Recognizing and Classifying Environmental Sounds

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. Environmental Sound Recognition
2. Foreground Events
3. Background Retrieval
4. Labels & Annotation
5. Future Directions

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

Columbia University in the City of New York

1. What is hearing for?

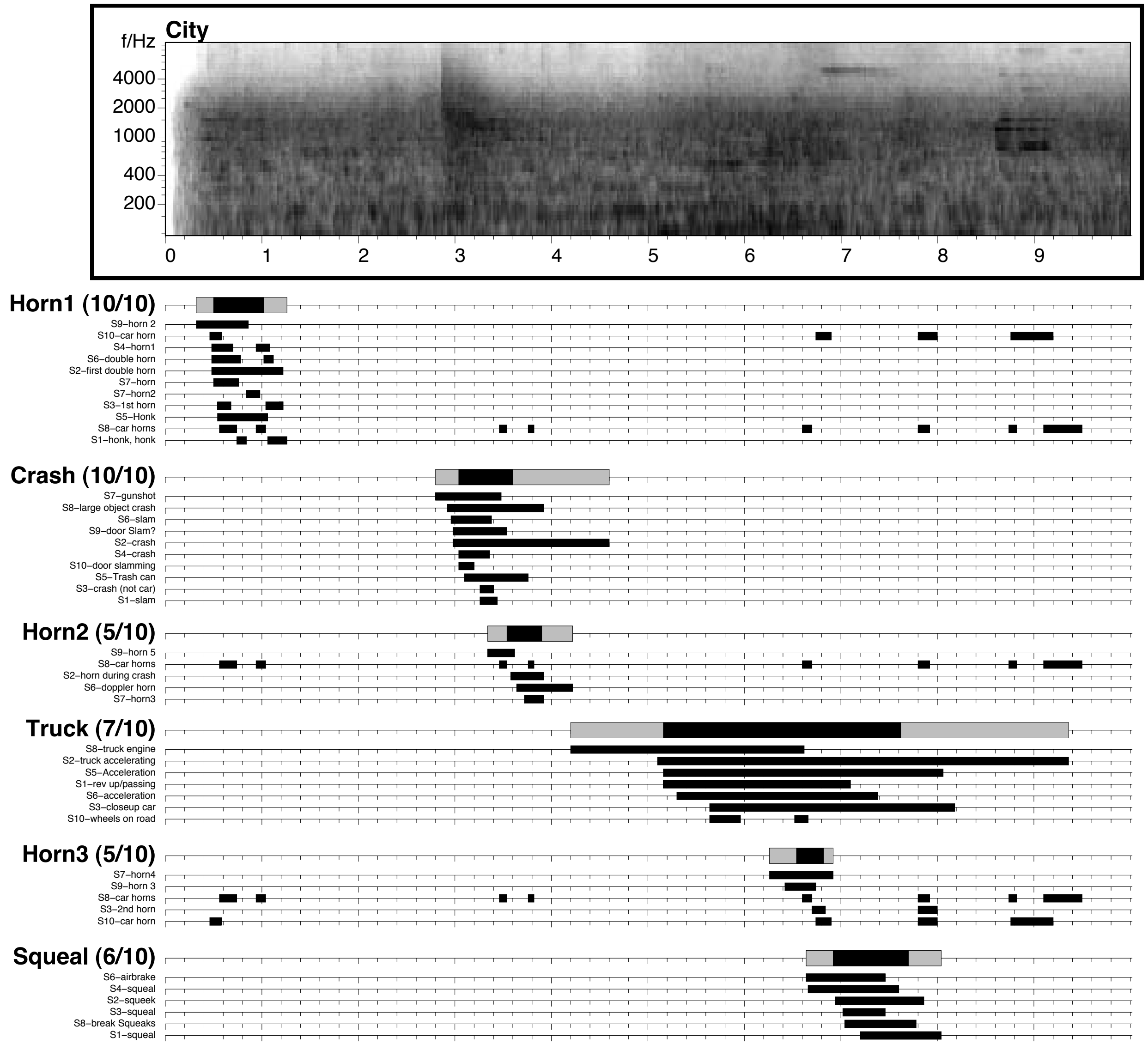
http://news.bbc.co.uk/2/hi/science/nature/7130484.stm

- Hearing = getting information from sound
  - predators/prey
  - communication
- Environmental sound recognition is fundamental

Environmental Sound Perception

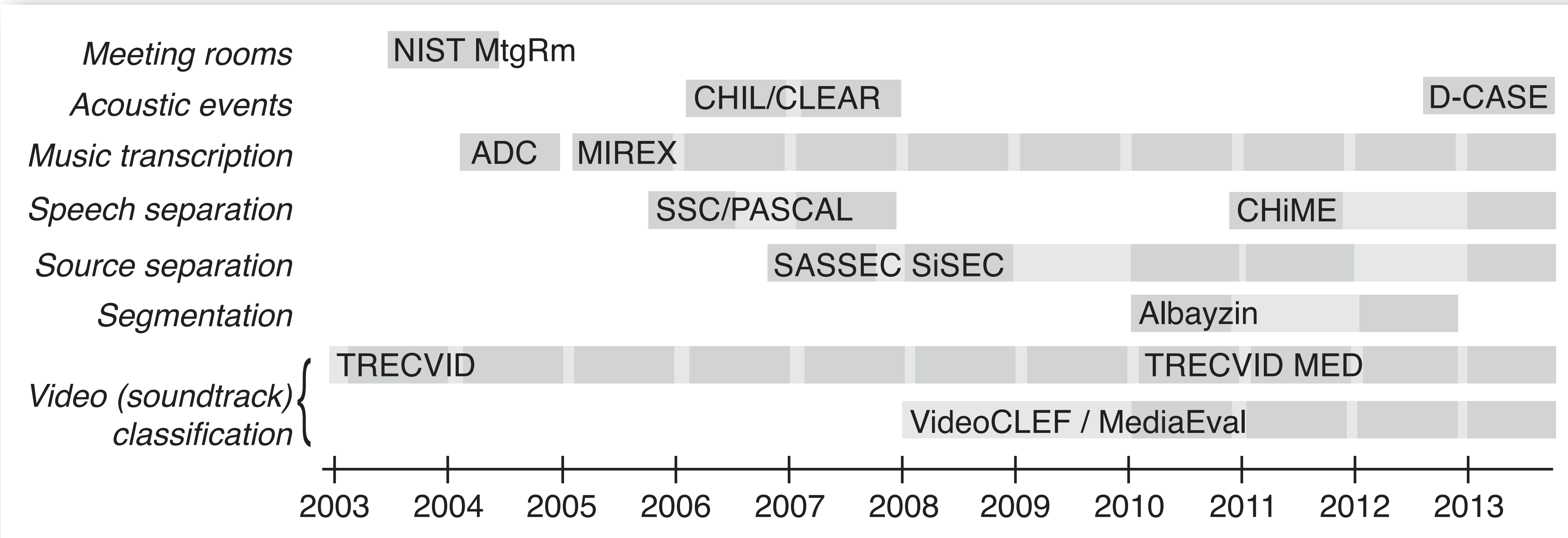
Ellis 1996

- What do people hear?
  - sources
  - ambience
- Mixtures are the rule



Sound Scene Evaluations

- Evaluations are good for research
  - help researchers, help funders
- A decade of evaluations:



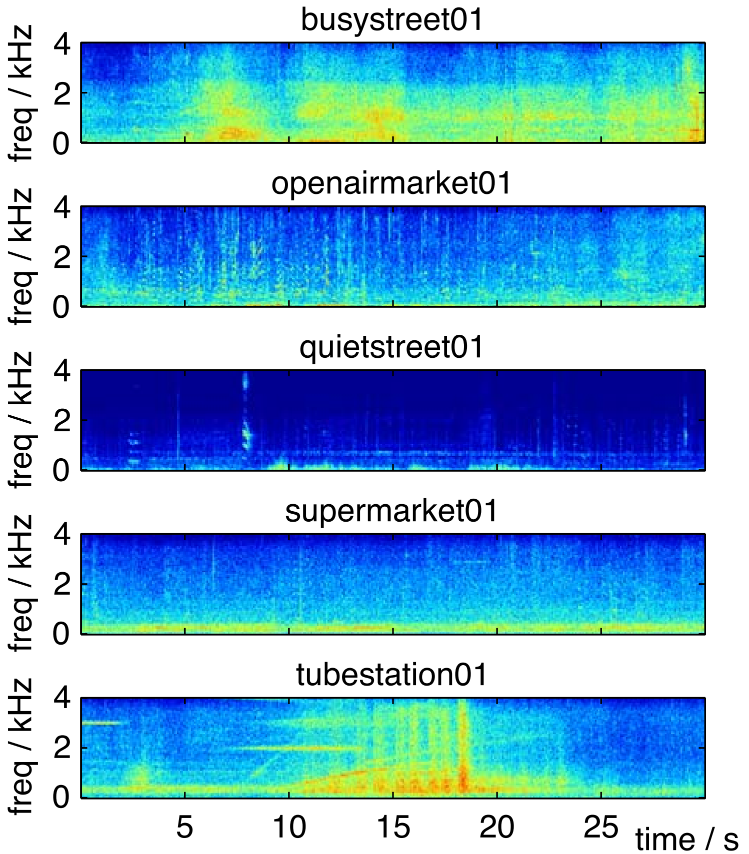
  - Metrics: SNR, Frame Acc, Event Error Rate, mAP

DCASE (2012-2013)

Giannoulis et al. 2013

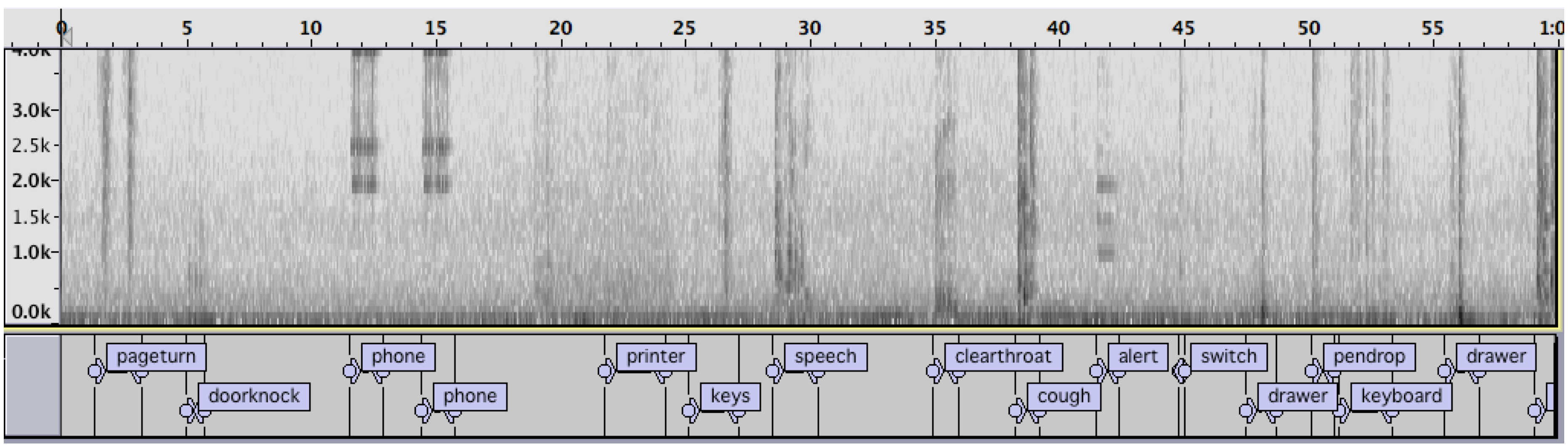
- 2012 IEEE AASP Challenge:
  "Detection and Classification of Acoustic Scenes and Events"
  - Systems submitted Mar 2013
  - Results at WASPAA, Oct 2013
  - 2 tasks...

