck

Foreground Event Recognition

Cotton, Ellis, Loui '11



- Transients = foreground events?
- Onset detector finds energy bursts
  - best SNR
- PCA basis to represent each
  - 300 ms x auditory freq
- "bag of transients"

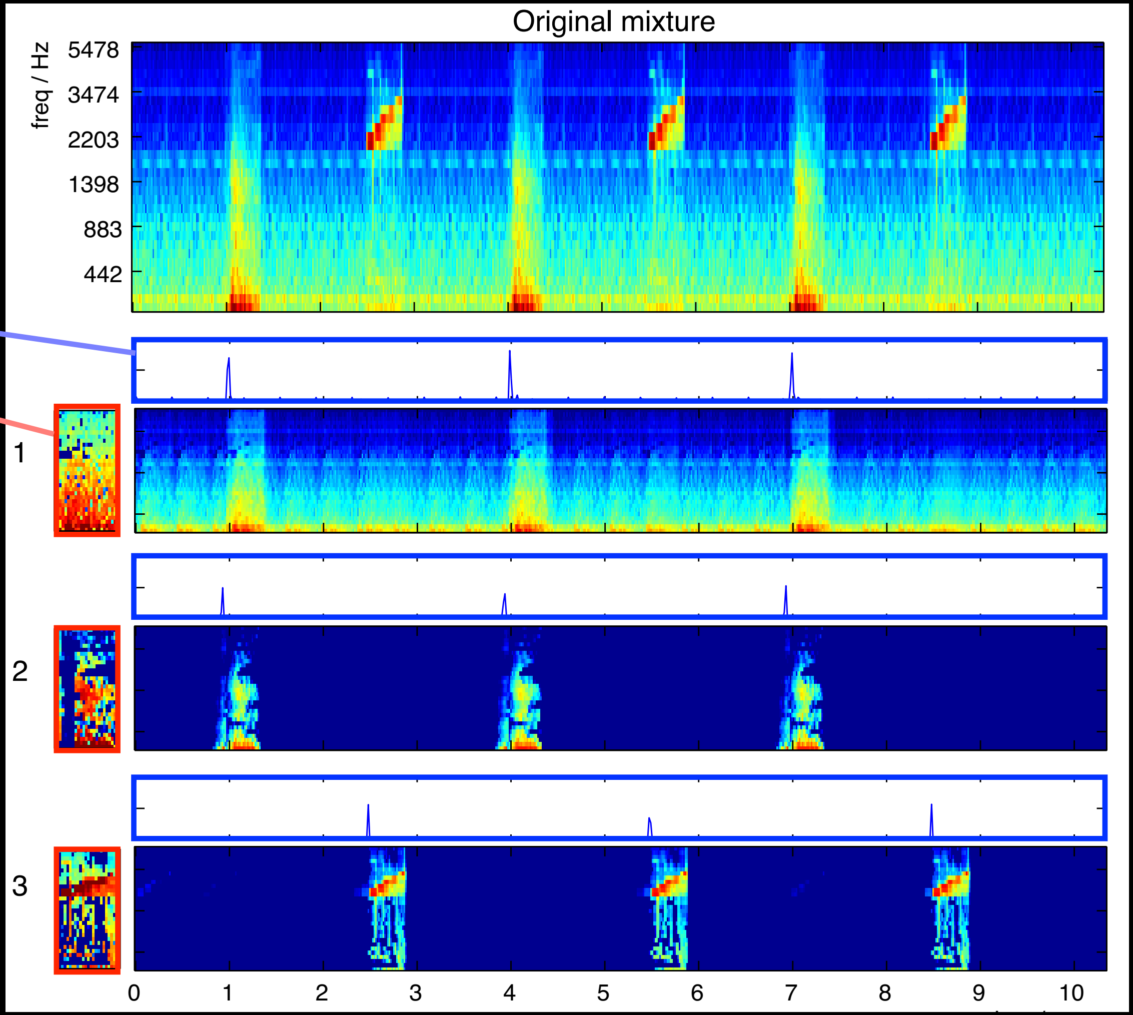
NMF Transient Features

Smaragdis & Brown ’03
Abdallah & Plumbley ’04
Virtanen ’07
Cotton & Ellis’ 11

- Decompose spectrograms into templates + activation

$$X = W \bullet H$$

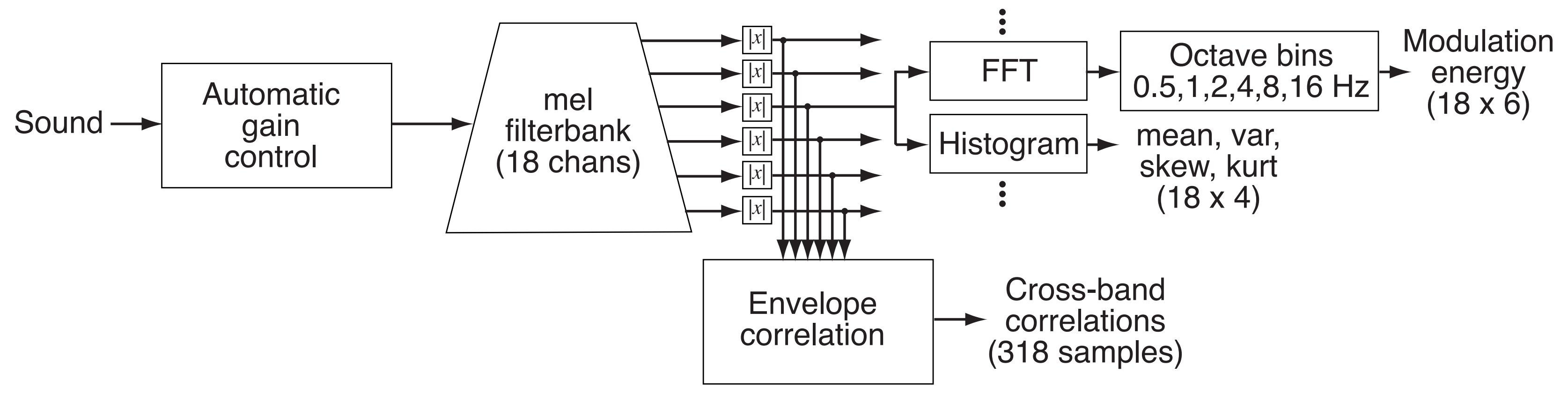
