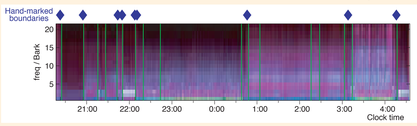
# Fingerprinting to Identify Repeated Sound Events in Long-Duration Personal Audio Recordings

James P. Ogle and Daniel P.W. Ellis • LabROSA, Columbia University • jpo2101@columbia.edu, dpwe@ee.columbia.edu

**Summary:** Body-worn audio recorders can collect huge "personal audio" archives of everything heard by the user, but navigating this data is a challenge. We investigate a noise-resistant audio fingerprint as a way to identify recurrent sound events. The fingerprint works well for data that is highly repeatable (e.g.phone rings) but not for more "organic" sounds (door closures etc.).

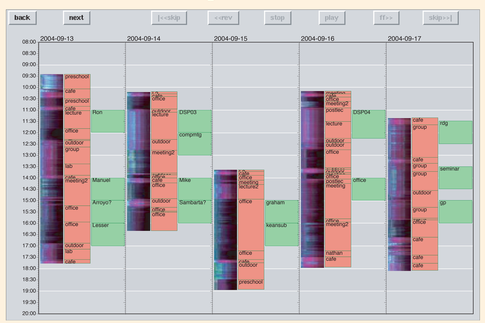
## Personal Audio Archives

- Consumer MP3 players (e.g. iRiver T10) can also record continuously for over 12 hours on a single rechargable AA battery
  - Easy to collect a "**personal audio archive**" of everything heard throughout the day
  - .. but finding anything in the recordings can take close to real-time
- We are researching ways get useful information from this data
  - e.g. automatic retrospective calendar of activities/locations
- Work so far addresses **segmenting and clustering** archives [Ellis & Lee 06]
  - works with time frames of 6..120 sec
  - investigates best features to capture background ambience
  - segmentation via BIC criterion (like speaker segmentation)



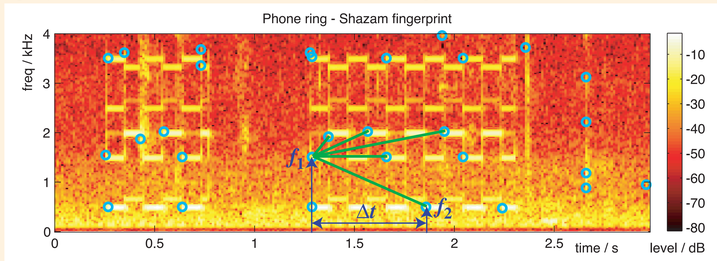

*Spectrogram-like representation of 8 hour recording shows energy (intensity), spectral flatness (saturation) and variance (hue) in 1 min windows .*

  - cluster recurring ambiences/environments with spectral clustering
- This work looks for **repeating foreground events** based on fingerprinting
  - repeating events may be relevant to user e.g. phone rings, theme songs
  - data-mining: repeats can be identified without user intervention
  - want to find repeats despite changes in channel and background (unlike exact repeats of [Johnson et al 00, Kashino et al 03, Herley 06])
- Vision is for interactive browser/calendar displaying multiple sources of information gleaned from recordings and other sources




## Sound Event Fingerprints

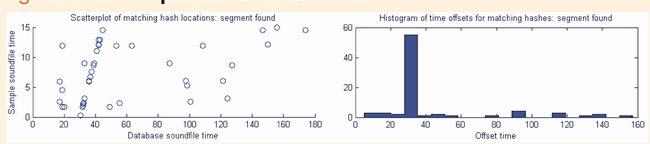
- To find repeating events in the long-duration recordings, we use the fingerprinting technique from Shazam [Wang 02, 06]