- Task 1: Scene classification
  - 10 classes x 10 examples x 30 s
    - street, supermarket, restaurant, office, park ...
  - evaluate on 100 examples
    (~1 hour total)
    by classification accuracy



DCASE Event Detection

- Task 2: Event detection ("office live")
  - 16 events x 20 training examples (~ 20 min total)
    knock, laugh, drawer, keys, phone ...
  - evaluate on ~15 min (?)
  - metrics: frame-level AEER
    & event-level precision-recall



TRECVID MED (2010-2013)

Over et al. 2011

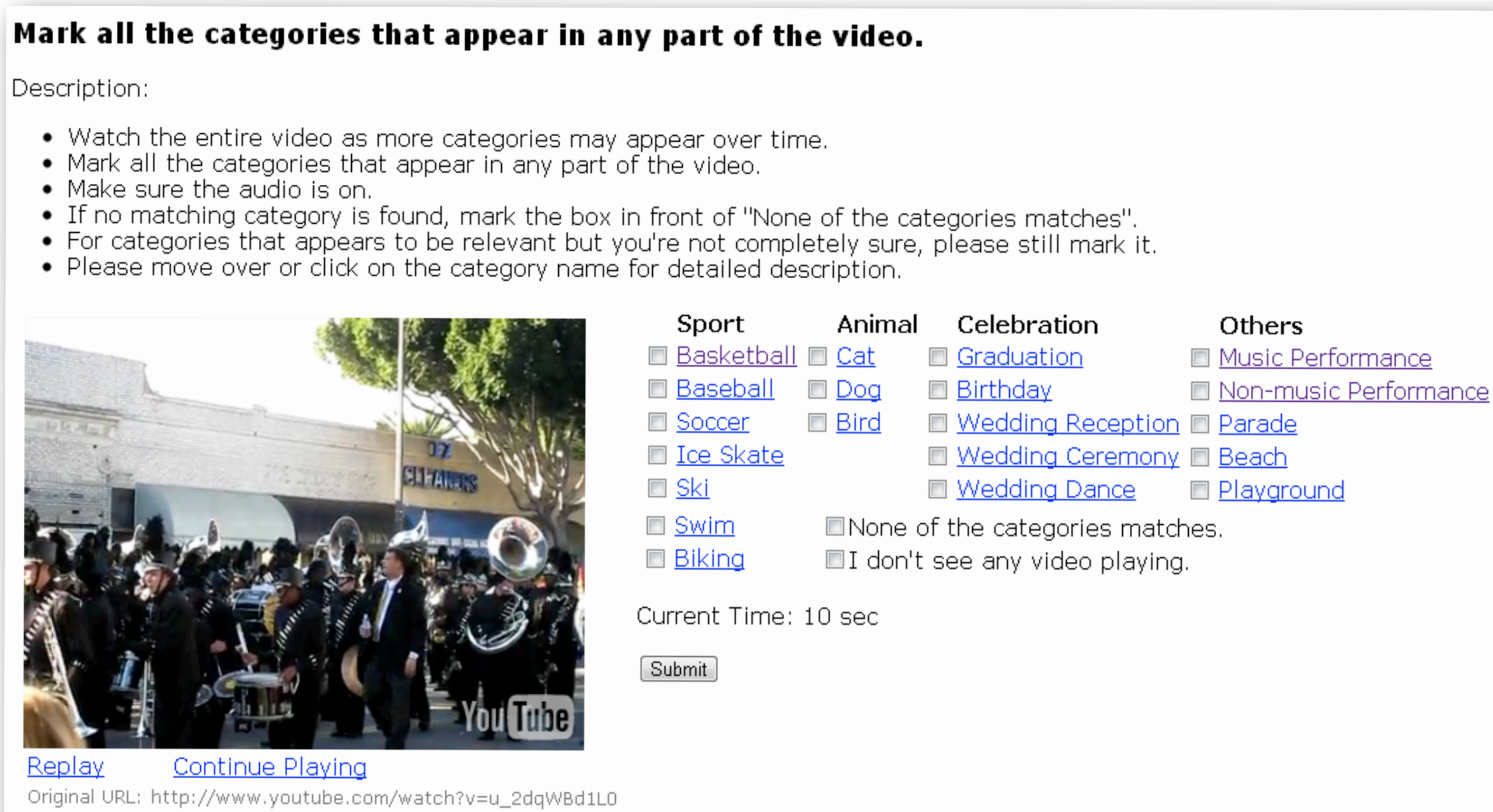
- "Multimedia Event Detection"
  - e.g. MED2011: 15 events x 200 example videos (~60s)
    - Making a sandwich, Birthday party, Parade, Flash mob
  - evaluate by mean Average Precision over 10k-100k videos (200-2,000 hours)
  - audio and video ...
    - participants have annotated ~1000 videos (> 10 h)

E009 Getting a Vehicle Unstuck

Consumer Video Dataset

Y-G. Jiang et al. 2011

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


Mark all the categories that appear in any part of the video.

Description:

- Watch the entire video as more categories may appear over time.
- Mark all the categories that appear in any part of the video.
- Make sure the audio is on.
- If no matching category is found, mark the box in front of "None of the categories matches".
- For categories that appears to be relevant but you're not completely sure, please still mark it.
- Please move over or click on the category name for detailed description.

Sport	Animal	Celebration	Others
Basketball	Cat	Graduation	Music Performance
Baseball	Dog	Birthday	Non-music Performance
Soccer	Bird	Wedding Reception	Parade
Ice Skate		Wedding Ceremony	Beach
Ski		Wedding Dance	Playground
Swim			
Biking			

None of the categories matches.
I don't see any video playing.

Current Time: 10 sec

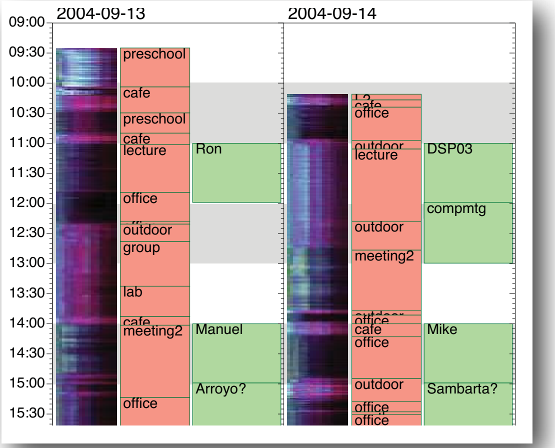
Submit

Replay Continue Playing

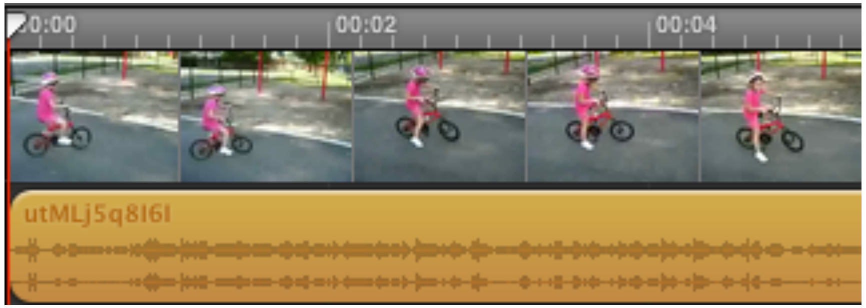
Original URL: http://www.youtube.com/watch?v=u_2dqWBd1L0

Environmental Sound Motivations

- Audio Lifelog
  Diarization



- Consumer Video
  Classification & Search

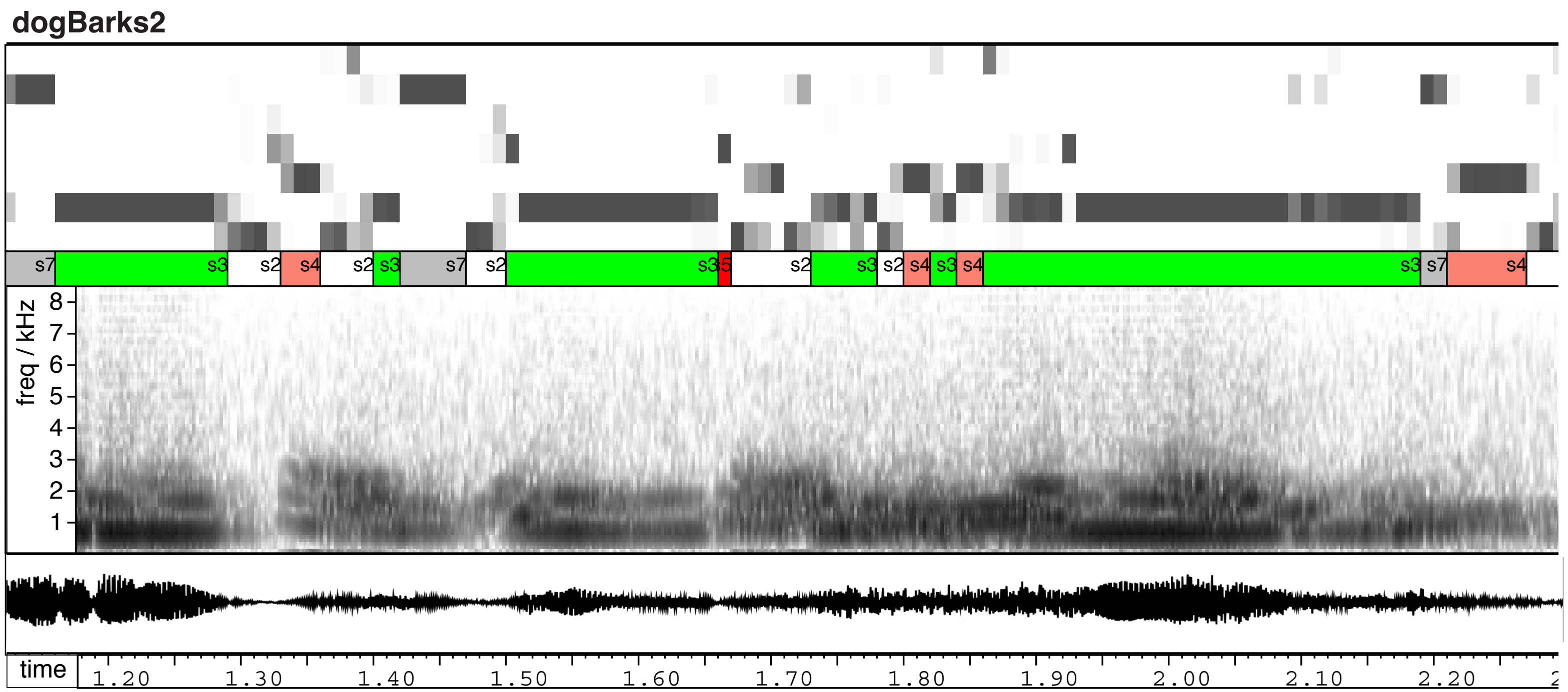


- Real-time hearing prosthesis app
- Robot environment sensitivity
- Understanding hearing

2. Foreground Event Recognition

Reyes-Gomes & Ellis 2003

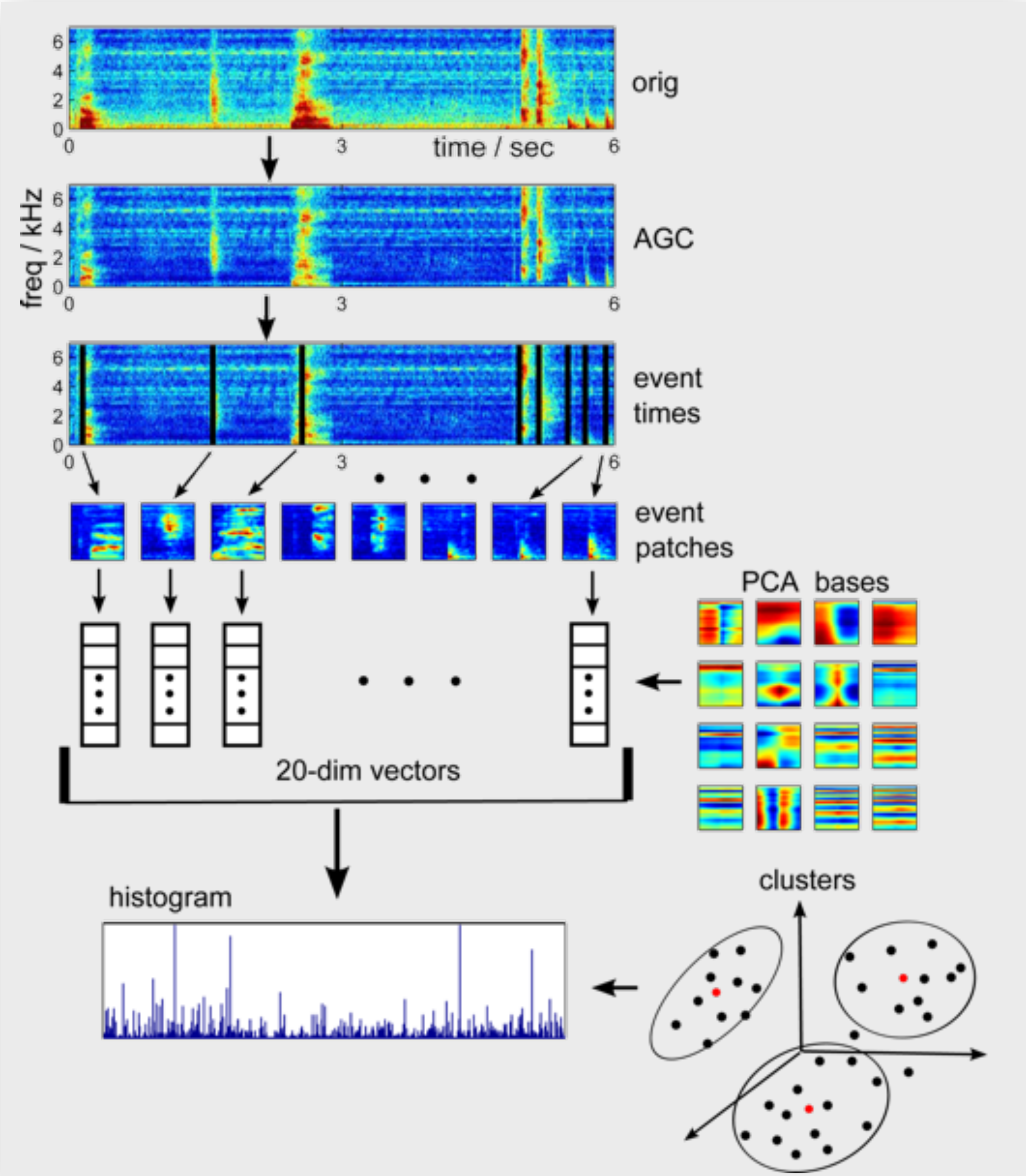
- "Events" are what we hear / notice
- ASR approach?