- well-behaved gradient descent
- 2D patches
- sparsity control
- computation time…



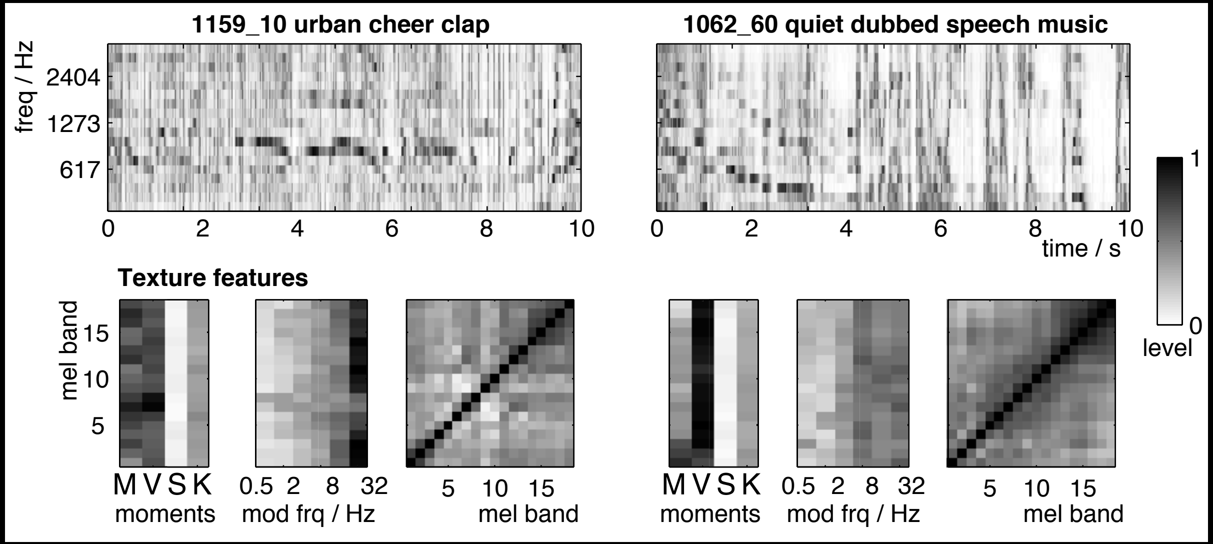
Background Retrieval

McDermott et al. '09
Ellis, Zheng, McDermott '11

- Classify soundtracks by statistics of ambience
- E.g. Texture features


Sound
Automatic gain control
mel filterbank (18 chans)
|x|
FFT
Octave bins 0.5,1,2,4,8,16 Hz
Modulation energy (18 x 6)
Histogram
mean, var, skew, kurt (18 x 4)
Envelope correlation
Cross-band correlations (318 samples)

