Name That Tune: Stream Audio Fingerprinting

Chris Burges, John Platt, Jon Goldstein, Erin Renshaw

Microsoft Research
Outline

• What is audio fingerprinting?
• Robust Audio Recognition Engine Demo
• Some ideas for applications
• Other companies doing audio fingerprinting
• How does RARE work?
• How well does RARE work?
What is Audio Fingerprinting?

• *Prepare database*: Given audio clips, extract a ‘fingerprint’ from each clip, store

• Clip is *not* changed, unlike watermarking

• *Run*: Compare ‘traces’, computed from an audio stream, against your database of stored fingerprints, to identify what’s playing. Confirm using second fingerprint.
Why Is This A Research Topic?
- *Audio Feature Extraction*

- Map high dim space of audio samples to low dim feature space
- Must be robust to likely input distortions
- Must be informative (different audio segments map to distant points)
- Must be efficient (use small fraction of resources on typical PC)
The RARE Fingerprinting System

- Identifies clips based on 6 second “fingerprints”
- 6 seconds mapped to 64 floats, generated once every 0.186 seconds (or 0.372 sec)
- Each 64-float vector checked against large database of fingerprints (e.g., 240,000)
- End-to-end system requires approx. 10% CPU on 746 MHz PC
- Confirmatory fingerprint can be used to add further robustness
How Confirmatory Fingerprint Works

FIND SINGLE CONFIRM FP

MATCH? YES WINNER!

IN DATABASE?

NO

NO

NO

NO

YES
RARE Demo

- 7 songs chosen at random, out of 239,209
- Play clean song, then play distorted version

Critical Mass: Good Morning: Clip to distort
The Glands: Free Jane: ‘Old Radio’ filter
Eddie Cantor: Joe Is Here: 8% Time Compress
Ella Fitzgerald: Somebody Loves Me: Add Reverb
Perl Bailey: They Didn’t Believe Me: Repeat Echo
Next: Stop, Drop and Roll: 20dB Noise Gate
Joy Division: Digital: Add treble to distort
Outline

• What is audio fingerprinting?
• Robust Audio Recognition Engine Demo
• Some possible applications
• Other companies doing audio fingerprinting
• How does RARE work?
• How well does RARE work?
Identifying Broadcast Music I

• You’re listening to FM radio at home

• You’d like to know what just played…
• …your eHome media center can tell you
Making WMP Smarter

• You’ve made your own CDs by mixing and matching songs from your collection

• You’d like Windows Media Player to tell you what’s playing!
Identifying Broadcast Music II

• You’re listening to FM radio on the road

• You’d like to know what just played, by holding a cell phone or PDA up to the source
Database Cleaning

• You’re managing a large database of audio

• You’d like to automatically identify and remove duplicates

• You’d like to add metadata to the rest
ASCAP wants to know what radio stations in a given area are playing.

Businesses want to know if the commercial they paid for aired, and when, and if it was ‘lexiconed’.

Content providers want to know when pirated content is played.
Outline

• What is audio fingerprinting?
• *Robust Audio Recognition Engine* Demo
• Some ideas for applications
• Other companies doing audio fingerprinting
• How does RARE work?
• How well does RARE work?
Other AF Efforts

• AUDIBLE MAGIC: Partnered with Loudeye, CMJ, SESAC.
• RELATABLE: Used by Napster. Uses 30 sec.
• MOODLOGIC: File based, p2p royalty tracking.
• PHILIPS: Cell-phone based music recognition service.
• SHAZAM: Cell-phone based, UK market
• MSN Music: File based
Why use RARE?

• RARE is flexible. Features are NOT hard-wired – they can easily be re-learned to handle new kinds of noise
• RARE is very robust and very fast
• This is Microsoft’s own technology: easily molded to different applications, we own the code and IP (2 patents filed)
Outline

- What is audio fingerprinting?
- *Robust Audio Recognition Engine* Demo
- Some ideas for applications
- Other companies doing audio fingerprinting
- How does RARE work?
- How well does RARE work?
Feature Extraction

- Downsample to 11.025 KHz mono
- Compute time-frequency map, take magnitudes
- De-Equalize
- Perceptual Thresholding
- Compute Robust Projections
- Collate Samples
  (Repeat)
- Output 64 floats every half-frame
De-Equalization

De-equalize by flattening the log spectrum.

