Subband Autocorrelation Features for Video Soundtrack Classification

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

{cvcotton,dpwe}@ee.columbia.edu http://labrosa.ee.columbia.edu/

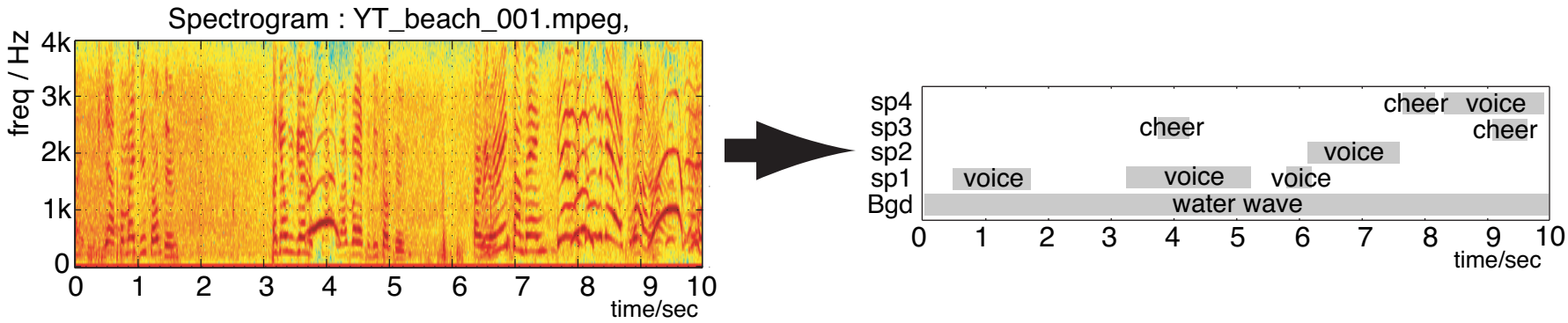
1. Soundtrack Classification
2. Auditory Model Features
3. Results & Future Work

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

Columbia University
In the City of New York

1. Soundtrack Classification

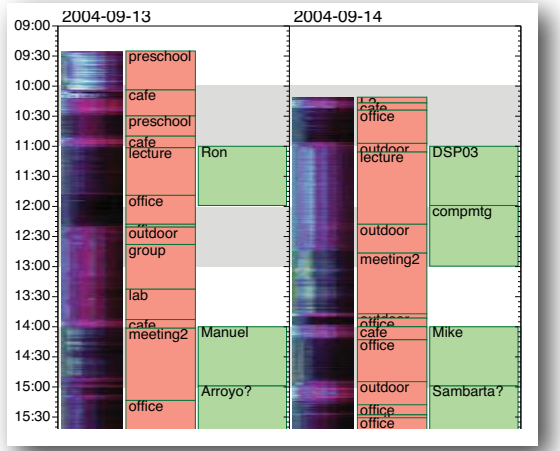
- Goal: Describe soundtracks with a vocabulary of user-relevant acoustic events/sources



- Challenges:
  - Defining acoustic event vocabulary
  - Overlapping sounds
  - Ground-truth training data
  - Classifier accuracy