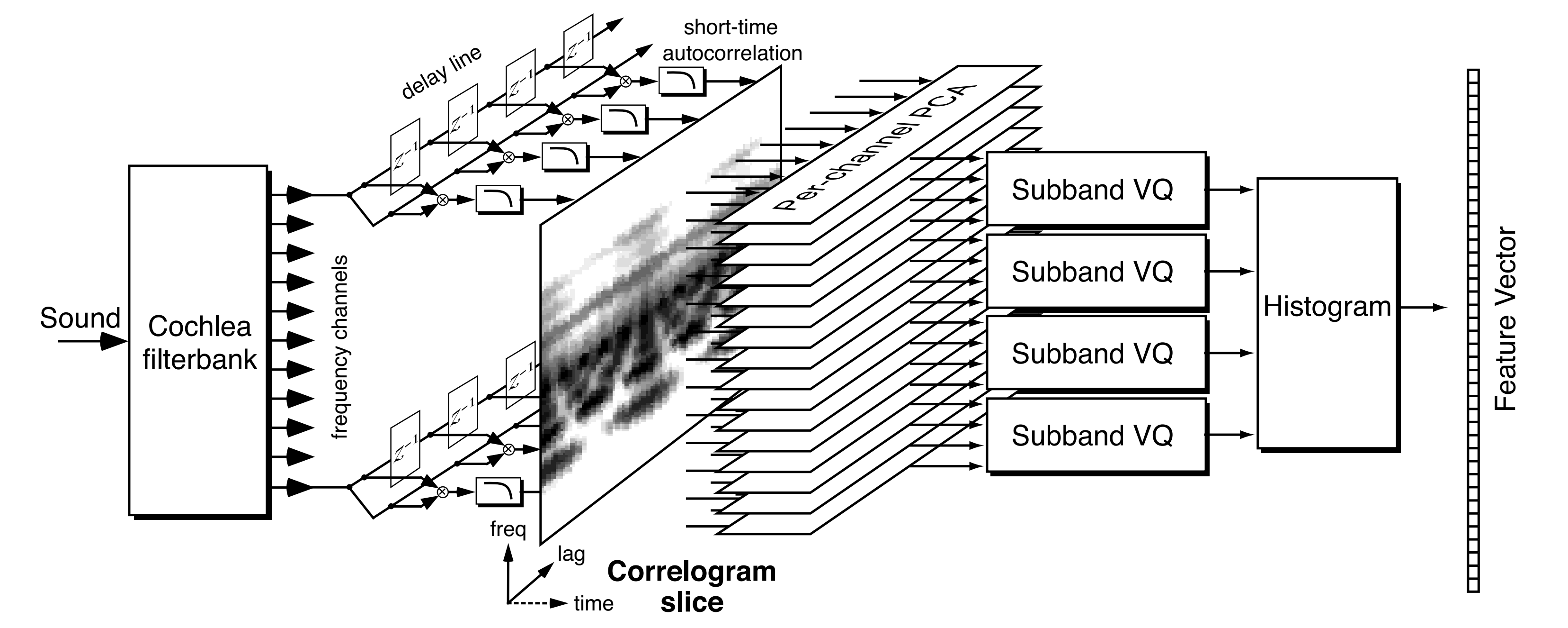
- Subband distributions
- Envelope cross-corrs