Before

After
Perceptual Thresholding

1 KHz

1.5 KHz

50 dB

3 KHz

So: remove coefficients that are below a perceptual threshold to lower unwanted variance.
Feature Extraction, cont.

- Output 64 floats every half-frame
- Downsample to 11.025 KHz mono
- Compute time-frequency map, take magnitudes
- De-Equalize
- Perceptual Thresholding
- Compute Robust Projections
- Collate Samples
  (Repeat)
- Output 64 floats every half-frame
Distortion Discriminant Analysis

372ms, step every 186ms

OPCA

6.1s, step every 186ms

OPCA

2048

64 64

32

64

OPCA

64
Oriented PCA

Want $\delta_1$ much smaller than $\delta_2$
Oriented PCA, Details

• The variance of the signal along unit vector $n$ is $n'C_1n$, where $C_1$ is the signal covariance matrix.

• The sum of squared differences of (signal-noisy signal) projected along $n$ is $n'C_2n$, where $C_2$ is the correlation matrix of the ‘noise’.

• Find $n$ that maximize $\frac{n'C_1n}{n'C_2n}$.
Distortion Discriminant Analysis

DDA is a convolutional neural network with weights trained using OPCA

372ms, step every 186ms

OPCA - can train with different noise!
RARE Training Data

- 200 20s audio segments, catted (67 min)
- Compute 11 distortions
- Perform one layer of OPCA
- Subtract mean projections of train data, norm noise to unit variance
- Repeat
- Use 10 20s segments as validation set to compute scale factor for final projections
The Distortions

Original

3:1 compress above 30dB

Light Grunge

Distort Bass

“Mackie Mid boost” filter

Old Radio

Phone

RealAudio© Compander
Outline

- What is audio fingerprinting?
- Robust Audio Recognition Engine Demo
- Some ideas for applications
- Other companies doing audio fingerprinting
- How does RARE work?
- How well does RARE work?
Test Results – Large Test Set

- 500 songs, 36 hours of music
- One stored fingerprint computed for every song, about 10s in, random start point
- Test alignment noise, false positive and negative rates, same with distortions
- Define distance between two clips as minimum distance between fingerprint of first, and all traces of the second
- Approx. $3.5 \times 10^8$ chances for a false positive
DDA for Alignment Robustness

Train second layer with alignment distortion:

![Graph showing the relationship between alignment shift and square distance differences. The graph includes two sets of data points, one for No Shift Training and another for With Shift Training. The x-axis represents the alignment shift in milliseconds, ranging from 0 to 180, while the y-axis shows the square distance differences ranging from 0 to 0.04. The data points are color-coded with blue crosses for No Shift Training and green circles for With Shift Training.]
Test Results: Positives

Fingerprints, sorted by distance

Normalized distance squared

- 372ms steps
- 186ms steps

Fingerprints, sorted by distance
Test Results: Negatives

Smallest ‘negative’ score: 0.14, largest ‘positive’ score: 0.026
Tests for Other Distortions

• Take all 500 test clips, grouped 10x50
• Add random alignment shift to all clips
• Distort each group (10 different distortions)
• RESULTS: false positive rate $4.10^{-6}$, false negative rate 0.2%
• But that was without using confirmation!
Tests for Other Distortions, cont.

First 500 out of ¼ million
Lookup: Computational Cost

• Overall system uses < 10% CPU on 850MHz P3 for 240K fingerprints
• Current lookup can handle 160 streams on a quad 2 GHz machine.
• With local caching, server caching, local pruning, and taking advantage of P4 architecture, expect at least a factor of 10 faster: 1,600 streams / machine
Bitvector yields 30x Speedup

<table>
<thead>
<tr>
<th>1st Dimension</th>
<th>2nd Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>❯</td>
<td>❯</td>
</tr>
<tr>
<td>❯</td>
<td>❯</td>
</tr>
</tbody>
</table>

List of FPs compatible with bin

{Search these!}
Conclusions

• RARE gives Microsoft robust, fast, stream audio fingerprinting
• RARE is raring to go…
• For technical details: http://msrweb/~cburges
• Contact aliases: cburges, jplatt