Environmental Sound Applications

- Audio Lifelog
  Diarization



- Consumer Video
  Classification & Search

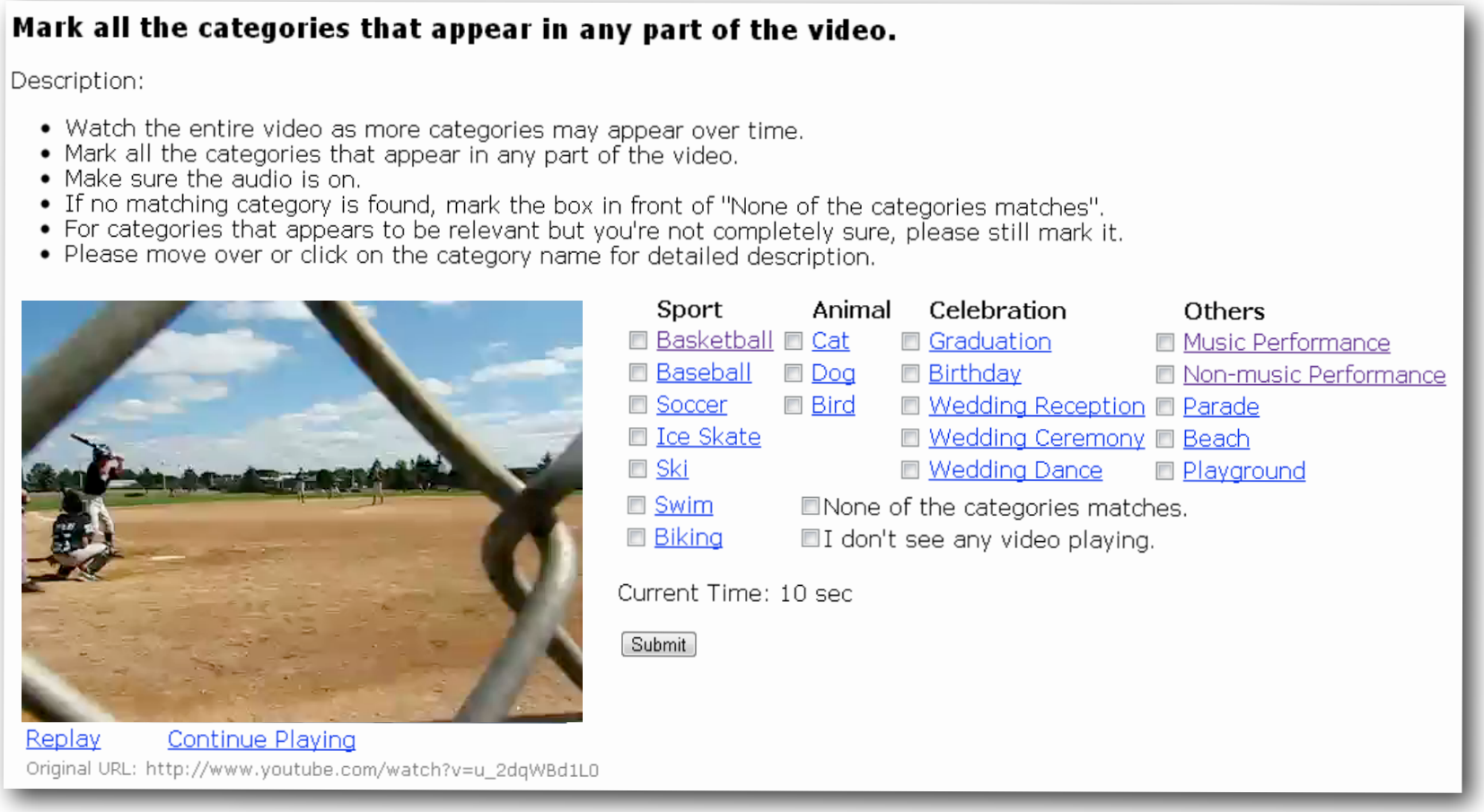
- Live hearing prosthesis app

- Robot environment sensitivity

Consumer Video Dataset

Y-G. Jiang et al. 2011

- Columbia Consumer Video (CCV) set
  - 9,317 videos / 210 hours
  - 20 concepts based on consumer user study
  - Labeled via Amazon Mechanical Turk



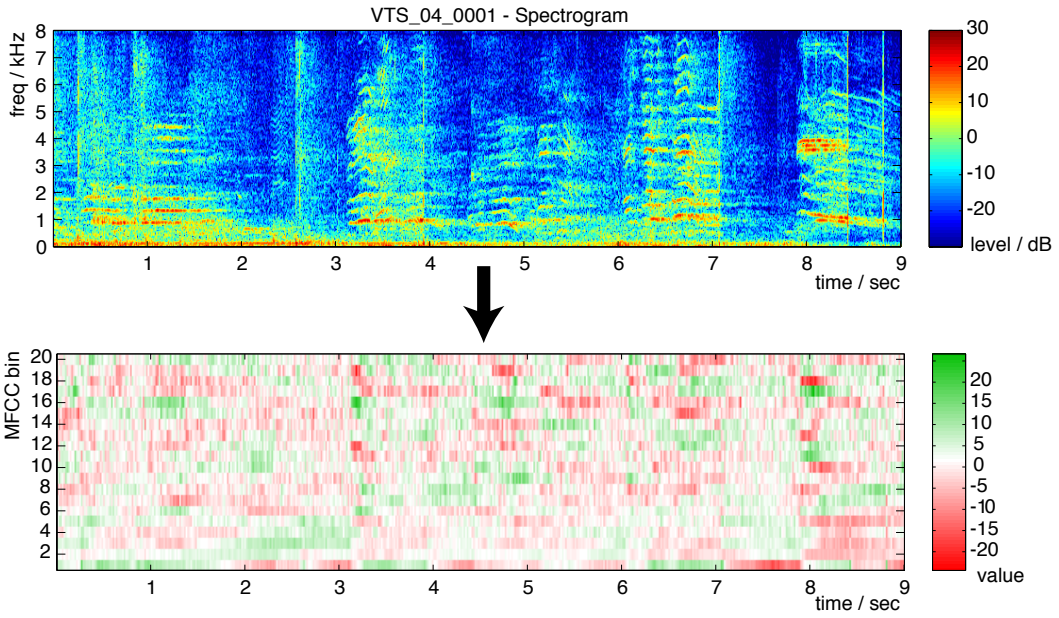
Soundtrack Classification

K. Lee & Ellis 2010

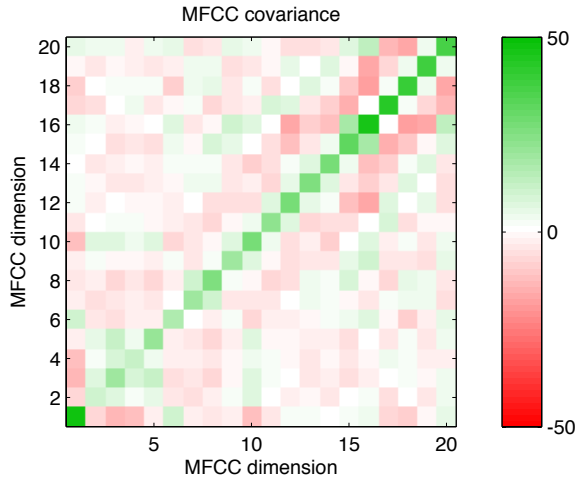
- **Baseline soundtrack classification system:**
  - divide sound into short frames (e.g. 30 ms)
  - calculate MFCC features for each frame
  - describe clip by statistics of frames (mean, covariance)
  - = "bag of features"