- events = words? what are subwords?
- need labeled data
- but ... mature tools are great

Transient Features

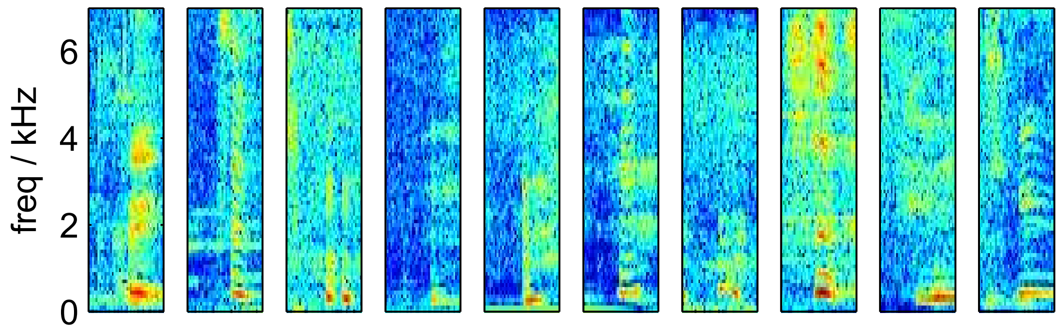
Cotton, Ellis, Loui '11

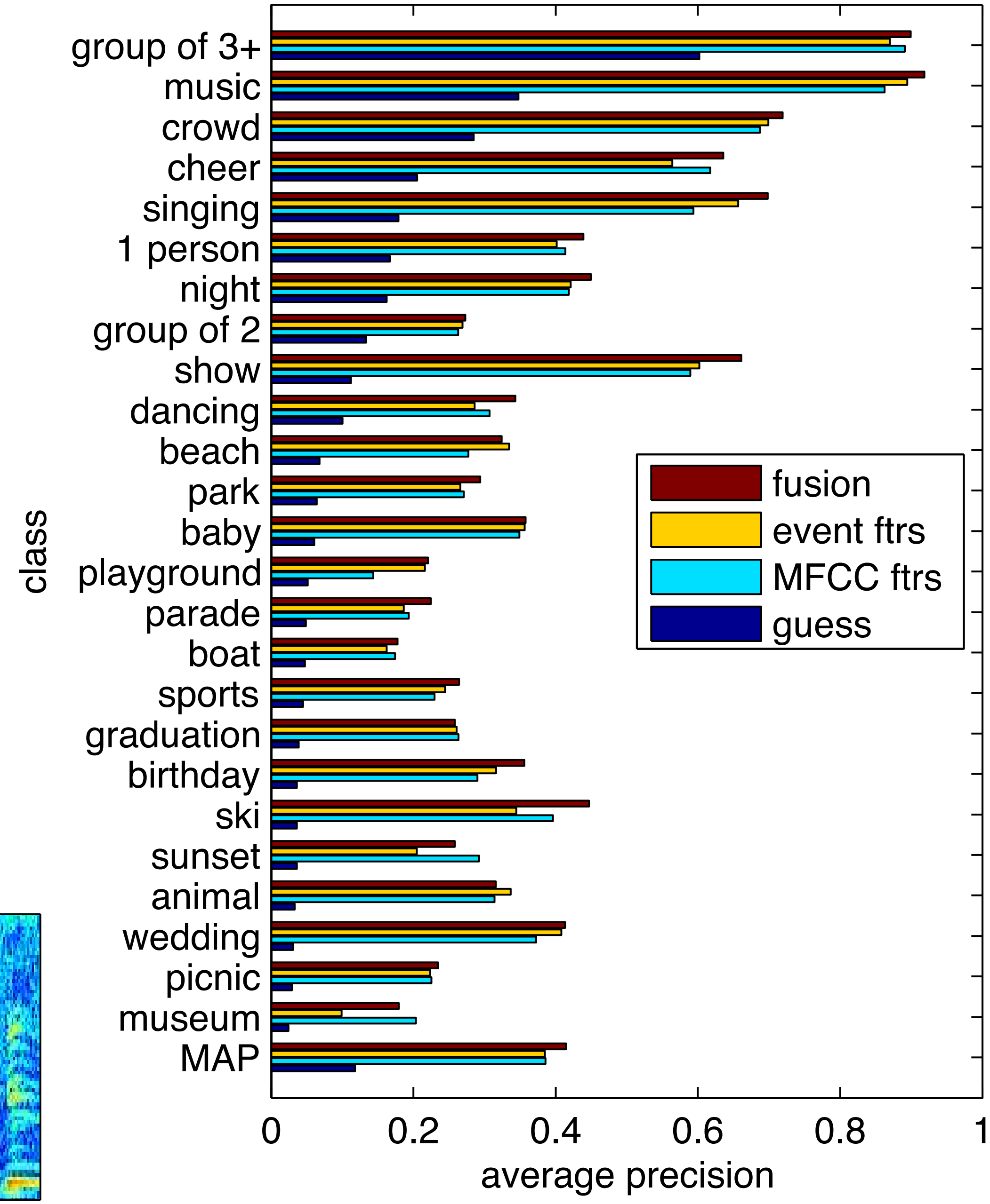


- Transients = foreground events?
- Onset detector finds energy bursts
  - best SNR
- PCA basis to represent each
  - 300 ms x auditory freq
- "bag of transients"

Transient Features

- Results show a small benefit
  - similar to MFCC baseline?
- Examine clusters
  - looking for semantic consistency...
  - link cluster to label





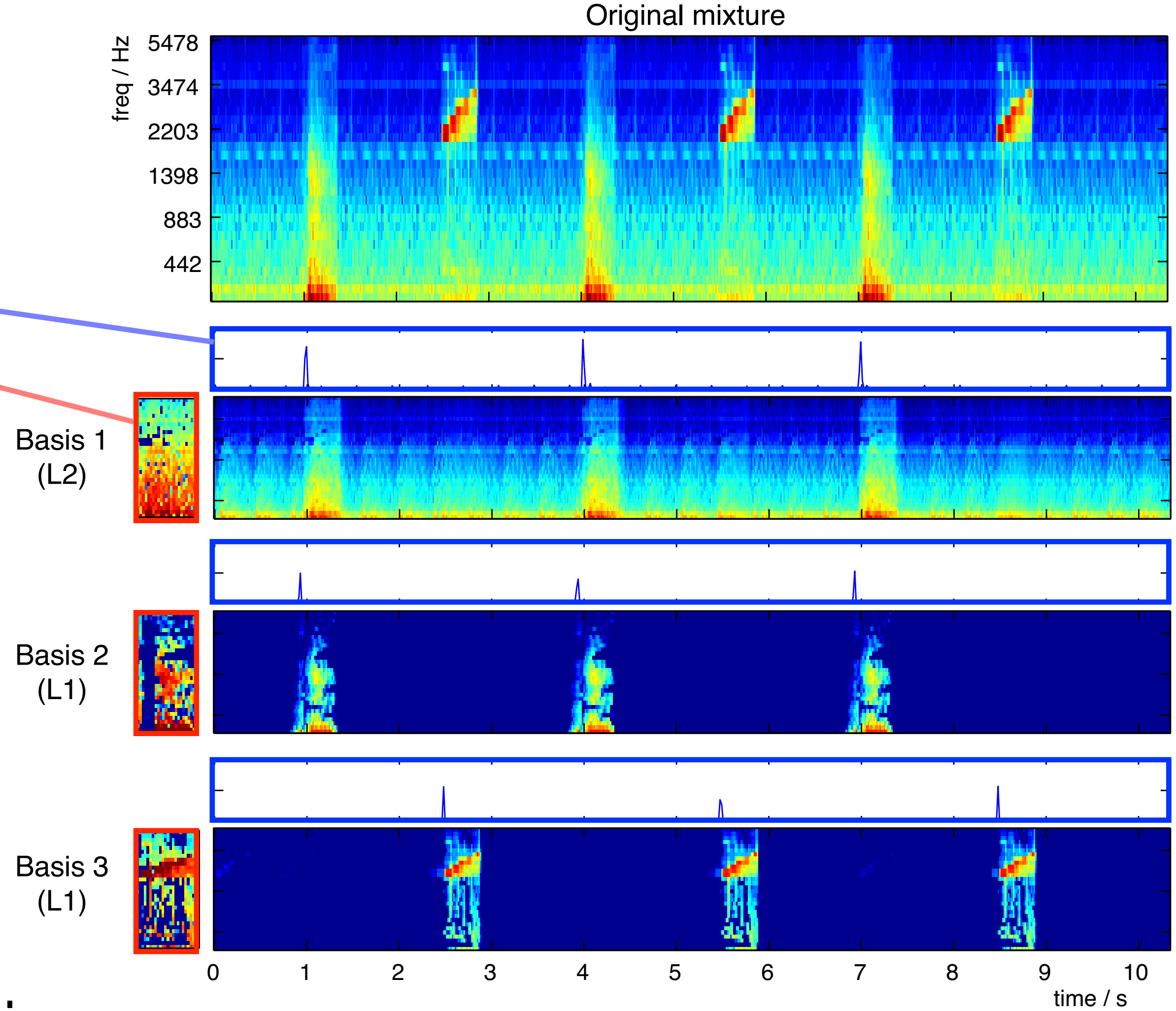
NMF Transient Features

Smaragdis & Brown '03
Abdallah & Plumbley '04
Virtanen '07

- Decompose spectrograms into templates + activation

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

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

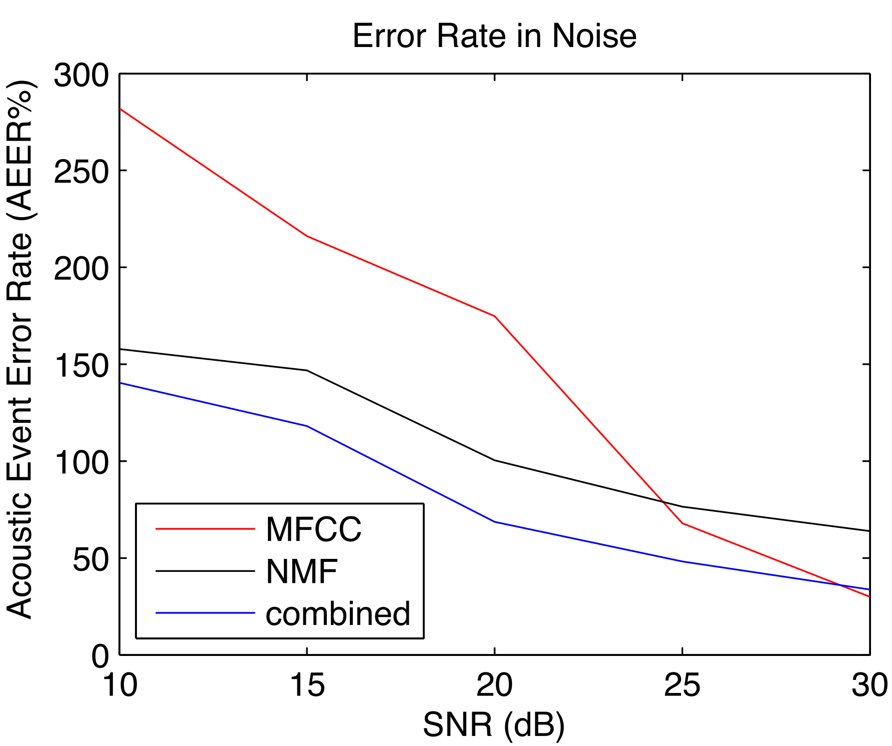


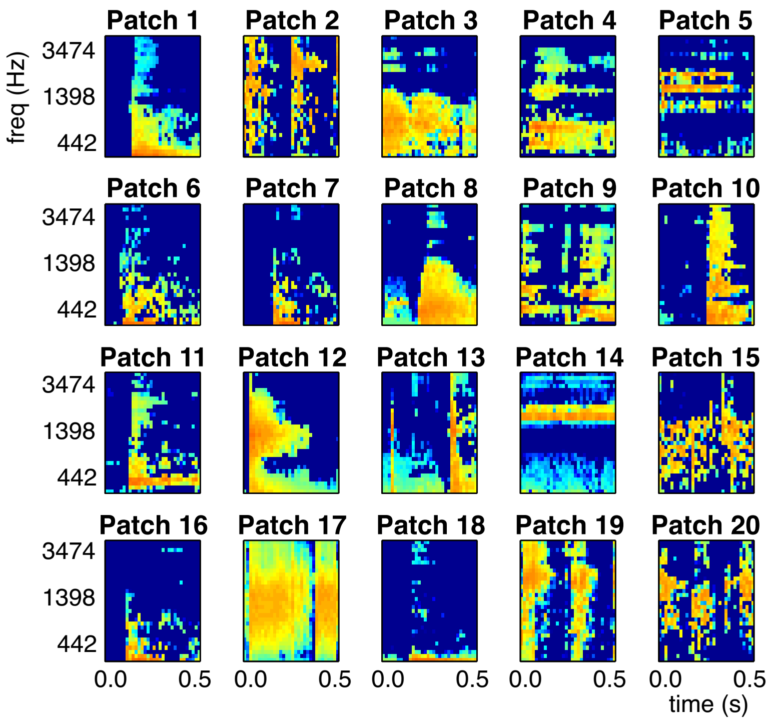
NMF Transient Features

Cotton & Ellis '11

- Learn 20 patches from CLEAR Meeting Room events

- Compare to MFCC-HMM detector

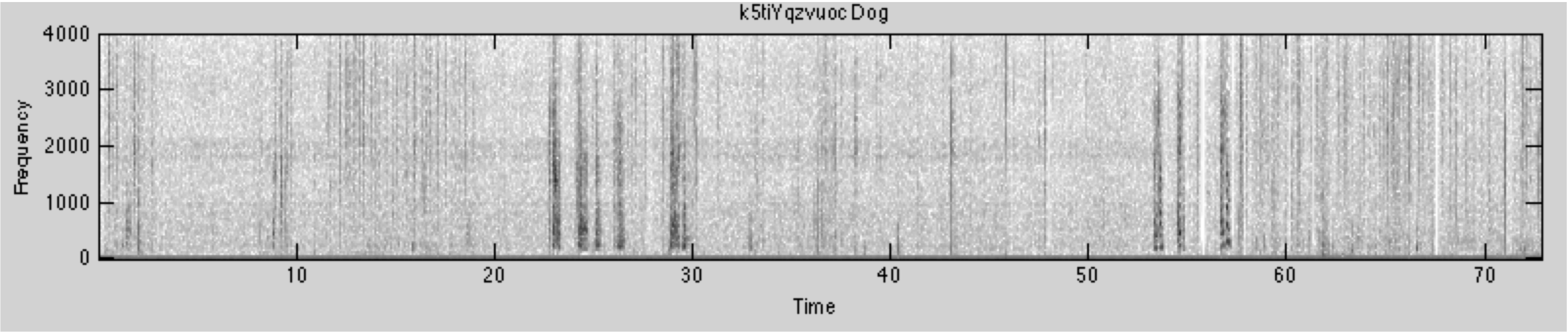




- NMF more noise-robust
  - combines well ...