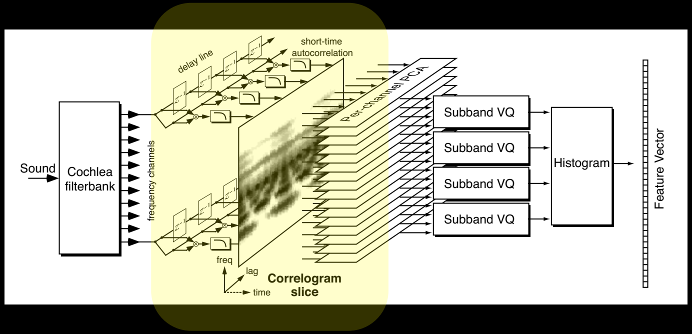
Auditory Model Features

Lyon et al. 2010
Lee & Ellis 2012
Cotton & Ellis 2013

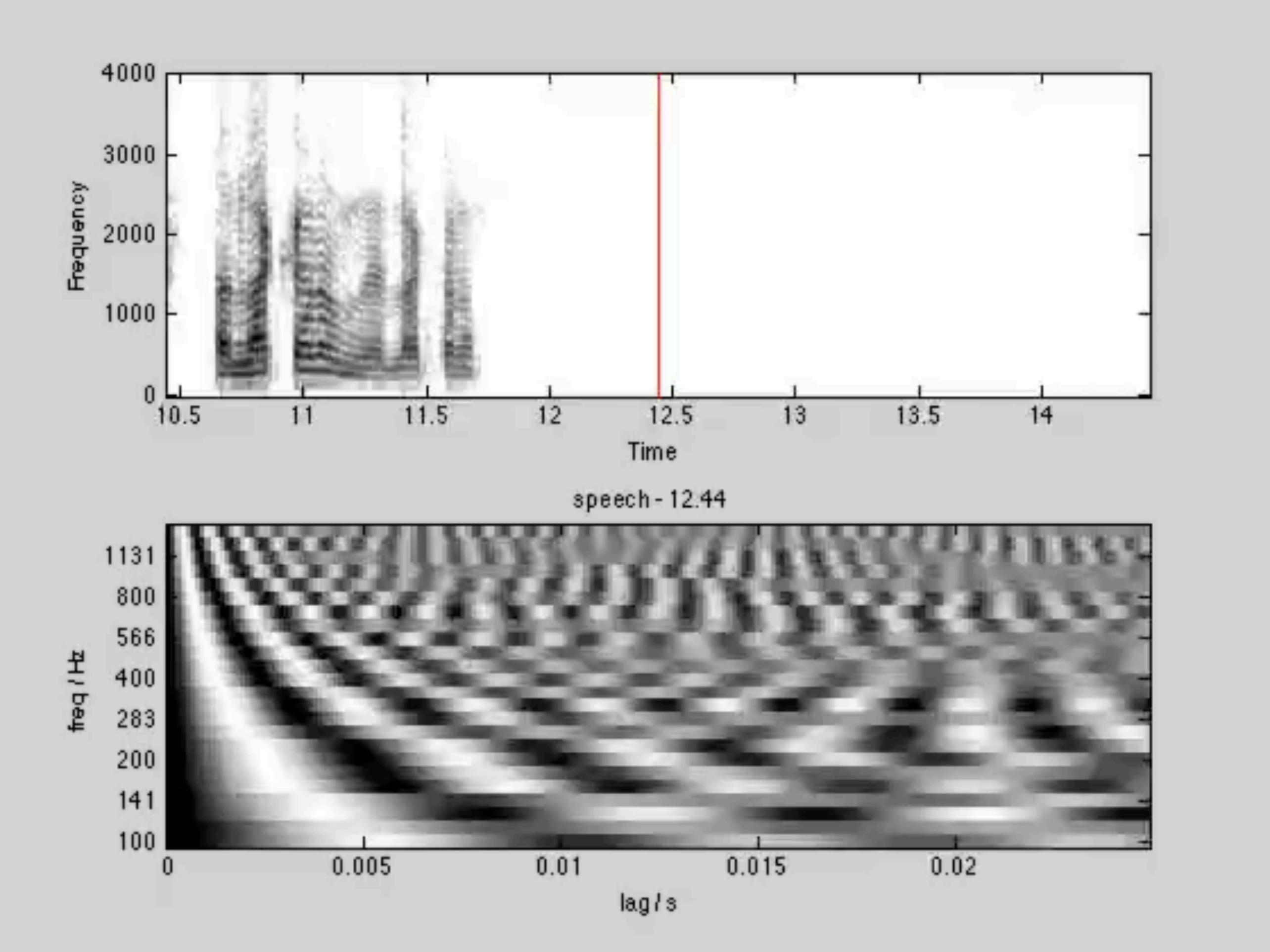
- **Subband Autocorrelation PCA**
  - Simplified version of autocorrelogram
  - 10x faster than Lyon original
- **Capture fine time structure in multiple bands**
  - information lost in MFCCs