Video Soundtrack

MFCC features



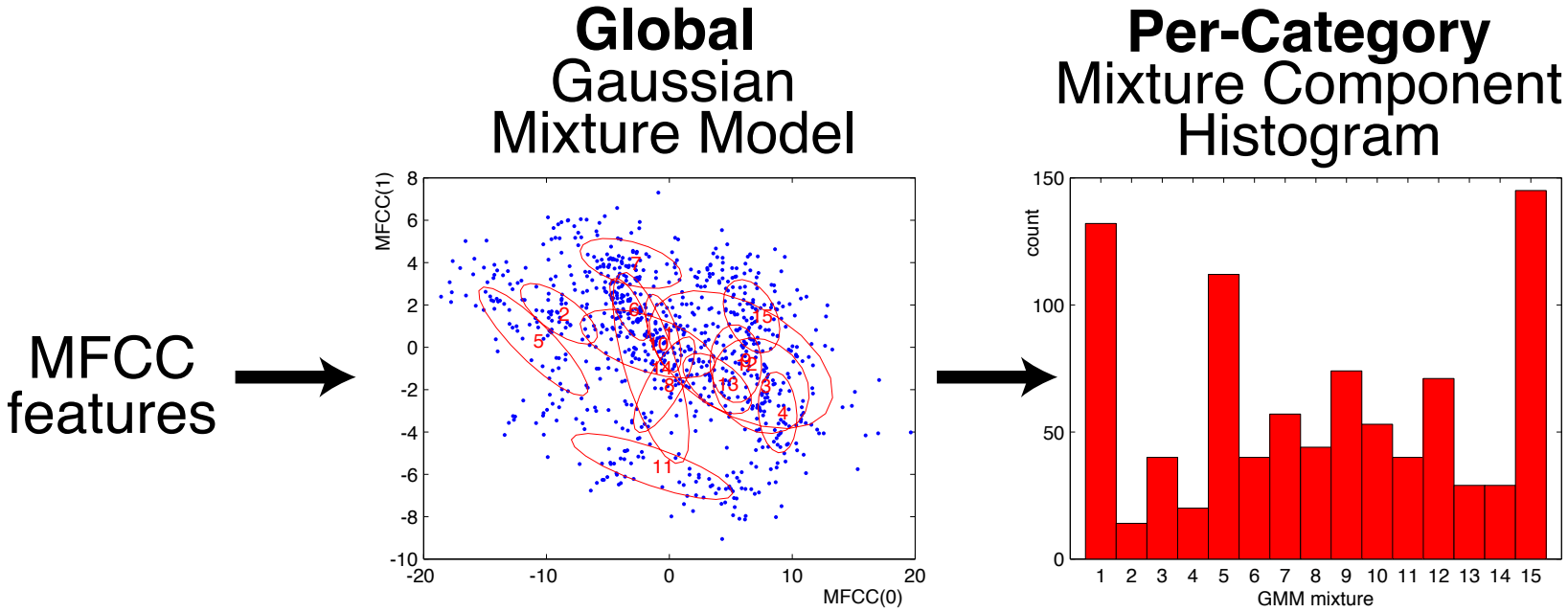
MFCC Covariance Matrix



- Classify by e.g. Mahalanobis distance + SVM

Codebook Histograms

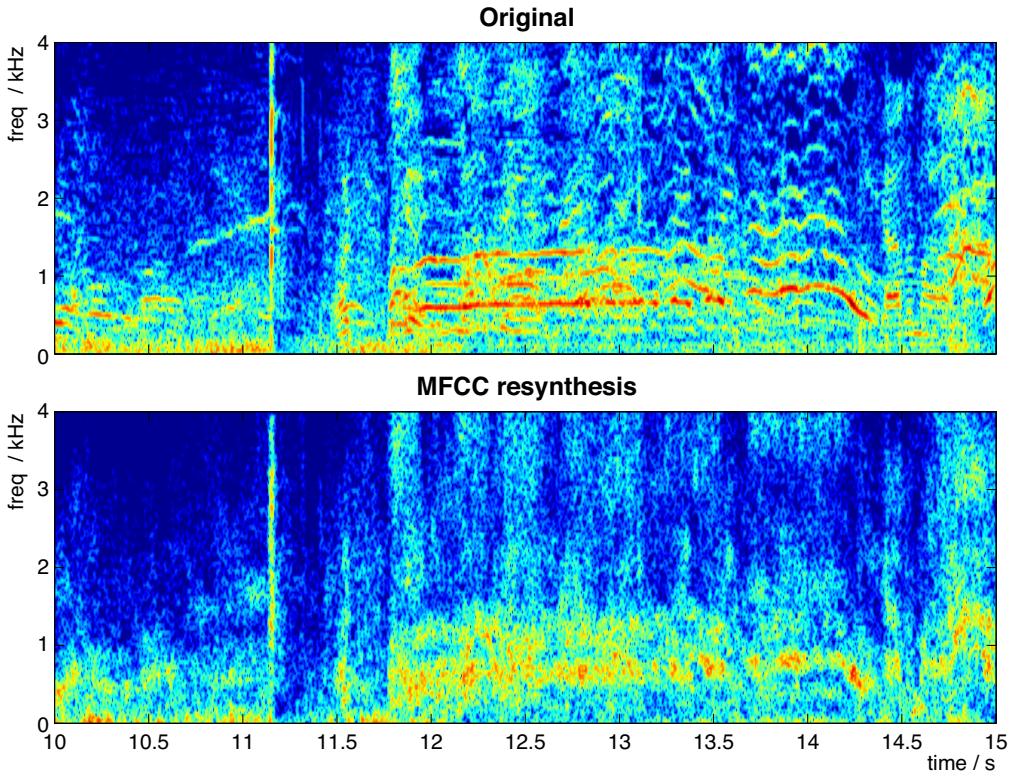
- Instead of Gaussian model, convert high-dim. distributions to multinomial
  - Vector Quantization (VQ)



- Classify by distance on histograms
  - KL, Chi-squared
  - + SVM

Baseline Limitations

- MFCCs model broad spectral shape but strip away fine detail
  - transients, pitch, texture