Why Are Events Hard?

- Events are short

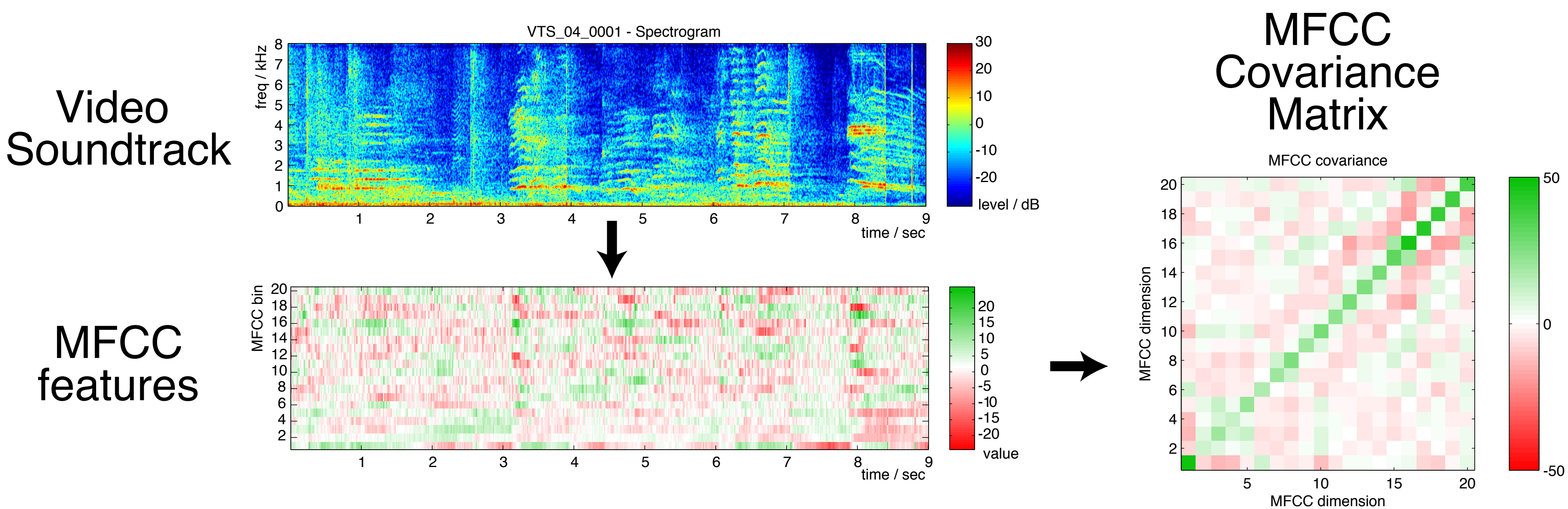


  - target sounds may occupy only a few % of time
- Events are varied
  - what is the vocabulary? what are the prototypes?
  - source & channel variability
- Critical information is in fine-time structure
  - onset transient etc.
  - poor match to classic frame-spectral-envelope features

3. Background Retrieval

K. Lee & Ellis 2010

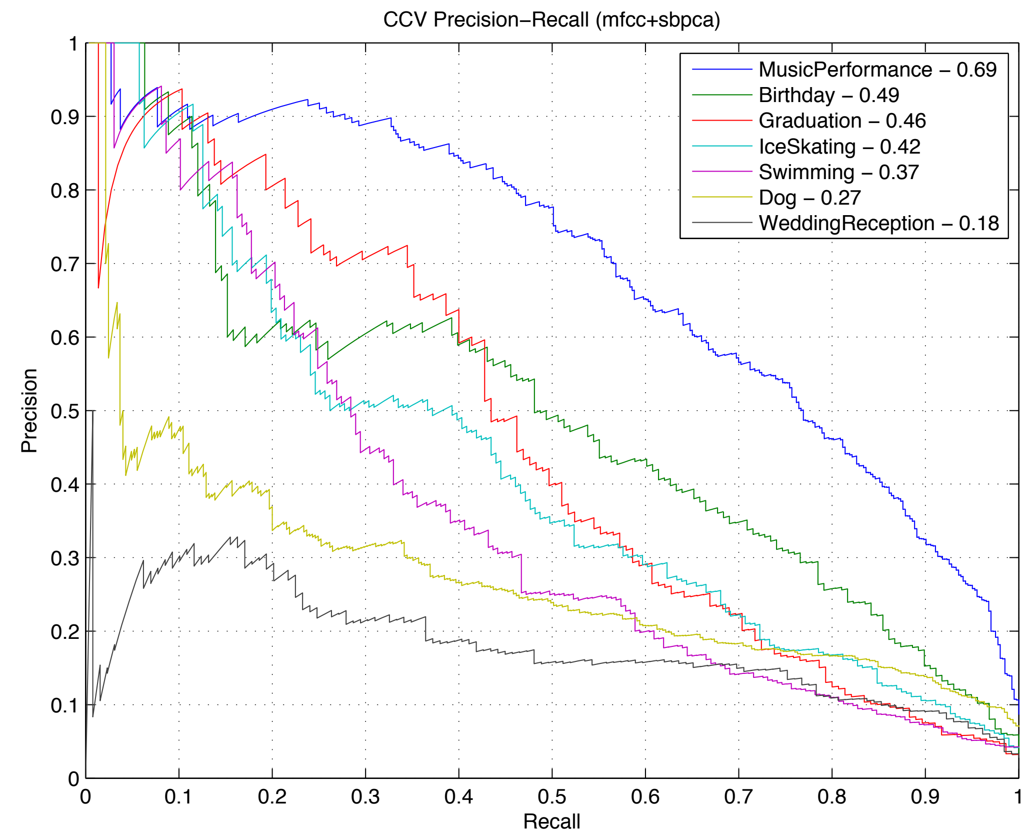
- Baseline for soundtrack classification
  - divide sound into short frames (e.g. 30 ms)
  - calculate features (e.g. MFCC) for each frame
  - describe clip by statistics of frames (mean, covariance)
  - = "bag of features"