  - Prominent peaks – **landmarks** – are selected in a spectrogram, thresholded to have a roughly constant rate in six frequency bands
  - Each landmark is **paired** with up to 9 neighbors nearby in time-frequency
  - Each pair gives a **combinatorial hash** defined by the frequencies of the two landmarks and the time between them $\{f_1, f_2, \Delta t\}$
    - quantizing each component to 6 bits gives $2^{18}$ (262,144) distinct hashes
  - An **index file** records the times when each hash occurs (a multi-hour recording has an index of <10MB)

```
2007-04-11-0839.idx
  00|00|00: 7012.45 11052.33 96384.28
  00|00|01: 123.11 125.87 23004.66 61993.83
  00|00|02:
  00|00|03: 71552.34 101663.03
  ...
```

  - Multiple hashes occurring around two time locations with the same relative timing indicate a **repeated sound event**




### Key Advantages of Shazam Fingerprint:

- No time framing to influence the hash (unlike [Burges et al 03, Herley 06])
- Spectral peaks make hashes almost invariant to background noise
- Missing any single hash does not preclude matching
- Lower bound number of matching hashes allows precision/recall tradeoff

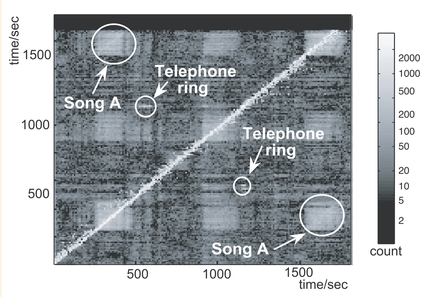
## Finding Repeated Events

- To find possible earlier instances of events in current window:
  - retrieve times of all earlier instances of current hashes (fast because store is indexed by hash value)
  - make a histogram of relative timings
  - look for large peak → repeated event
    ... nearly **constant time** search

| Archive Length (min) | 60 | 120 | 180 | 390 |
|---|---|---|---|---|
| Search time (ms) | 21 | 31 | 37 | 131 |

### Example

- Histogram of # shared hashes in 5 sec windows for 30 min personal audio recording, with two instances of a phone ring and three plays of music recording "Song A".




## Evaluation

| Category | Recall | | Precision | |
|---|---|---|---|---|
| Production Audio | 29/30 | 97% | 29/34 | 85% |
| Alert Sounds | 45/65 | 69% | 45/45 | 100% |
| Organic Sounds | 0/20 | 0% | 0/20 | 0% |

| SNR / dB | 3 | -3 | -9 | -15 | -21 |
|---|---|---|---|---|---|
| Recall | 100% | 89% | 89% | 89% | 56% |

- Exactly-repeating sounds (alarms, recordings) are detected well; "organic" sounds (speech, door closing) are not.
- Search for particular event (telephone ring) shows excellent resistance to background noise.

### References

[Burges et al 03] C.J.C. Burges, J.C. Platt, S. Jana "Distortion discriminant analysis for audio fingerprinting" IEEE Tr. Speech and Audio Proc., 11(3):165–174, 2003.

[Ellis & Lee 06] D.P.W. Ellis and K. Lee "Accessing minimal-impact personal audio archives" IEEE MultiMedia Magazine, 13(4):30–38, 2006.

[Herley 06] C. Herley "ARGOS: Automatically extracting repeating objects from multimedia streams" IEEE Tr. Multimedia, 8(1):115–129, 2006.

[Johnson et al 00] S. Johnson & P. Woodland "A Method for Direct Audio Search with Applications to Indexing and Retrieval" Proc. ICASSP, III:1427–1430, 2000.

[Kashino et al 03] K. Kashino, T. Kurozumi, H. Murase "A quick search method for audio and video signals based on histogram pruning" IEEE Tr. Multimedia, 5(3):348–357, 2003.

[Wang 02] A. Wang "An industrial-strength audio search algorithm" Proc. ISMIR 2003.

[Wang 06] A. Wang "The Shazam music recognition service" Comm. ACM, 49(8):44–48, Aug 2006.