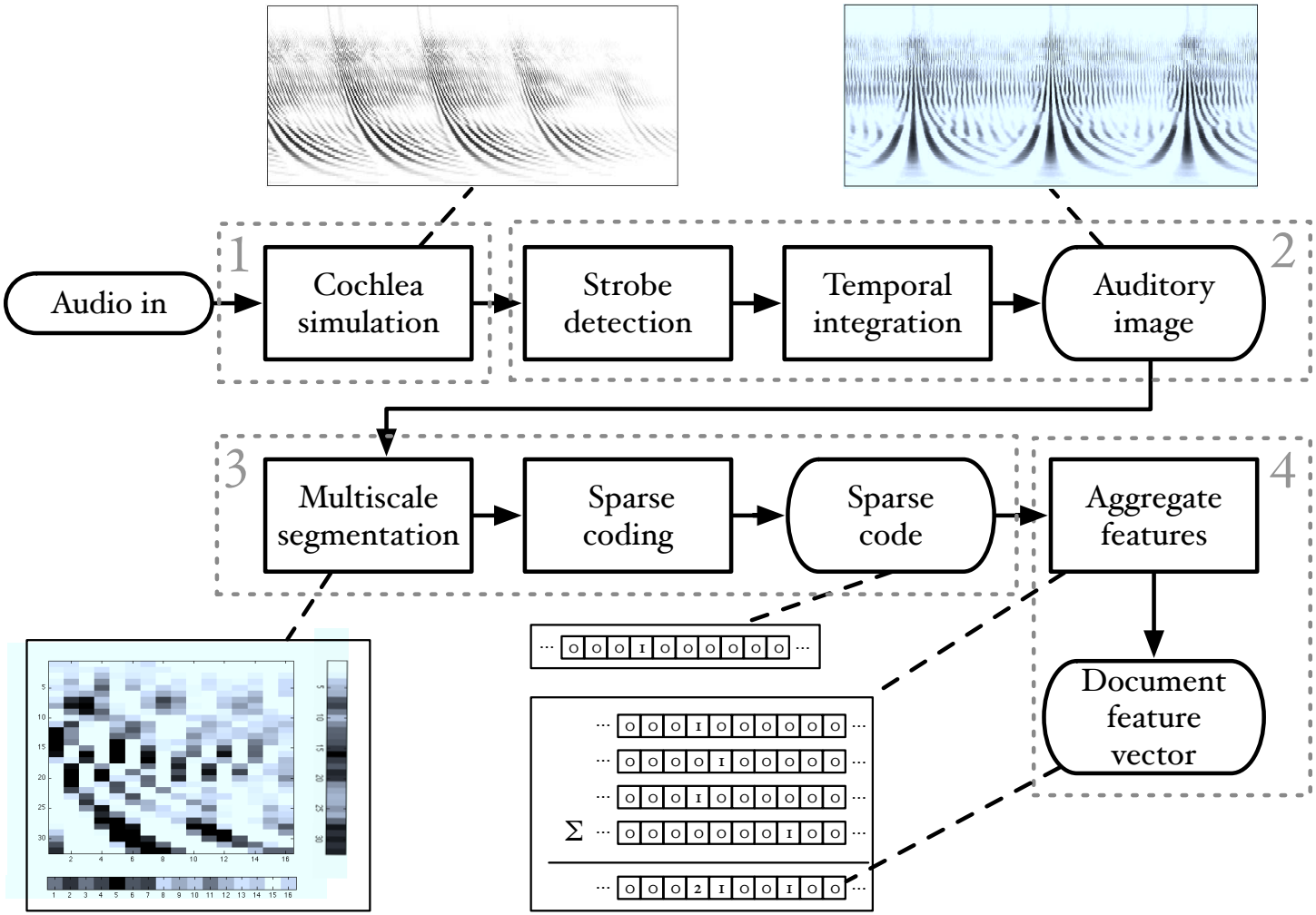


- BoF loses temporal structure...
- Channel & Noise

2. Auditory Model Features

Lyon et al. 2010

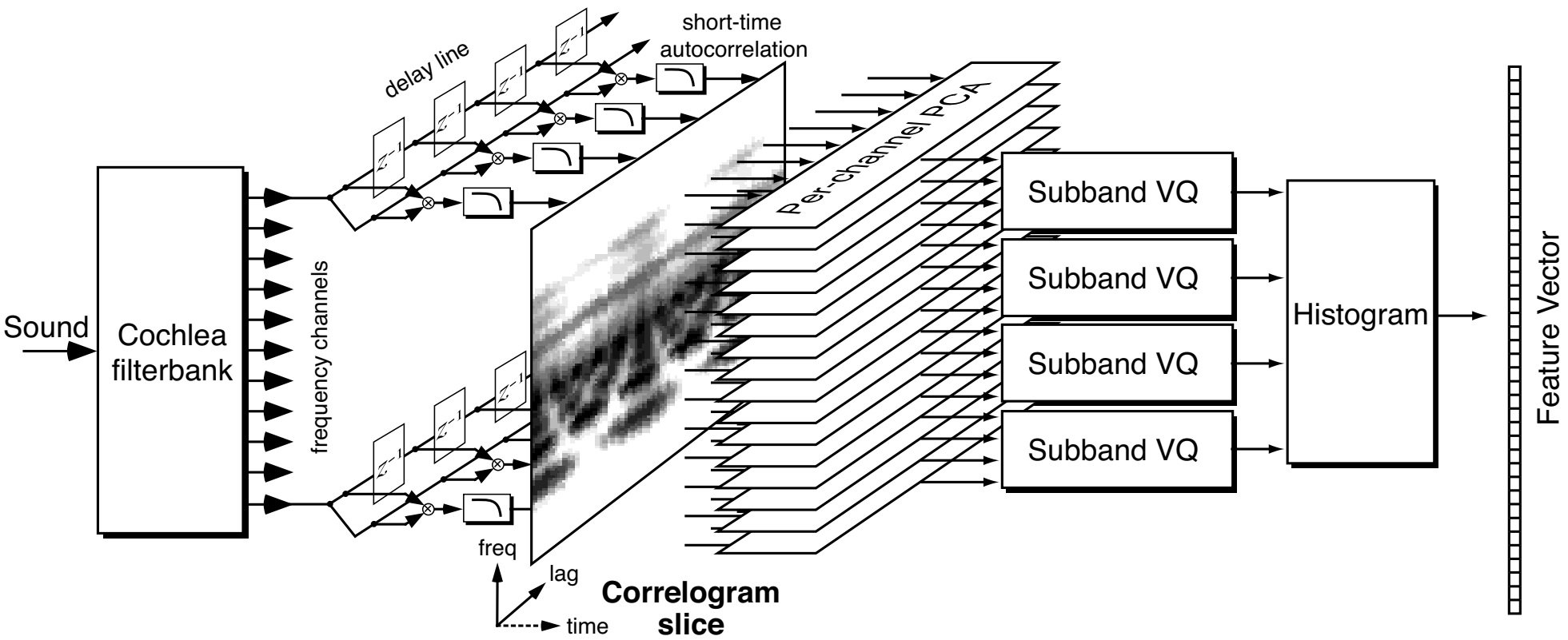
- Lyon's Stabilized Auditory Image (SAI) model
  - fine structure stabilized by 'strobing' on transients



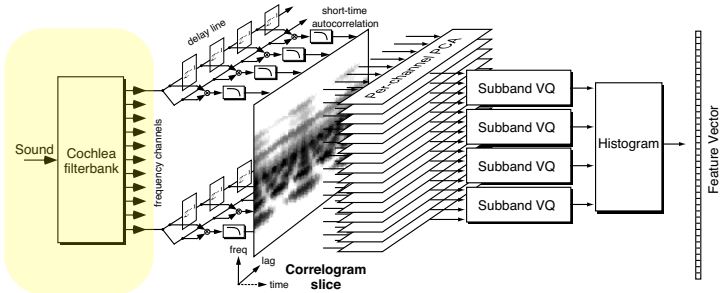
Subband Autocorrelation (SBPCA)