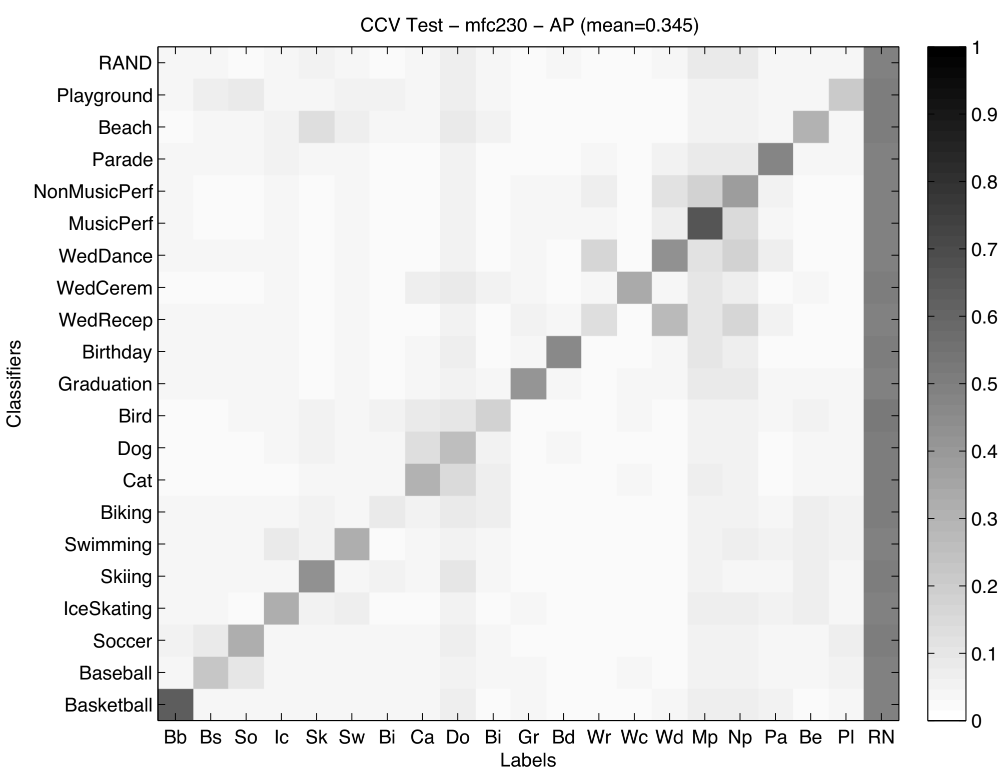


- Classify by e.g. Mahalanobis distance + SVM

Retrieval Evaluation

- Rank large test set by match to category
- Precision-Recall

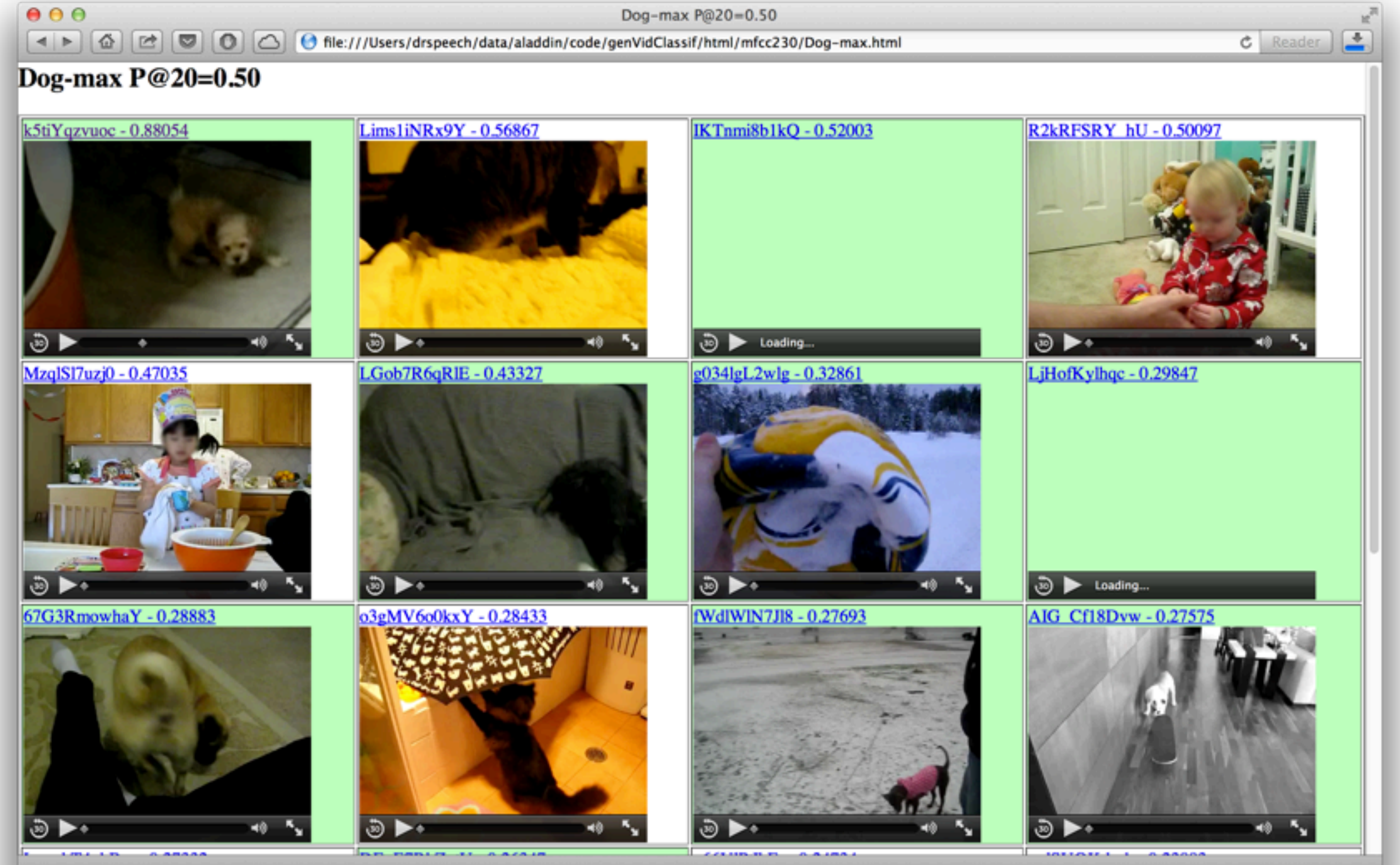




- mean Average Precision

Retrieval Examples

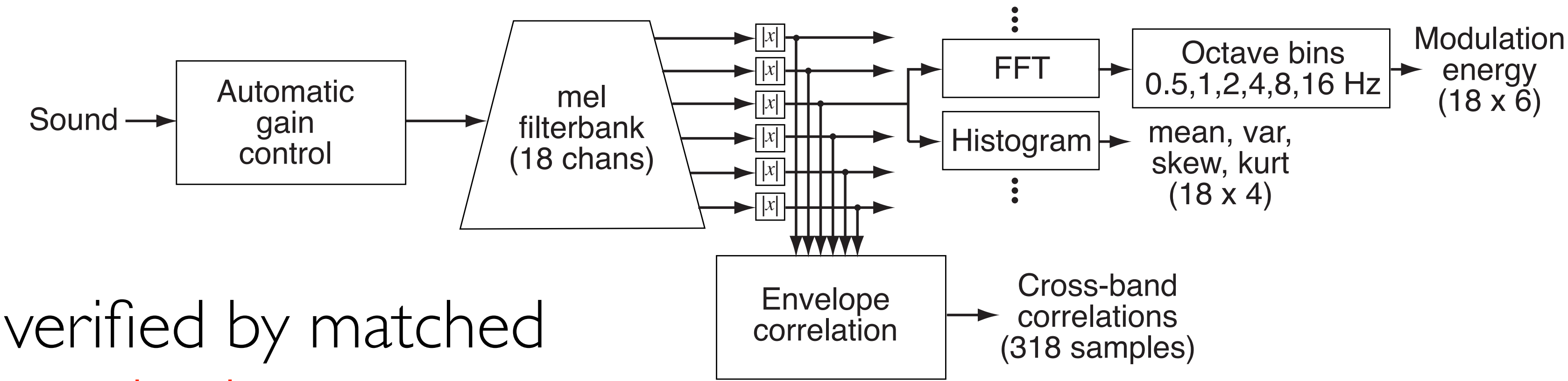
- High precision for top hits (in-domain)



Sound Texture Features

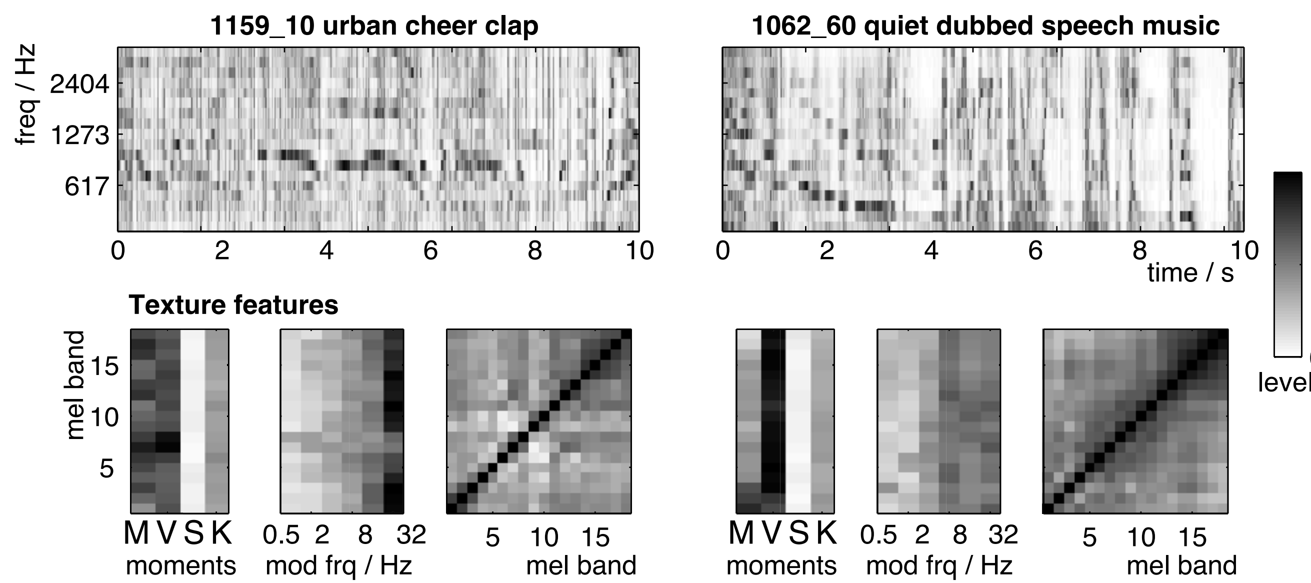
McDermott et al. '09
Ellis, Zheng, McDermott '11

- Characterize sounds by perceptually-sufficient statistics



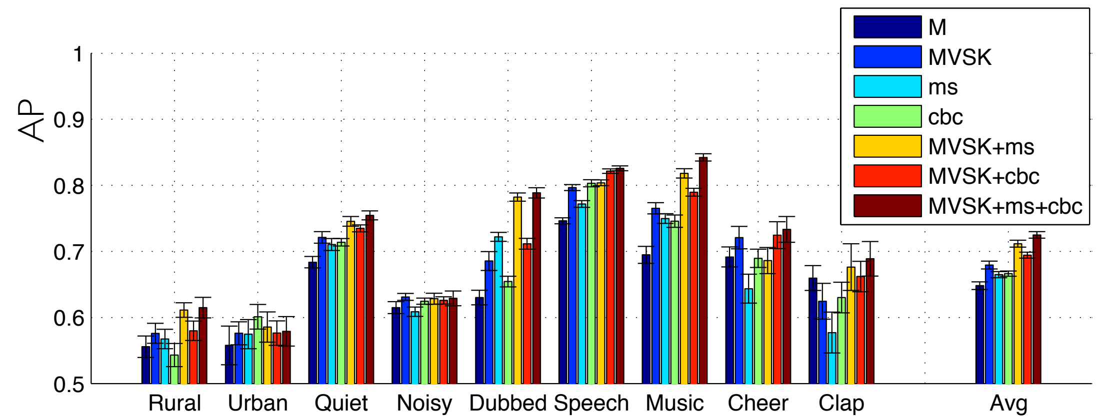
  - .. verified by matched resynthesis

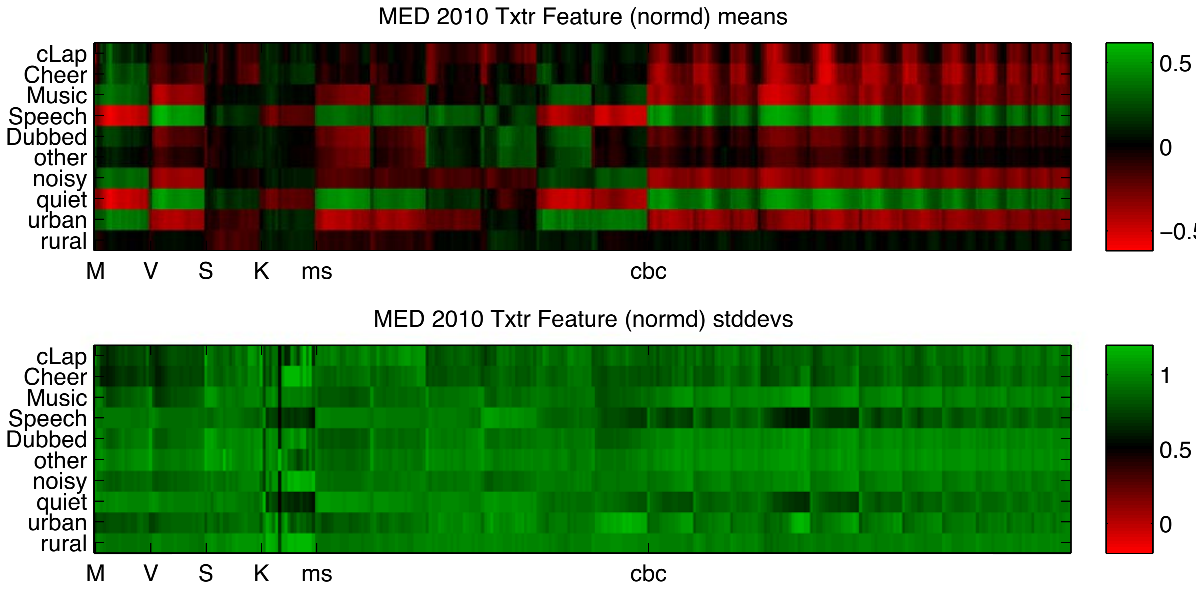
- Subband distributions & env x-corrs
  - Mahalanobis distance ...



Sound Texture Features

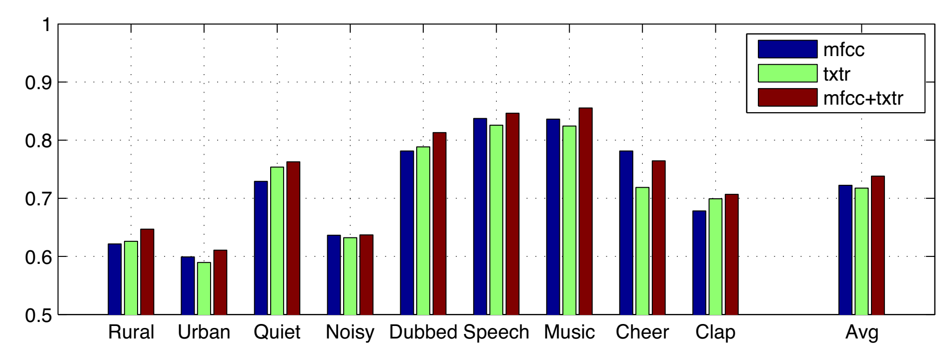
- Test on MED 2010 development data
  - 10 audio-oriented manual labels





- Per-class stats
  - relate dimensions to classes?

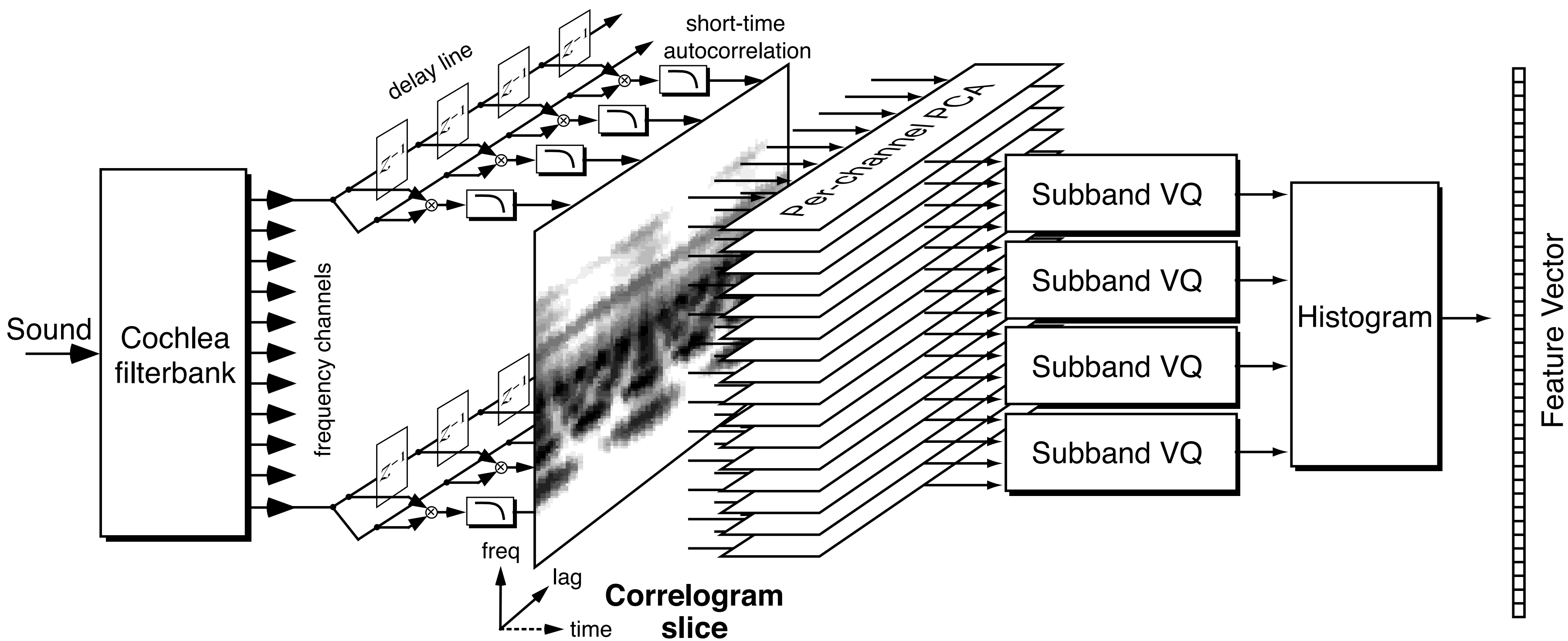
- Perform ~ same as MFCCs
  - covariance ~ texture?