Subband Autocorrelation



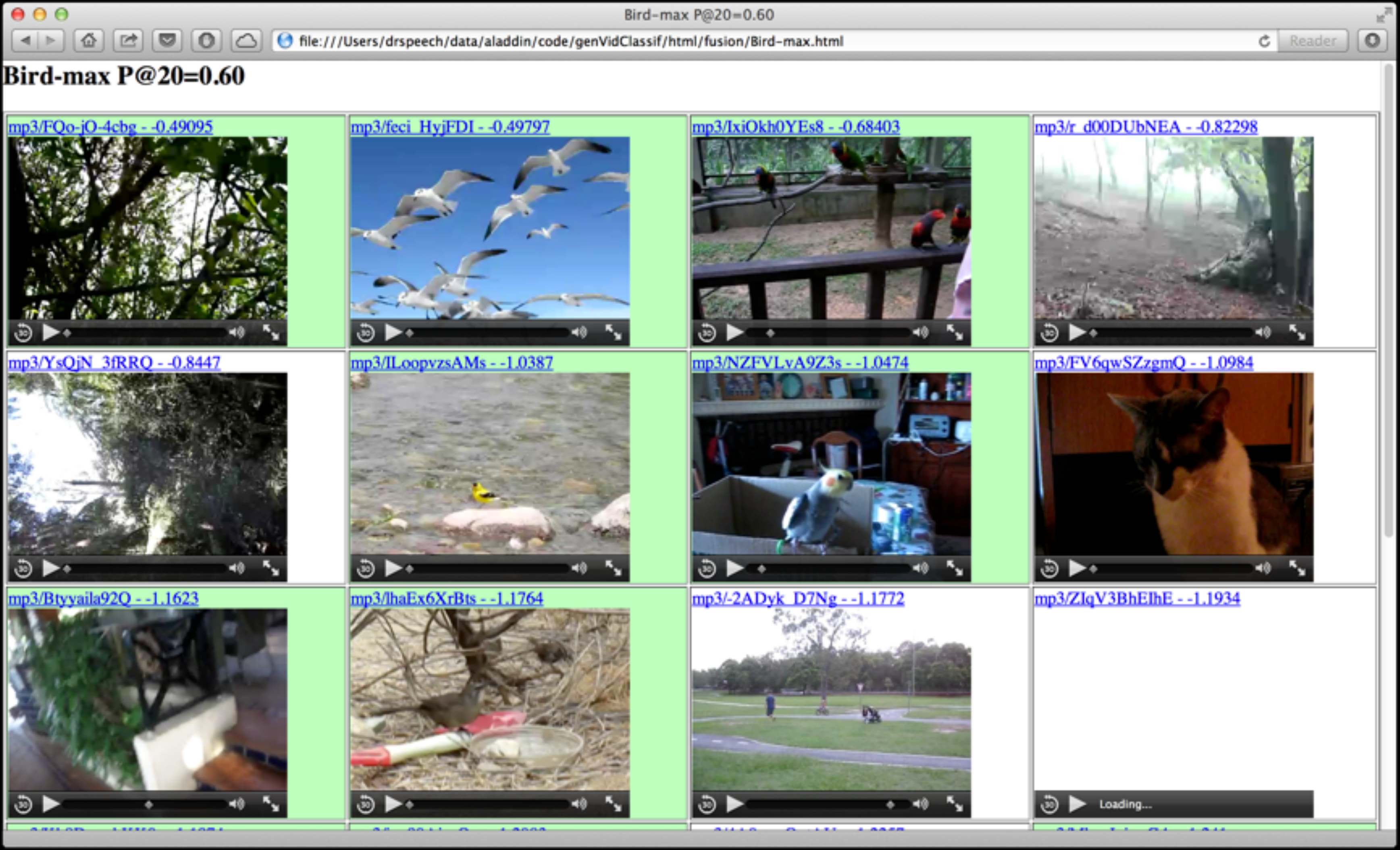
- Autocorrelation stabilizes fine time structure



- 25 ms window, lags up to 25 ms
- calculated every 10 ms
- normalized to max (zero lag)