B-S Lee & Ellis 2012

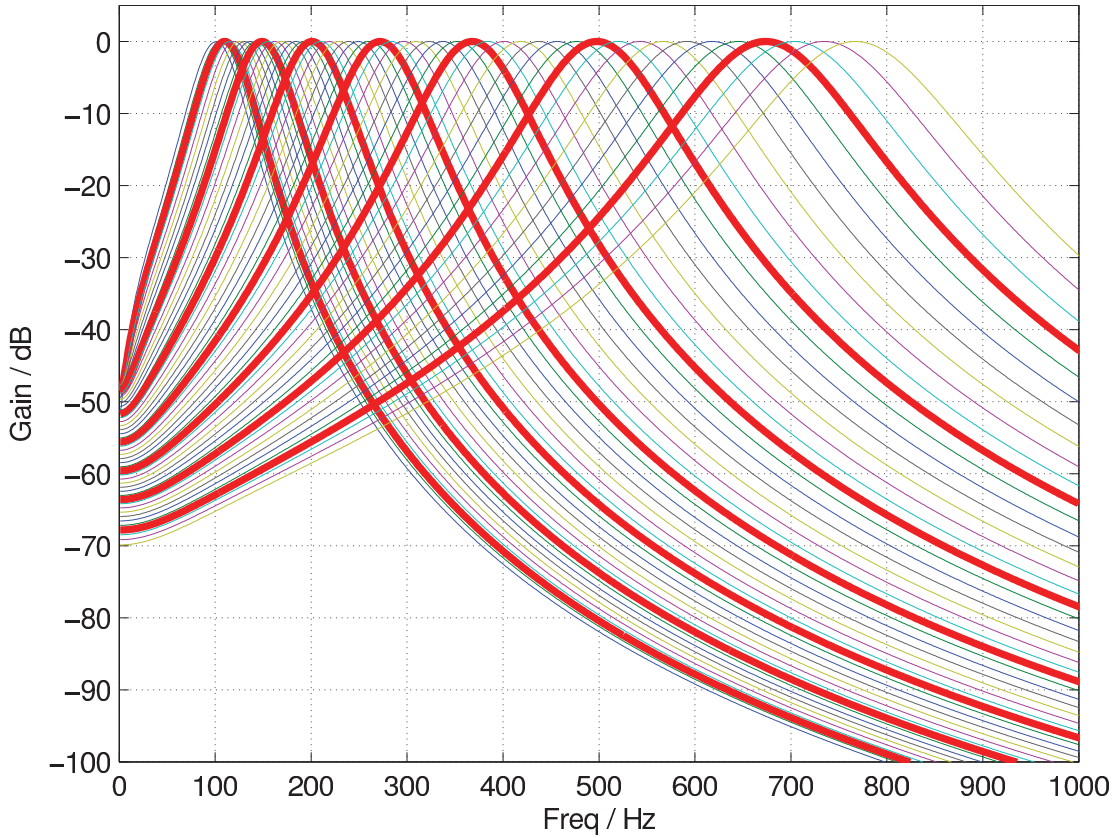
- Simplified version of Lyon model
  - 10x faster (RTx5 → RT/2)
- Captures fine time structure in multiple bands
  - .. the information that is lost in MFCC features



Filterbank

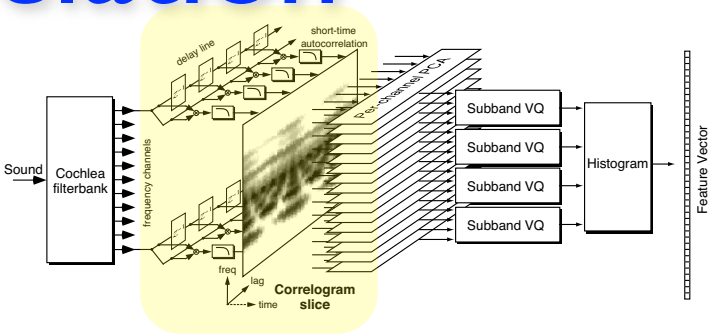


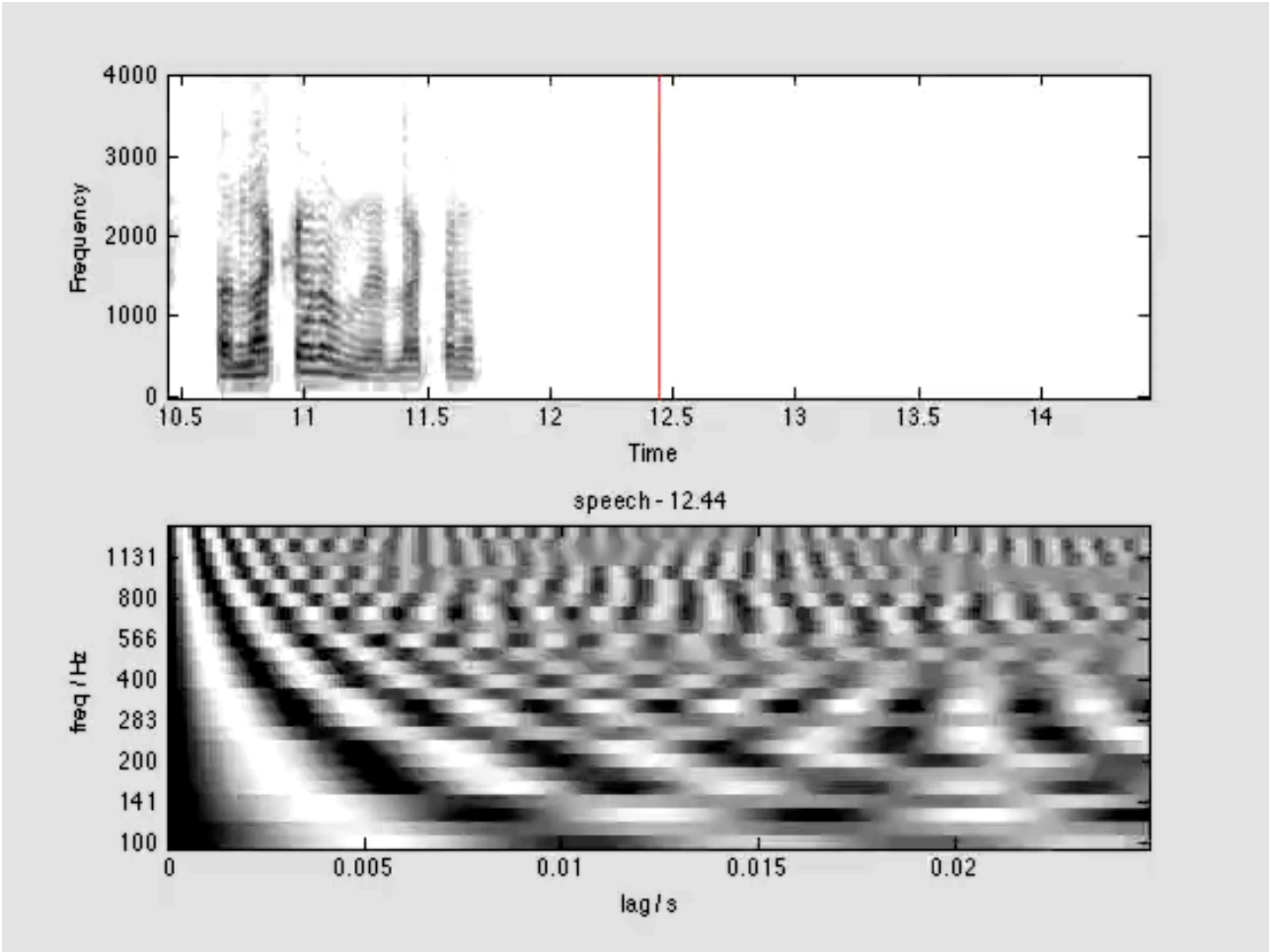
- Simple four-pole, two-zero bandpass filters
- Constant-Q, log-spacing
  - very rough approximation to cochlea