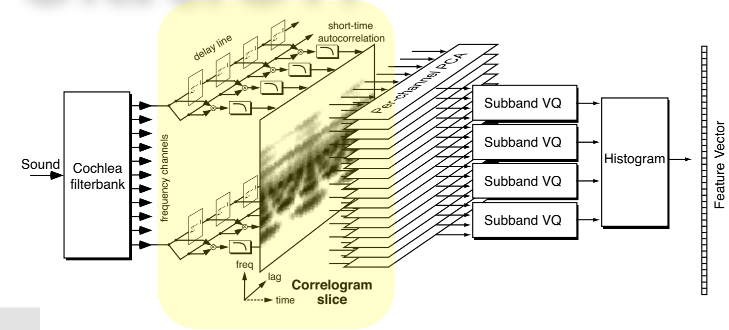
Auditory Model Features

Lyon et al. 2010
Cotton & Ellis 2013

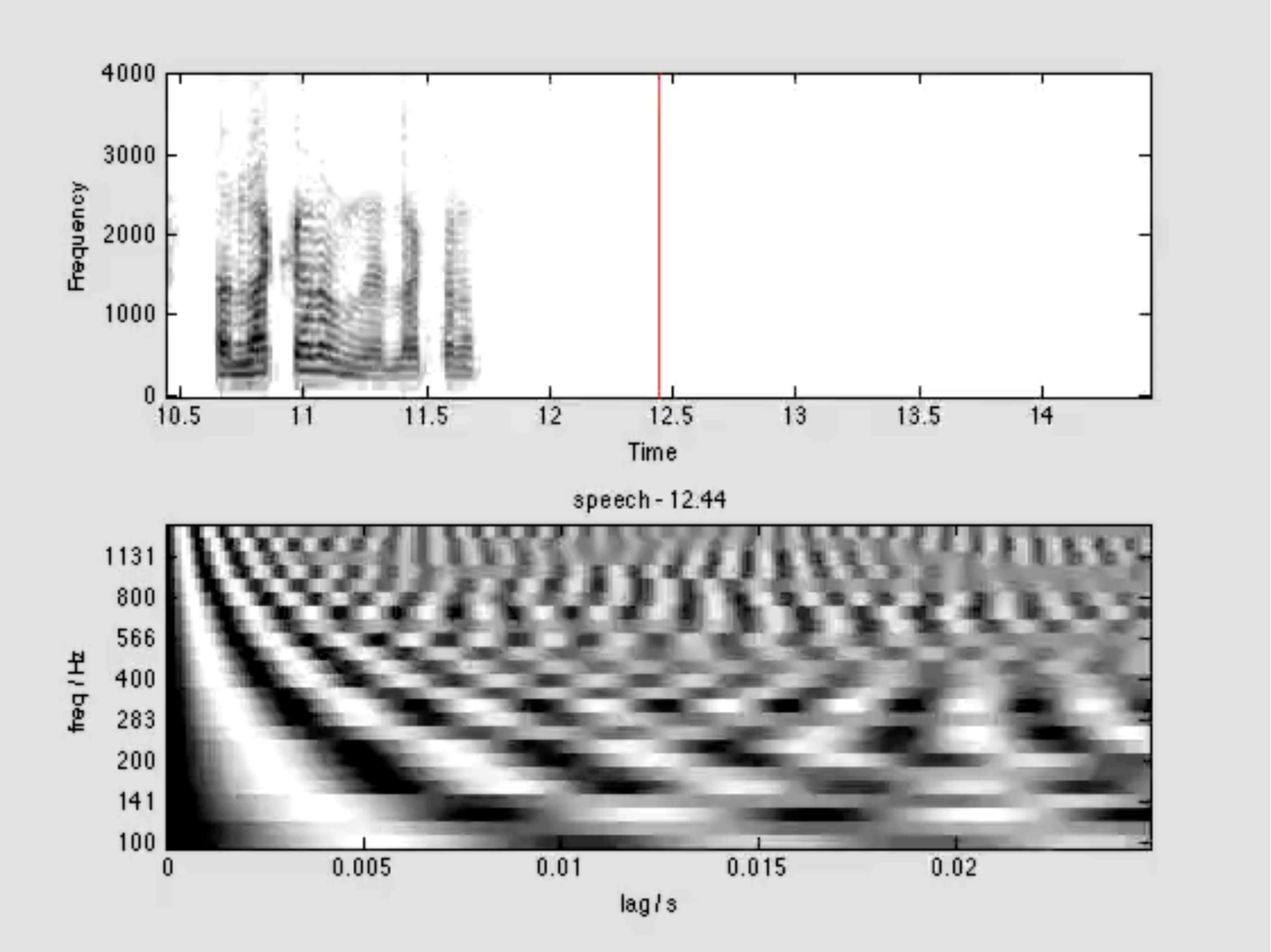
- **Subband Autocorrelation PCA (SBPCA)**
  - Simplified version of Lyon et al. system
  - 10x faster (RTx5 → RT/2)
- **Captures fine time structure in multiple bands**
  - .. missing in MFCC features



Subband Autocorrelation



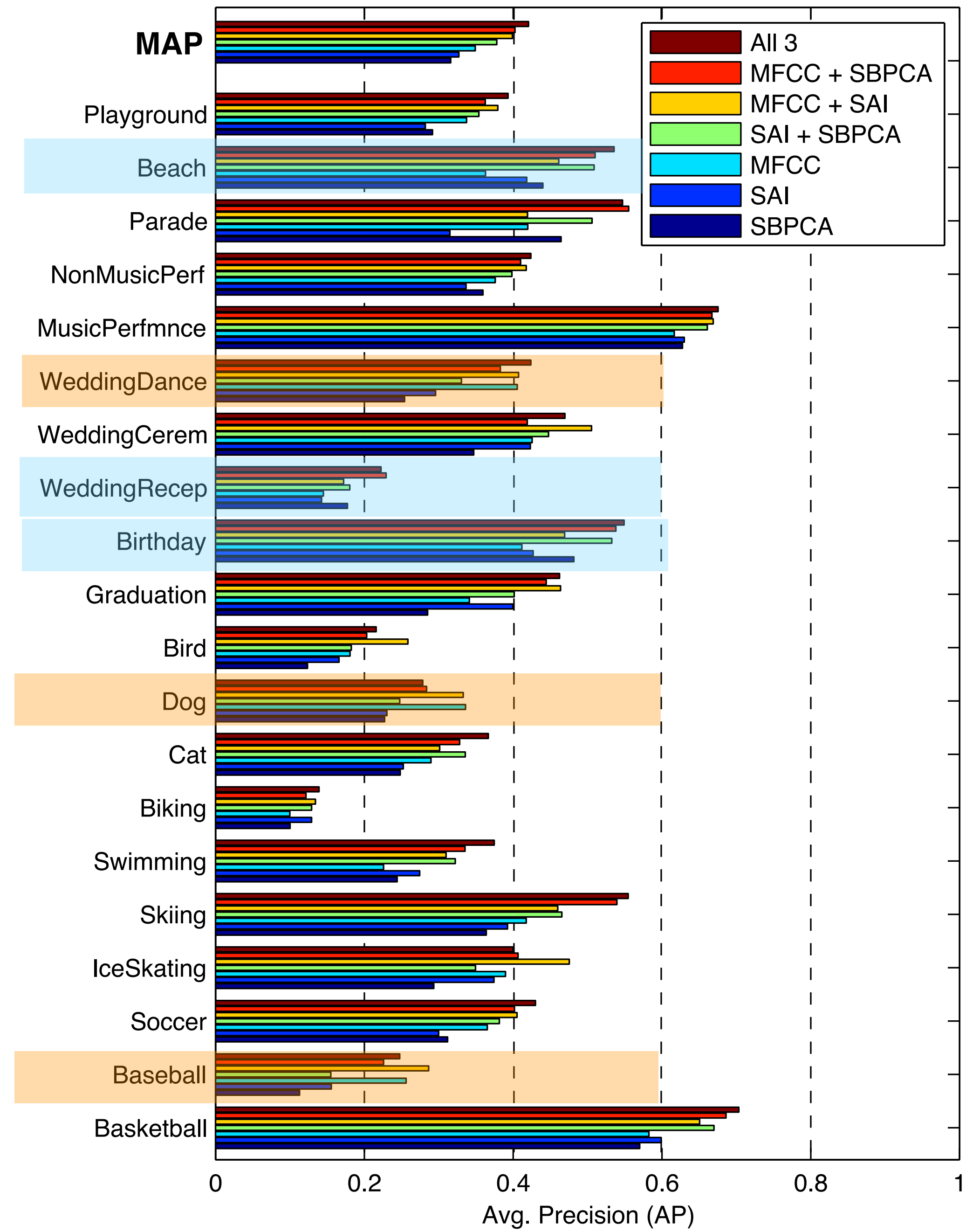
- Autocorrelation stabilizes fine time structure



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

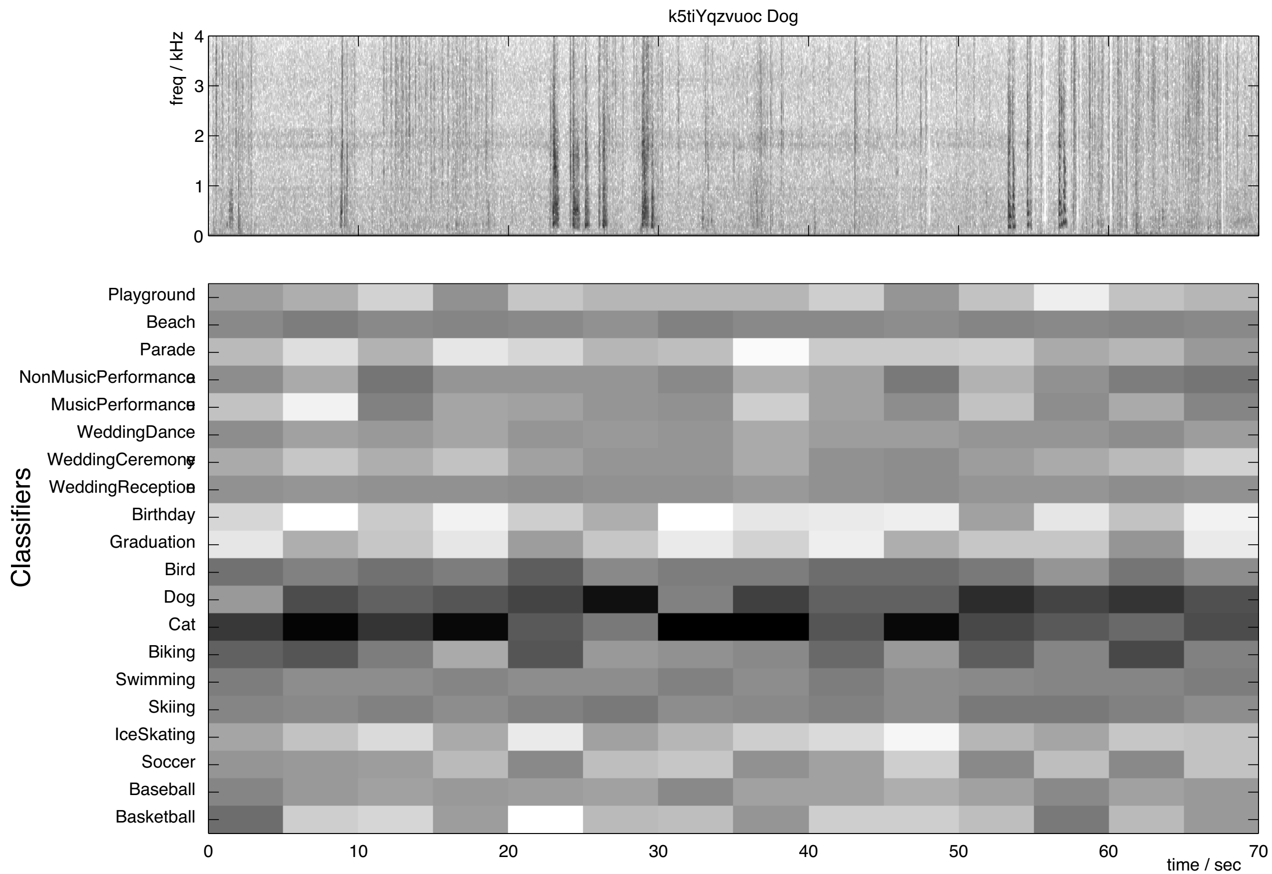
Auditory Model Feature Results

- 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



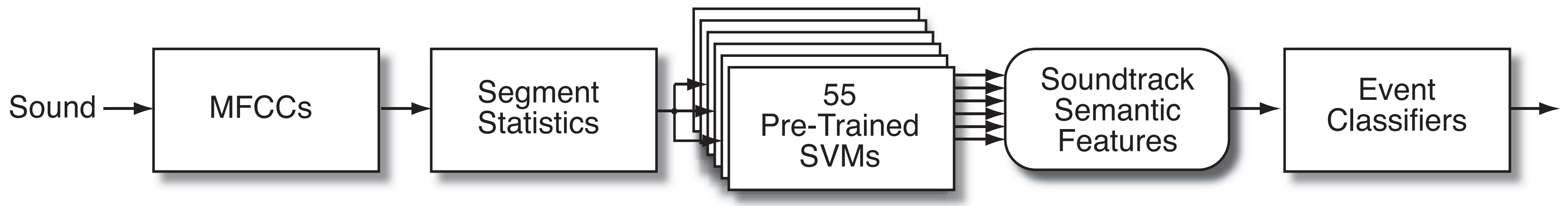
What is Being Recognized?

- **Soundtracks represented by global features**
  - MFCC covariance, codebook histograms
  - What are the critical parts of the sound?