Retrieval Examples

- High precision for in-domain top hits



Bird-max P@20=0.60

file:///Users/drspeech/data/aladdin/code/genVidClassif/html/fusion/Bird-max.html

Bird-max P@20=0.60

mp3/FQo-jO-4cbg - -0.49095

mp3/feci_HyjFDI - -0.49797

mp3/IxiOkh0YEs8 - -0.68403

mp3/r_d00DUbNEA - -0.82298

mp3/YsQjN_3fRRQ - -0.8447

mp3/ILoopvzsAMs - -1.0387

mp3/NZFVLvA9Z3s - -1.0474

mp3/FV6qwSZzgmQ - -1.0984

mp3/Btyyaila92Q - -1.1623

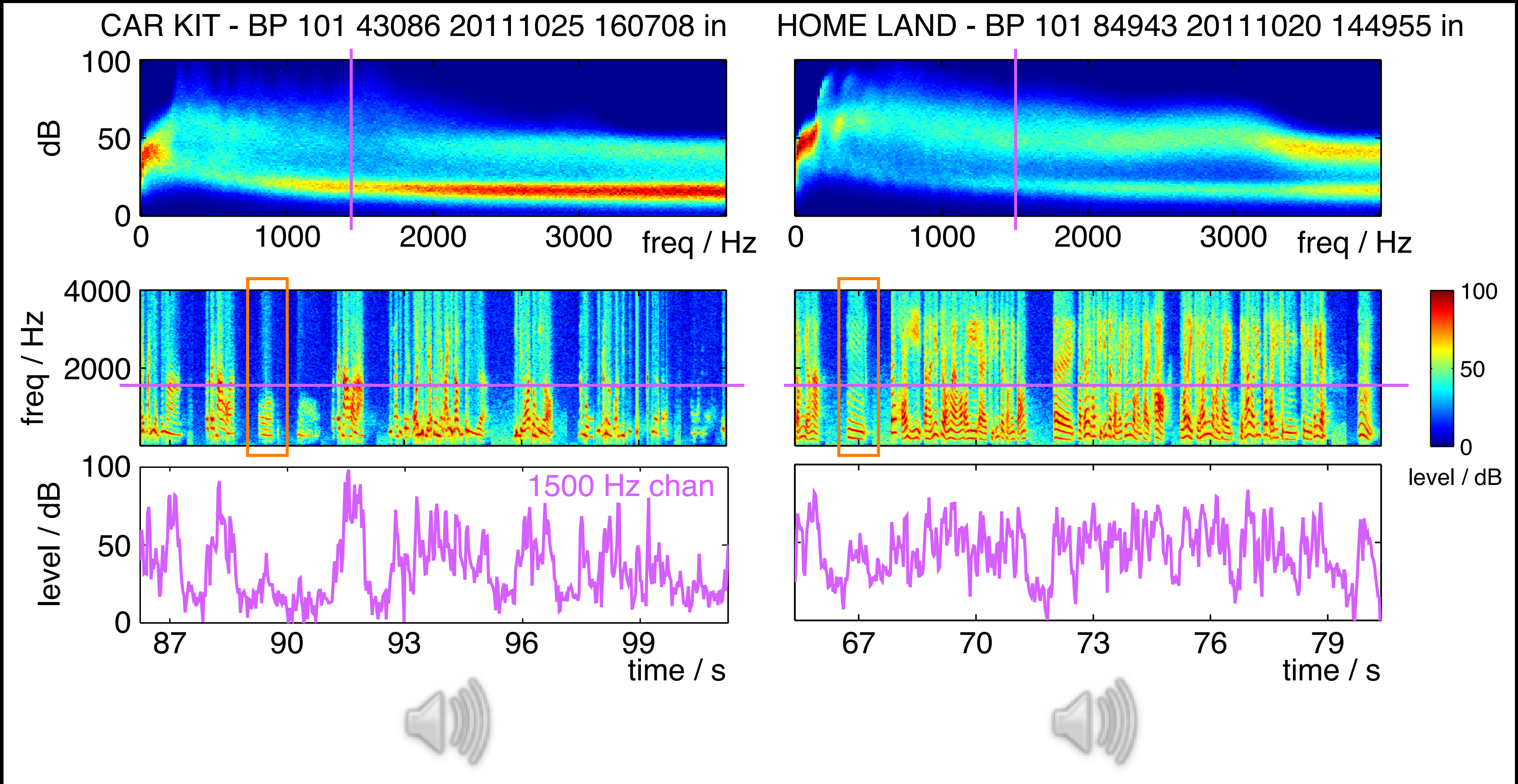
mp3/lhaEx6XrBts - -1.1764

mp3/-2ADyk_D7Ng - -1.1772

mp3/ZIqV3BhEIhE - -1.1934

3. Speech Enhancement

- Noisy speech scenarios
  - Ambient recording (background noise)
  - Communication channel (processing distortion)