Subband Autocorrelation

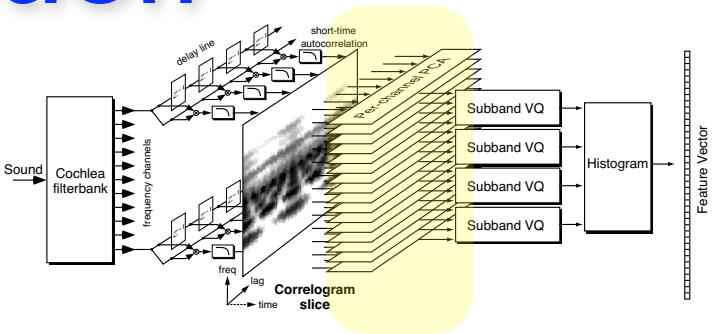
- Autocorrelation stabilizes fine time structure





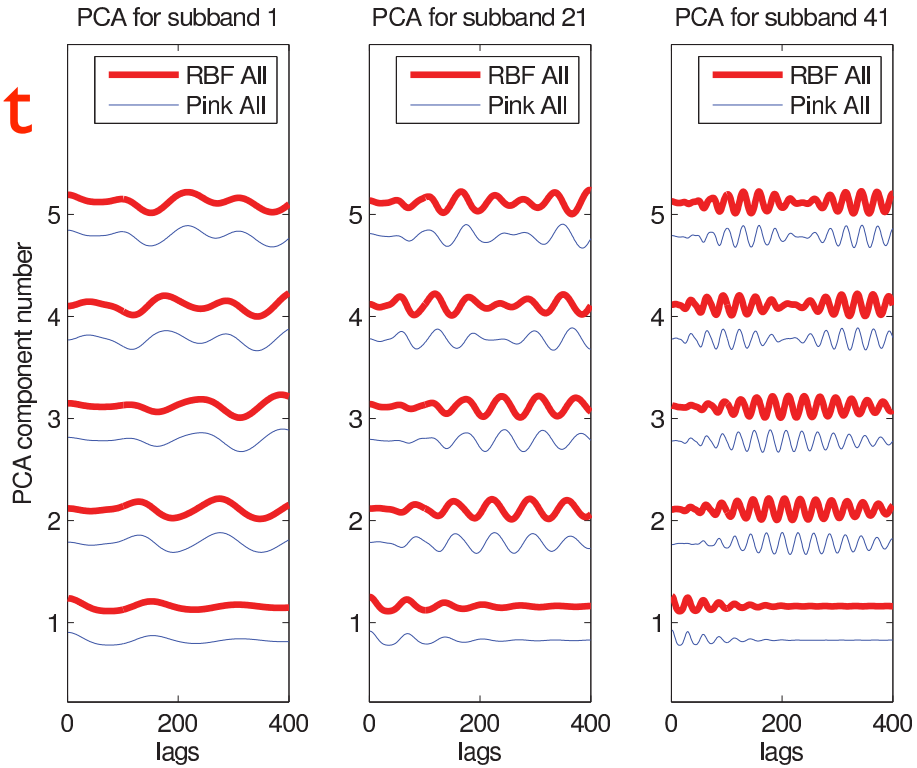
- 25 ms window, lags up to 25 ms
- calculated every 10 ms
- normalized to max (zero lag)

PCA Summarization

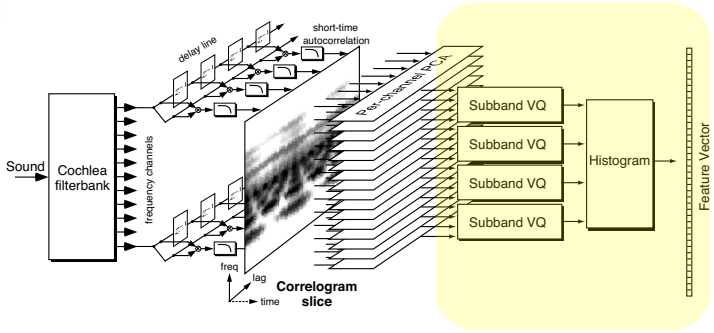


- **Autocorrelation of bandpass signals is constrained**
  - full image = 24 bands x 200 lags
  - but intrinsic information is much lower