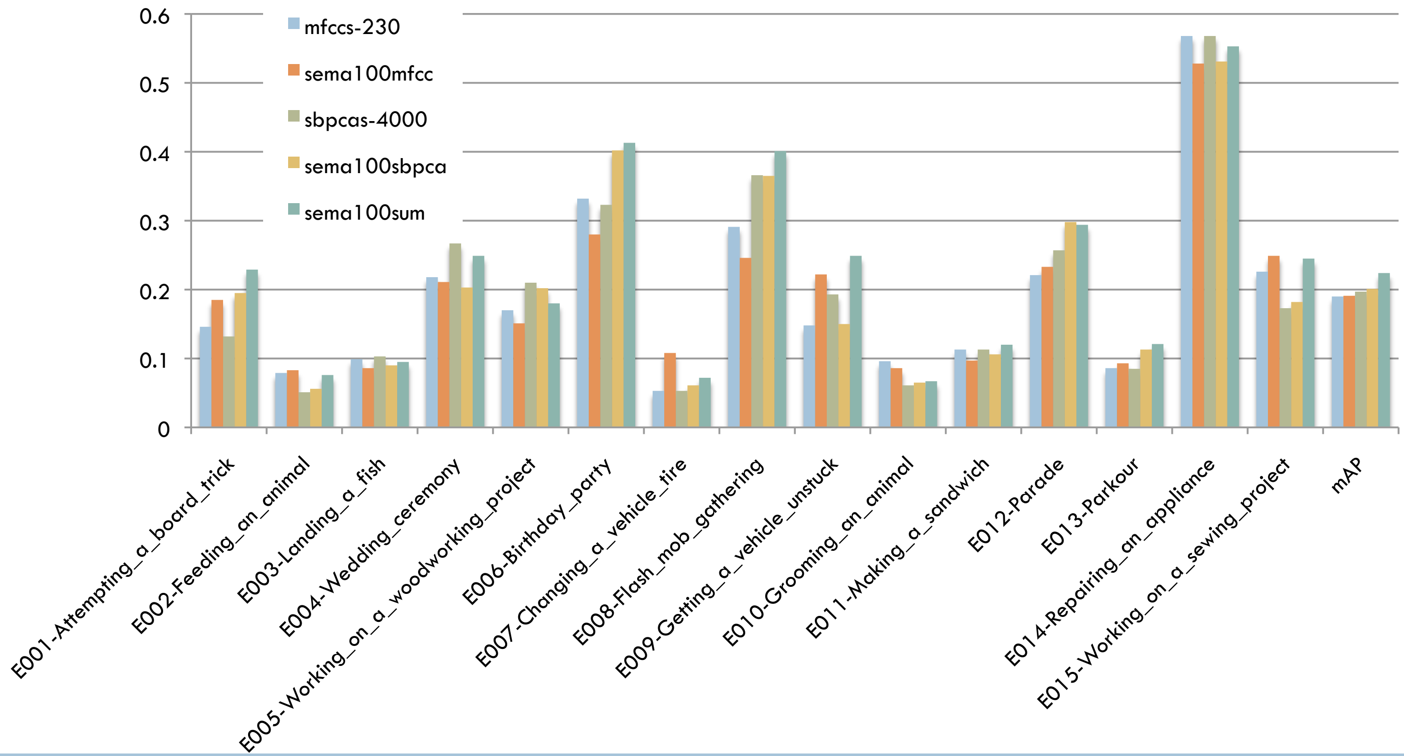
Semantic Audio Features

- Train classifiers on related labeled data



  - defines a new "semantic" feature space

- Use for target classifier
  - or combo



4. Labels & Annotation

Burger et al. '12

- "Semantic Features" are a promising approach
  - but we need good coverage...
  - how to learn more categories?

- Annotation is expensive
  - fine time annotation
    > 10x real-time
  - a few hours are available

- What to label?
  - generic vs. task-specific

animal	singing	clatter
anim_bird	music_sing	rustle
anim_cat	music	scratch
anim_ghoat	knock	hammer
anim_horse	thud	washboard
human_noise	clap	applause
laugh	click	whistle
scream	bang	squeak
child	beep	tone
mumble	engine_quiet	sirene
speech	engine_light	water
speech_ne	power_tool	micro_blow
radio	engine_heavy	wind
white_noise	cheer	
other_creak	crowd	

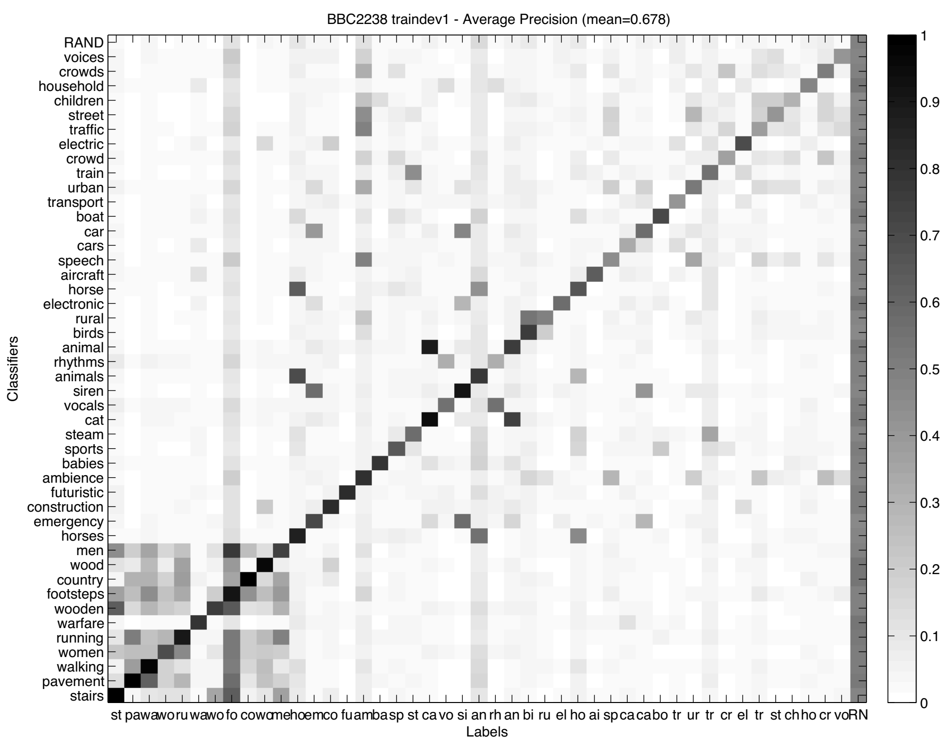
BBC Audio Semantic Classes

- BBC Sound Effects Library
  - 2238 tracks (60 h)
  - short descriptions

```
SFX001-04-01    Wood Fire Inside Stove  5:07
SFX001-05-01    City Skyline City Skyline    9:46
SFX001-06-01    High Street With Traffic, Footsteps
SFX001-07-01    Car Wash Automatic, Wash Phase Inside R
SFX001-08-01    Motor Cycle Yamaha Rd 350: Motor Cycle
SFX001-09-01    Motor Cycle Yamaha Rd 350, Rider Runs U
```

```
290 footsteps          59 men
267 on                 59 general
240 animals            55 switch
197 ambience           53 starts
193 interior           53 crowds
...                    ...
```