RPCA Enhancement

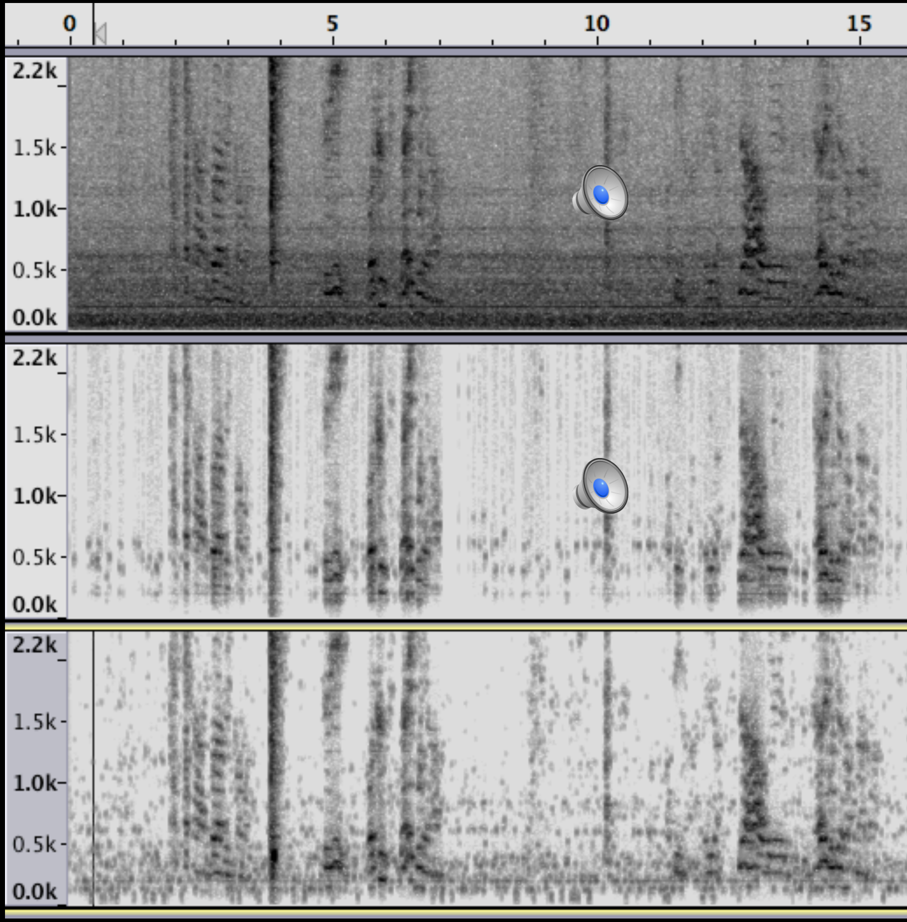
Chen, McFee & Ellis '14

- Decompose spectrogram into sparse + low-rank
- Sparse activation H of dictionary W

$$\min_{H,L,S} \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \lambda_S \|S\|_1$$
$$+ \mathcal{I}_+(H)$$
$$\text{s.t. } Y = WH + L + S$$

- ASR benefits:

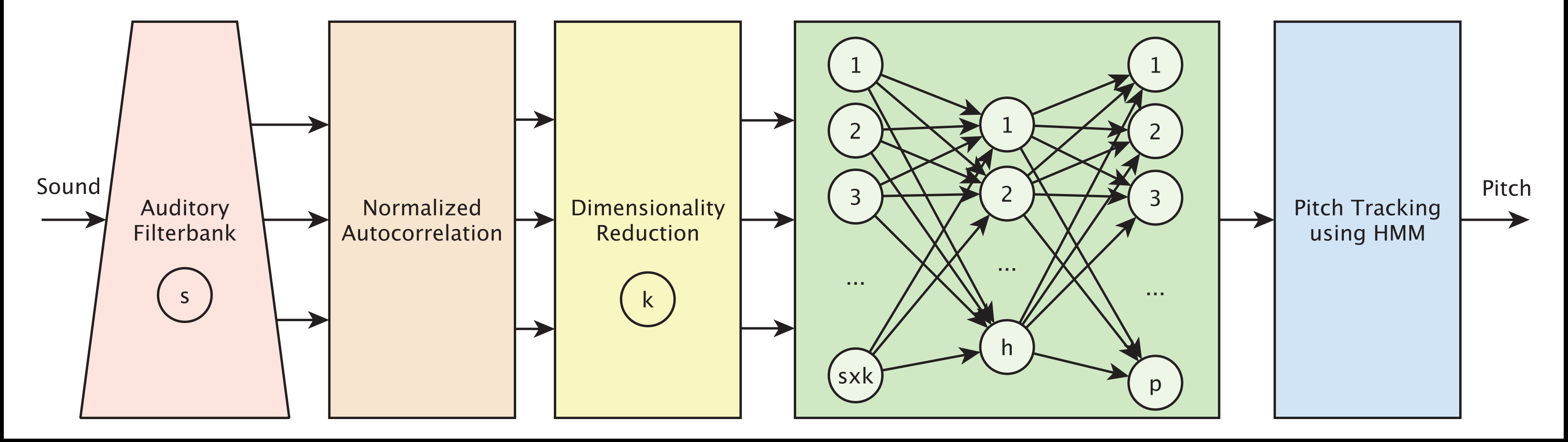
	C	S	D	I
Orig	6.8	10.6	82.6	0.7
RPCA	10.8	36.5	52.7	0.5
wie+RPCA	10.4	40.1	49.5	2.1