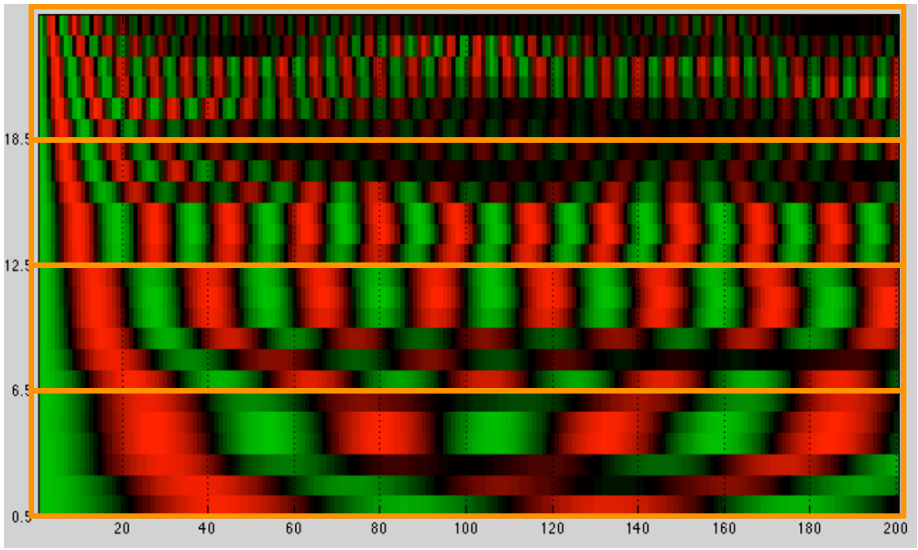
- **Reduce dimensionality with Principal Component Analysis (PCA)**
  - calculated on individual bands to retain separability
  - per-subband PCA bases have little dependence on training material
  - 10 dimensions adequate



Subband VQ



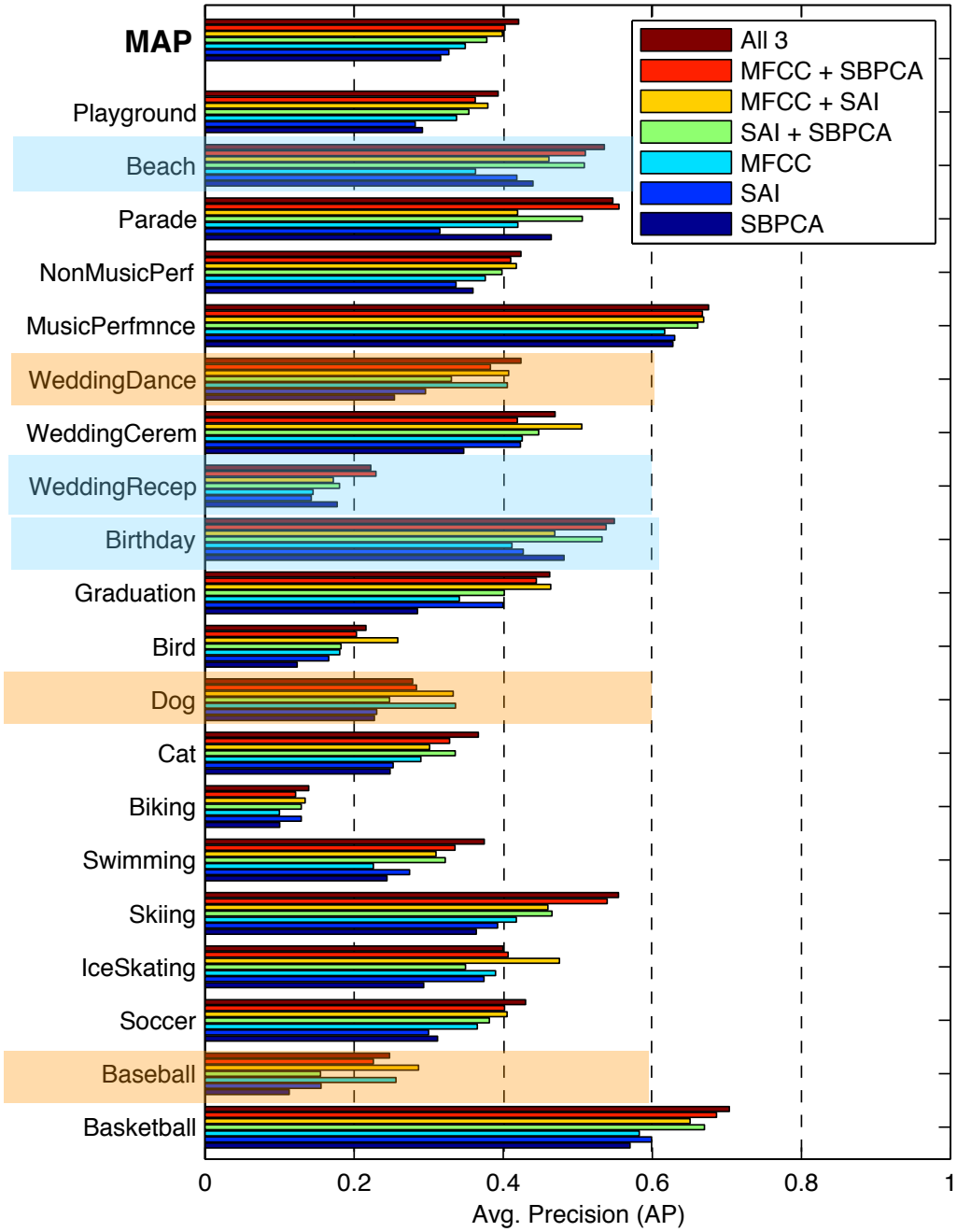
- Summarize features across segments by VQ histogram
- 4 separate bands to provide overlap resistance
  - non-overlapping groups of 6 subbands x 10 PCA coefs



- Each band quantized into 1000 codewords; whole soundtrack → codeword histogram
  - 4000 dimensions, sparse
- SVM classifier with Chi^2 kernel