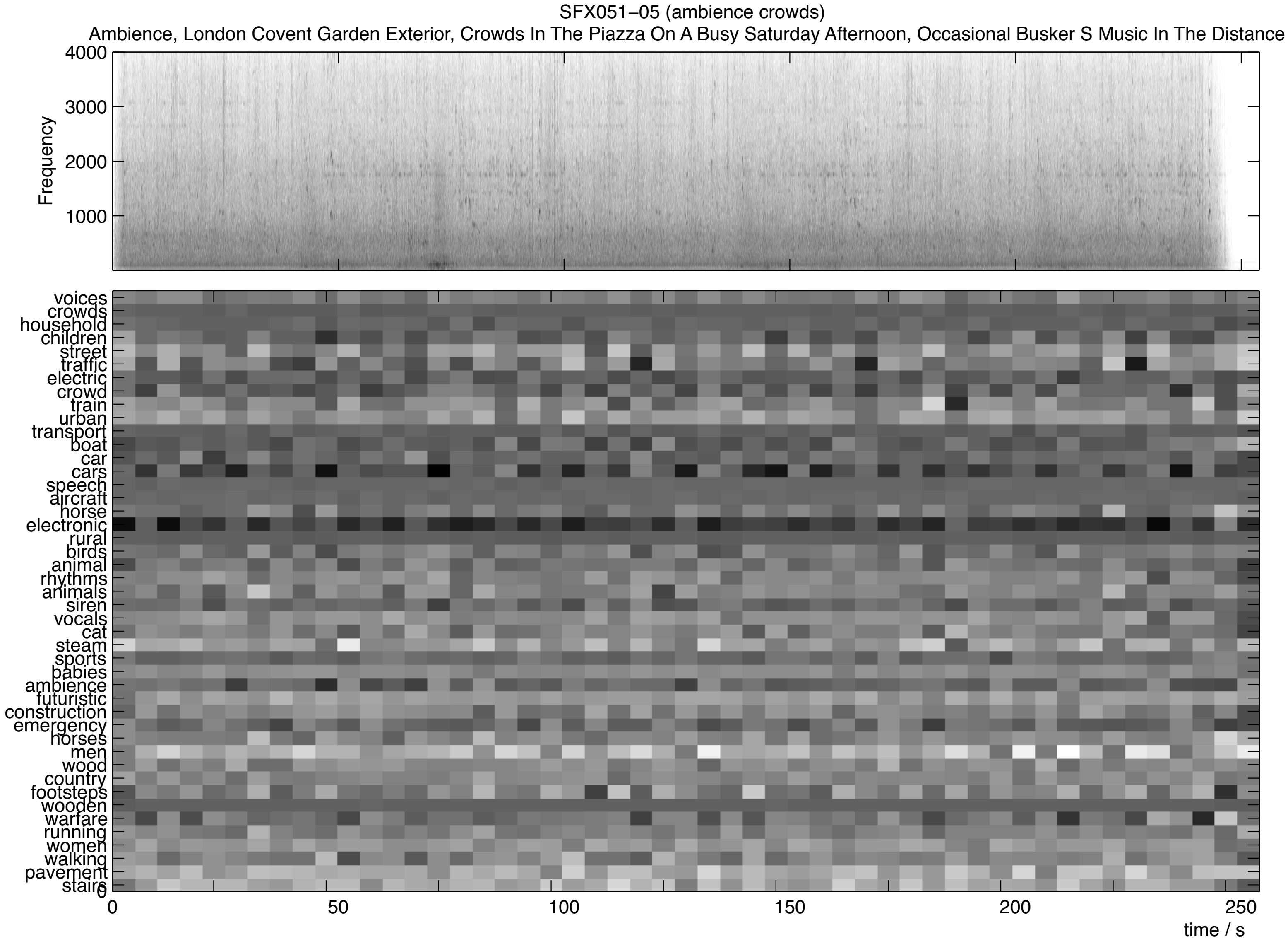
- Use top 45 keywords



- Added as "semantic units"
  - some redundancy visible in mutual APs

BBC Audio Semantic Classes

- Limited semantic correspondence

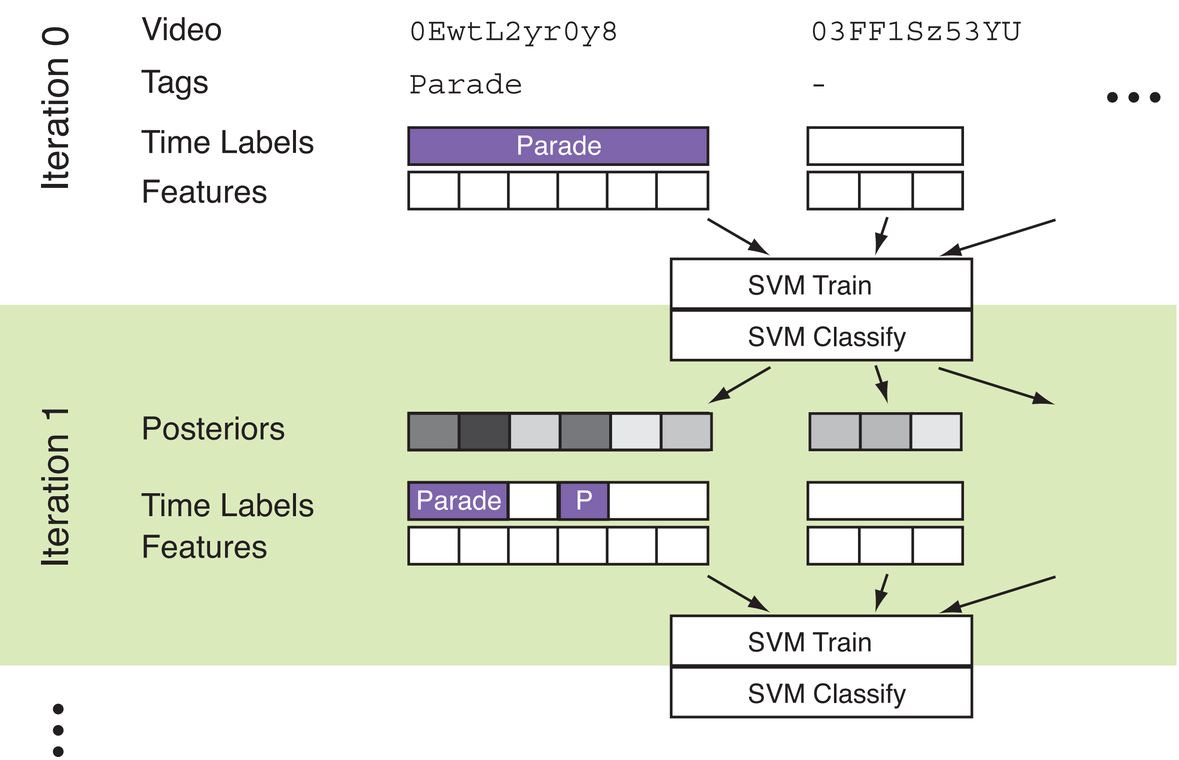


Label Temporal Refinement

K Lee, Ellis, Loui '10

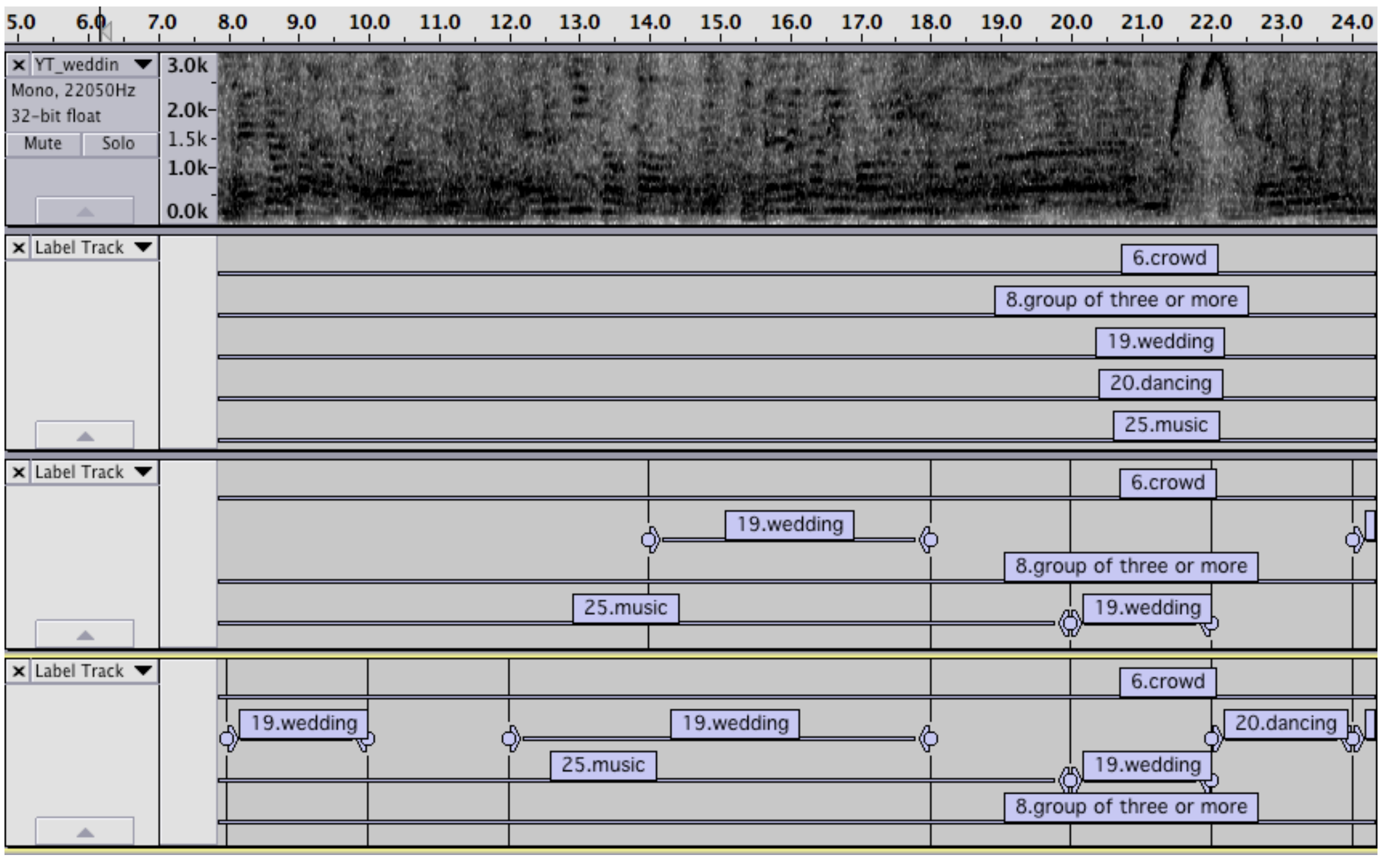
- Audio Ground Truth at coarse time resolution
  - better-focused labels give better classifiers?
  - but little information in very short time frames

- Train classifiers on shorter (2 sec) segments?
  - Initial labels apply to whole clip
  - Relabel based on most likely segments in clip
  - Retrain classifier



Label Temporal Refinement

- Refining labels is "Multiple Instance Learning"
  - "Positive" clips have at least one +ve frame
  - "Negative" clips are all –ve
- Refine based on previous classifier's scores
  - threshold from CDFs of +ve and –ve frames
  - mAP improves ~10% after a few iterations



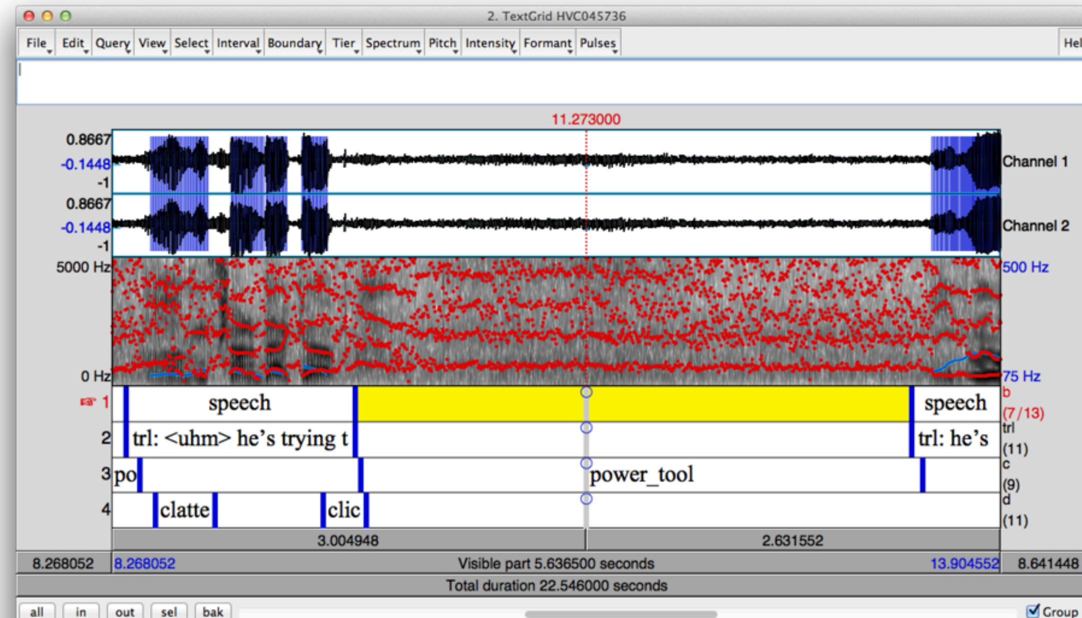
5. Future: Tasks & Metrics

- Environmental sound recognition: What is it good for?
  - media content description ("Recounting")
  - environmental awareness

- What are the right ways to evaluate?
  - task-specific metrics: AEER, F-measure
  - downstream tasks: WER, mAP
  - real applications: archive search, aware devices

Labels & Annotations

- Training data: quality vs. quantity
  - quality costs:
    - DCASE ~ 0.3 h
    - TRECVID MED (Aladdin) ~ 10 h
  - quantity always wins

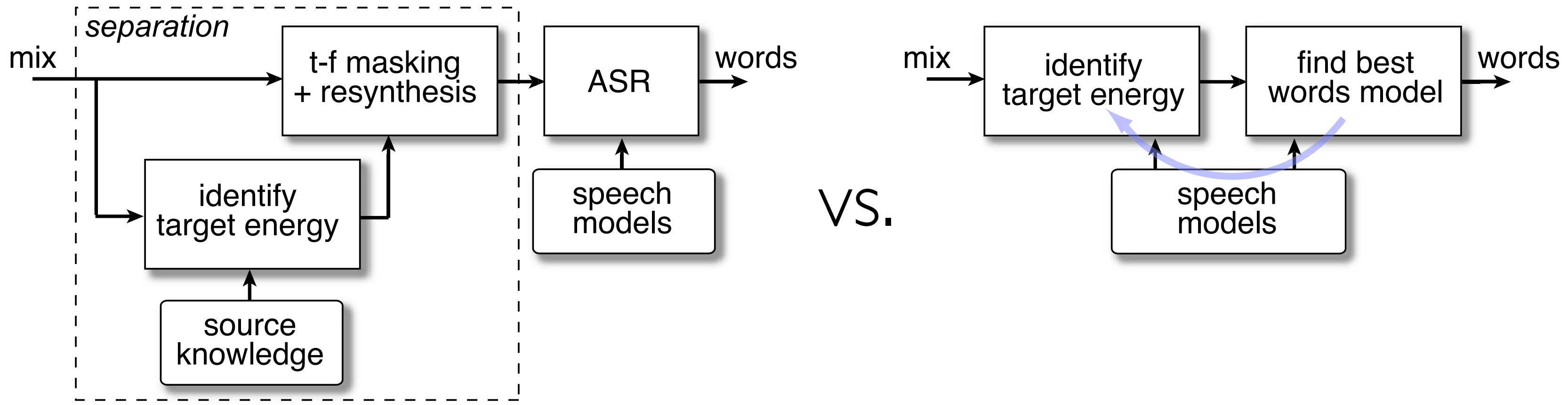


Susanne Burger CMU

- Opportunistic labeling
  - e.g. Sound Effects library, subtitles ...
  - need refinement strategies
- Existing annotations indicate interest

Source Separation

- Separated sources makes event detection easy
  - "separate then recognize" paradigm



  - integrated solution more powerful...
- Environmental Source Separation is ill-defined
  - relevant "sources" are listener-defined
  - environment description addresses this
  - Environment recognition for source separation

Summary

- (Machine) Listening:
  Getting useful information from sound

- Foreground event recognition
  ... by focusing on peak energy patches

- Background sound retrieval
  ... from long-time statistics

- Data, Labels, and Task
  ... what are the sources of interest?

References 1/2

- Samer Abdallah, Mark Plumbley, “Polyphonic music transcription by non-negative sparse coding of power spectra,” *ISMIR*, 2004, p. 10-14.
- Susanne Burger, Qin Jin, Peter F. Schulam, Florian Metze, “Noisemes: Manual annotation of environmental noise in audio streams,” Technical Report LTI-12-017, CMU, 2012.
- Courtenay Cotton, Dan Ellis, & Alex Loui, “Soundtrack classification by transient events,” *IEEE ICASSP*, Prague, May 2011.
- Courtenay Cotton & Dan Ellis, “Spectral vs. Spectro-Temporal Features for Acoustic Event Classification,” *IEEE WASPAA*, 2011, 69-72.
- Courtenay Cotton and Dan Ellis, “Subband Autocorrelation Features for Video Soundtrack Classification,” *IEEE ICASSP*, Vancouver, May 2013, 8663-8666.
- Dan Ellis, “Prediction-driven computational auditory scene analysis.” Ph.D. thesis, MIT Dept of EECS, 1996.
- Dan Ellis, Xiaohong Zheng, Josh McDermott, “Classifying soundtracks with audio texture features,” *IEEE ICASSP*, Prague, May 2011.
- Dimitrios Giannoulis, Emmanouil Benetos, Dan Stowell, Mathias Rossignol, Mathieu Lagrange and Mark Plumbley, “Detection and Classification of Acoustic Scenes and Events,” Technical Report EECSRR-13-01, QMUL School of EE and CS, 2013.
- Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Dan Ellis, Alex Loui, “Consumer video understanding: A benchmark database and an evaluation of human and machine performance,” *ACM ICMR*, Apr. 2011, p. 29.

References 2/2

- Keansub Lee, Dan Ellis, Alex Loui, “Detecting local semantic concepts in environmental sounds using Markov model based clustering,” *IEEE ICASSP*, 2278-2281, Dallas, Apr 2010.
- Keansub Lee & Dan Ellis, “Audio-Based Semantic Concept Classification for Consumer Video,” *IEEE TASLP.* 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.
- Josh McDermott, Andrew Oxenham, Eero Simoncelli, “Sound texture synthesis via filter statistics,” *IEEE WASPAA*, 2009, 297-300.
- Paul Over, George Awad, Jon Fiscus, Brian Antonishek, Martial Michel, Alan Smeaton, Wessel Kraaij, Wessel, Georges Quénot, “TRECVID 2010 - An overview of the goals, tasks, data, evaluation mechanisms, and metrics,” Technical Report, NIST, 2011.
- Manuel Reyes-Gomez & Dan Ellis, “Selection, Parameter Estimation, and Discriminative Training of Hidden Markov Models for General Audio Modeling,” *IEEE ICME*, Baltimore, 2003, I-73-76.
- Paris Smaragdis & Judith Brown, “Non-negative matrix factorization for polyphonic music transcription,” *IEEE WASPAA*, 2003, p. 177-180.
- Tuomas Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria,” *IEEE TASLP* 15(3): 1066-1074, 2007.

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.