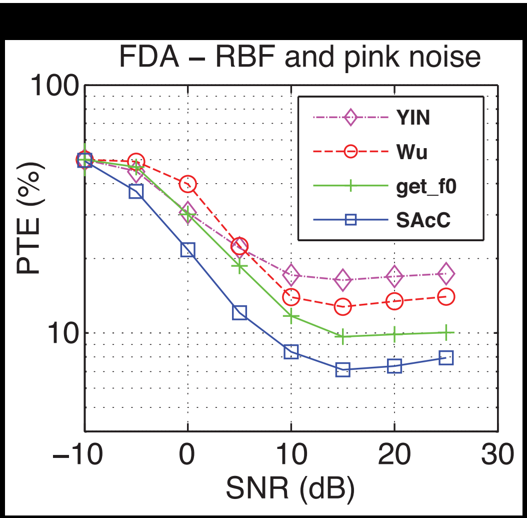
Classification Pitch Tracker

Lee & Ellis '12

- SAcC: MLP trained on noisy speech with ground-truth pitch track targets



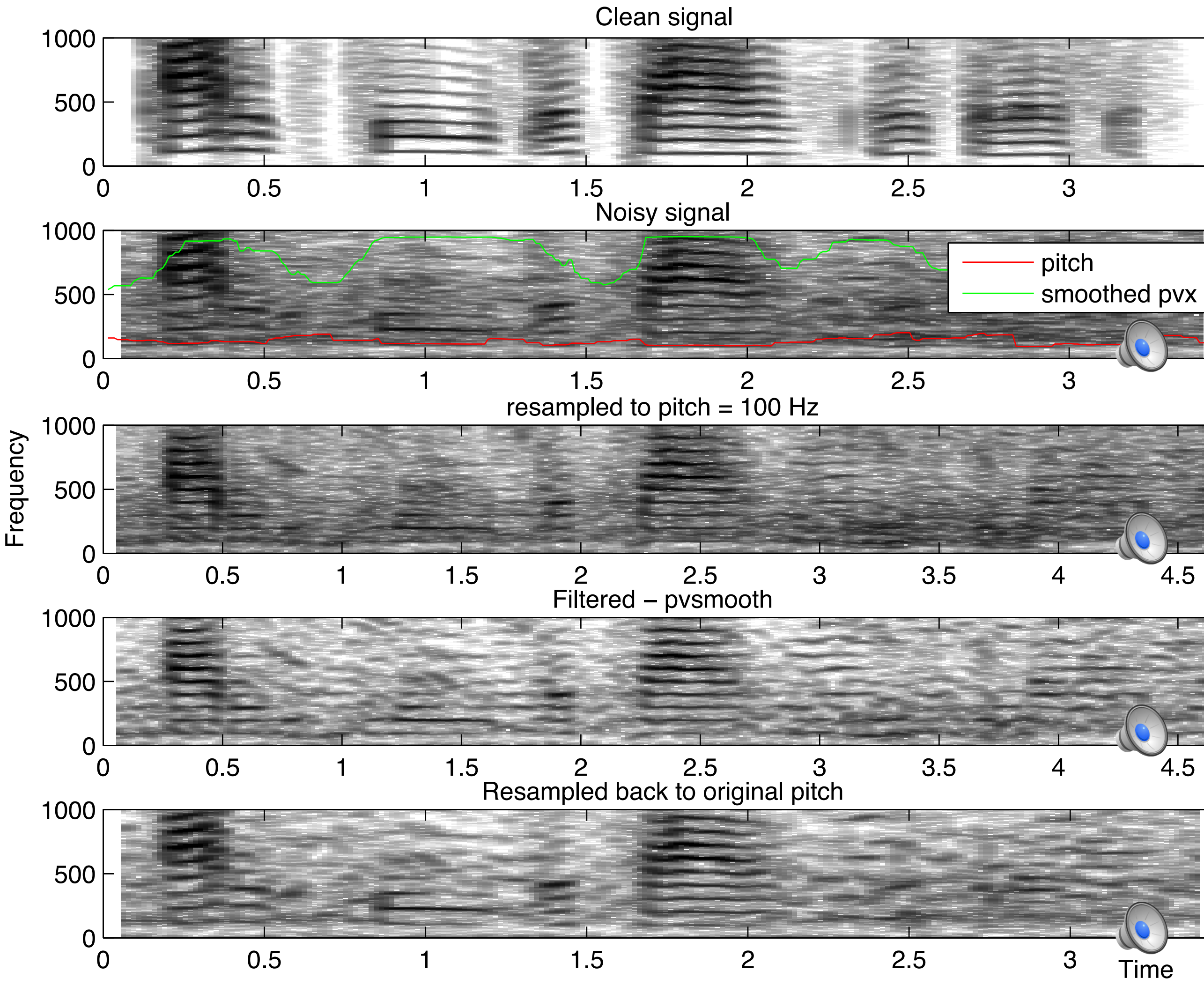
- Large benefits for in-domain noisy speech



Pitch-Normalized Enhancement

- Use noise-robust pitch tracker for enhancement?

- Normalize voice pitch
- Fixed-pitch enhancement
- Reimpose pitch



Summary

- Music
    - transcription, segmentation, …
    - alignment for ground truth

- Soundtracks
    - foreground events, background ambience

- Noisy Speech
    - classification pitch tracking
    - spectrogram enhancement