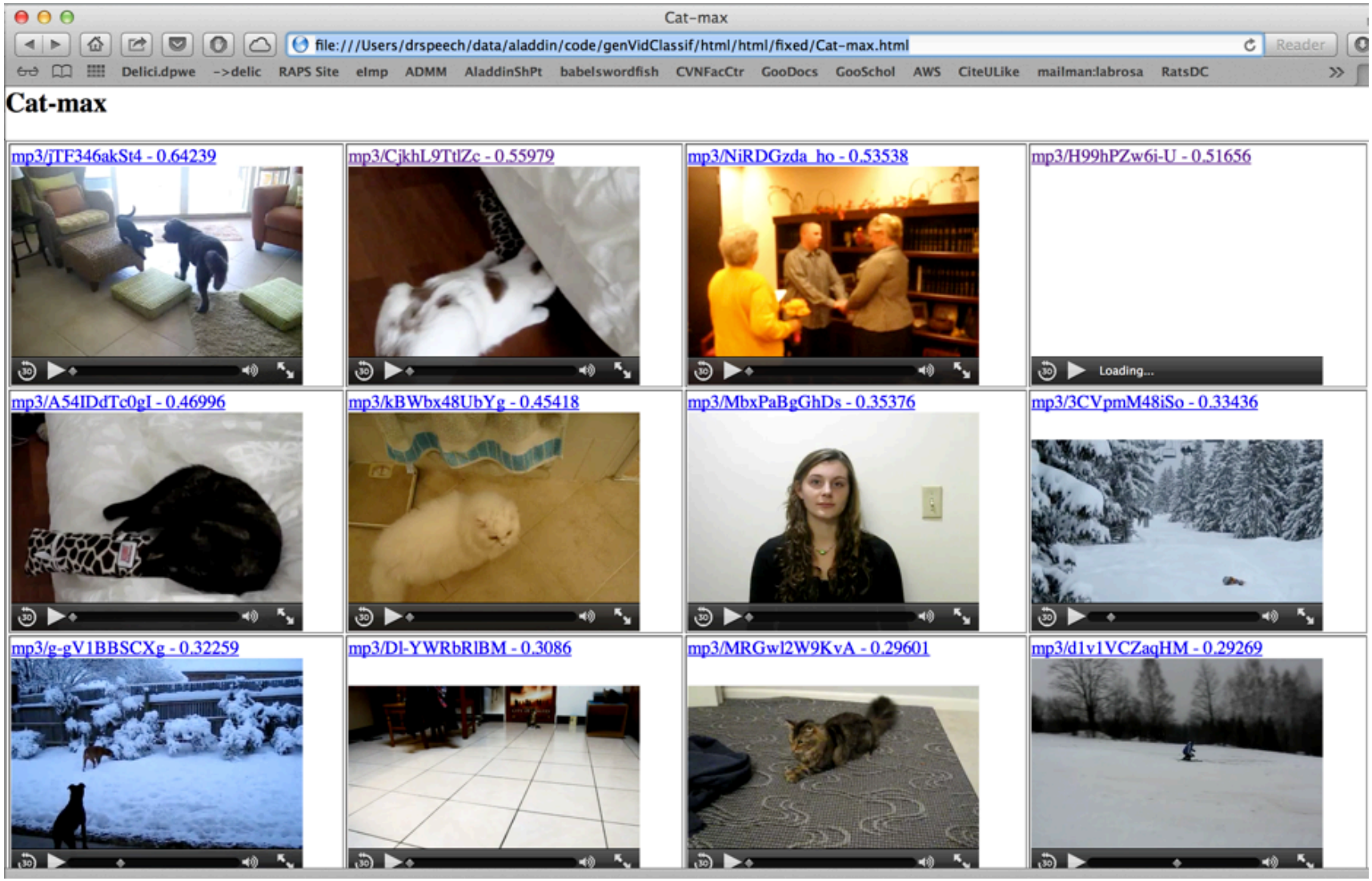
Auditory Model Feature Results

- SAI and SBPCA close to MFCC baseline
- Fusing MFCC and SBPCA improves mAP by 15% rel
  - mAP: 0.35 → 0.40
- Calculation time
  - MFCC: 6 hours
  - SAI: 1087 hours
  - SBPCA: 110 hours



Retrieval Examples

- Browsing results are sometimes surprising!

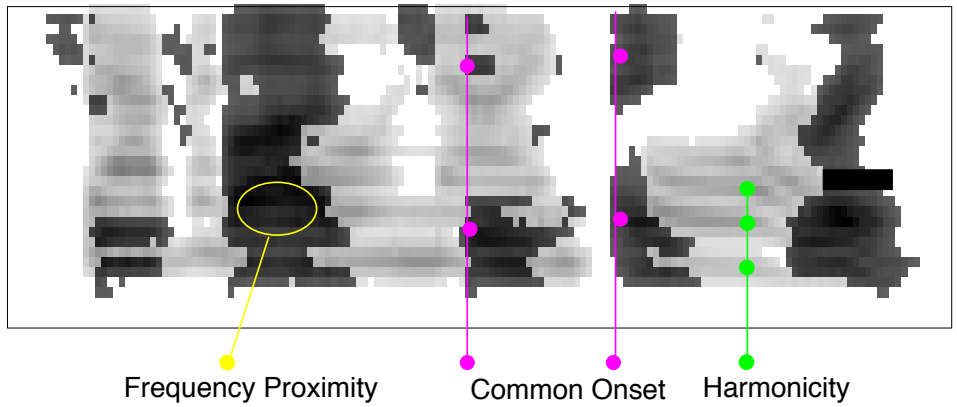


file:///u/drspeech/data/aladdin/code/genVidClassif/html/mfcc230/Cat-max.html

3. Open Issues

- Foreground & Events
  - transients at a coarser scale

- Better object/event separation
  - parametric models
  - spatial information?
  - computational auditory scene analysis...



Barker et al. '05

- Better Annotation
  - granularity in time & concept

Summary

- Soundtrack classification
  Acoustic properties of different events

- Standard model
  MFCC + bag-of-features + classifier

- Auditory model features
  to recapture the fine time structure
  - combination improves retrieval

code available:
http://labrosa.ee.columbia.edu/projects/calcSBPCA/

References

- Jon Barker, Martin Cooke, & Dan Ellis, "Decoding Speech in the Presence of Other Sources," *Speech Communication* 45(1): 5-25, 2005.
- Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A.C. Loui, "Consumer video understanding: A benchmark database and an evaluation of human and machine per- formance," in Proc. ACM International Conference on Multimedia Retrieval (ICMR), Apr. 2011, p. 29.
- Byung-Suk Lee & Dan Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in Proc. INTERSPEECH-12, Sept. 2012.
- Keansub Lee & Dan Ellis, "Audio-Based Semantic Concept Classification for Consumer Video," *IEEE Tr. Audio, Speech and Lang. Proc.* 18(6): 1406-1416, Aug. 2010.
- R.F. Lyon, M. Rehn, S. Bengio, T.C. Walters, and G. Chechik, "Sound retrieval and ranking using sparse auditory representations," Neural Computation, vol. 22, no. 9, pp. 2390–2416, Sept. 2010.

Acknowledgment

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20070. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon.

Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.