Semantic Audio Studio Tools and Techniques using MPEG-7

Dr. Michael Casey

Centre for Computational Creativity

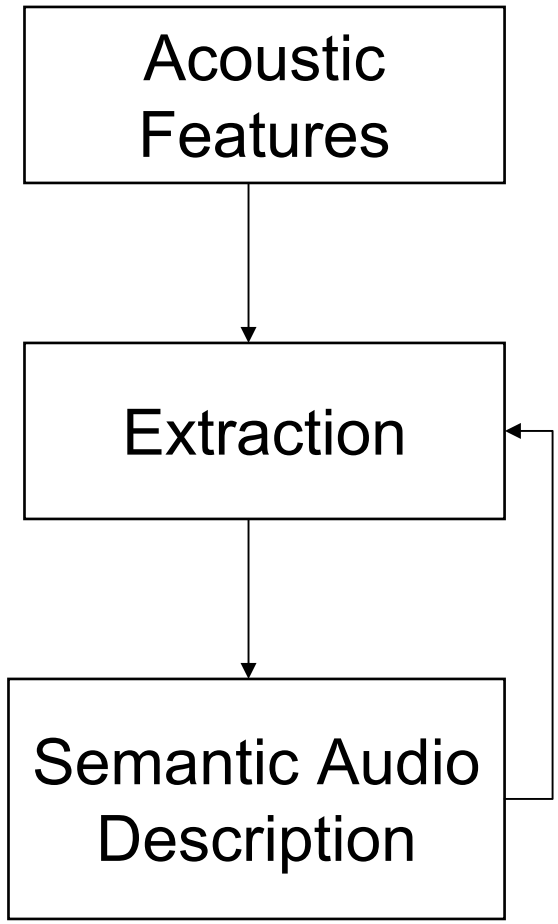
Department of Computing

City University, London

Overview

- MPEG-7 Tools
  - Low Level Audio Descriptors
  - Statistical Sound Models (Semantic ?)
- Music Unmixing
  - Independent Spectrogram Separation
- Sound Classification
  - Automatic label extraction
  - “Semantic” processing
- Segment Similarity, Structure Extraction Musaics
  - S-Matrix (Self-Similarity Matrix)
  - C-Matrix (Cross-Similarity Matrix)
  - Segment Replacement
  - Musaics

Semantic Audio Analysis



MPEG-7 Audio Descriptors

Header

```
<!-- DAFx 2003 MPEG-7 Audio Processing Examples -->
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
        <Description xsi:type="ContentEntityType">
                <MultimediaContent xsi:type="AudioType">
                        <Audio xsi:type="AudioSegmentType">
                                <MediaTime>
                                        <MediaTimePoint>T00:00:00</MediaTimePoint>
                                        <MediaDuration>PT15S450N1000F</MediaDuration>
                                </MediaTime>
                                <AudioDescriptor xsi:type="AudioWaveformType">
                                        <SeriesOfScalar hopSize="PT10N1000F" totalNumOfSamples="1545">
                                                <Min>
-6.10352e-05 -0.000274658 -0.0010376 -0.00161743 -0.00210571 -0.00216675 -0.00259399 -0.00473022 -0.00
1 -0.00299072 -0.0043335 -0.00762939 -0.00497437 -0.00683594 -0.00408936 -0.0071106 -0.00296021 -0.010
-0.0118713 -0.00765991 -0.00363159 -0.0100403 -0.0131226 -0.0138245 -0.0131226 -0.00772095 -0.00854492
-0.0020752 -0.00265503 -0.00158691 -0.00140381 -0.0010376 -0.000518799 -0.000610352 -0.000183105
 </Min>
                                                <Max>
 0.000335693 0.000671387 0.00152588 0.000396729 0.00177002 0.00125122 0.00286865 0.00119019 0.00280762
74658 0.0032959 0.0027771 0.00552368 0.00445557 0.00588989 0.00494385 0.00674438 0.00964355 0.0007324
 0.00210571 0.00167847 0.000671387 0.00109863 0.00112915 0.000488281 0.000274658
 </Max>
                                        </SeriesOfScalar>
                                </AudioDescriptor>
                        </Audio>
                </MultimediaContent>
        </Description>
</Mpeg7>
```

MPEG-7 Audio Descriptors

```
<!-- DAFx 2003 MPEG-7 Audio Processing Examples -->
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
        <Description xsi:type="ContentEntityType">
                <MultimediaContent xsi:type="AudioType">
                        <Audio xsi:type="AudioSegmentType">
                                <MediaTime>
                                        <MediaTimePoint>T00:00:00</MediaTimePoint>
                                        <MediaDuration>PT15S450N1000F</MediaDuration>
                                </MediaTime>
                                <AudioDescriptor xsi:type="AudioWaveformType">
                                        <SeriesOfScalar hopSize="PT10N1000F" totalNumOfSamples="1545">
                                                <Min>
-6.10352e-05 -0.000274658 -0.0010376 -0.00161743 -0.00210571 -0.00216675 -0.00259399 -0.00473022 -0.00
1 -0.00299072 -0.0043335 -0.00762939 -0.00497437 -0.00683594 -0.00408936 -0.0071106 -0.00296021 -0.010
-0.0118713 -0.00765991 -0.00363159 -0.0100403 -0.0131226 -0.0138245 -0.0131226 -0.00772095 -0.00854492
-0.0020752 -0.00265503 -0.00158691 -0.00140381 -0.0010376 -0.000518799 -0.000610352 -0.000183105
                                                </Min>
                                                <Max>
0.000335693 0.000671387 0.00152588 0.000396729 0.00177002 0.00125122 0.00286865 0.00119019 0.00280762
74658 0.0032959 0.0027771 0.00552368 0.00445557 0.00588989 0.00494385 0.00674438 0.00964355 0.0007324
0.00210571 0.00167847 0.000671387 0.00109863 0.00112915 0.000488281 0.000274658
                                                </Max>
                                        </SeriesOfScalar>
                                </AudioDescriptor>
                        </Audio>
                </MultimediaContent>
        </Description>
</Mpeg7>
```

Segments

MPEG-7 Audio Descriptors

```
<!-- DAFx 2003 MPEG-7 Audio Processing Examples -->
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
        <Description xsi:type="ContentEntityType">
                <MultimediaContent xsi:type="AudioType">
                        <Audio xsi:type="AudioSegmentType">
                                <MediaTime>
                                        <MediaTimePoint>T00:00:00</MediaTimePoint>
                                        <MediaDuration>PT15S450N1000F</MediaDuration>
                                </MediaTime>
                                <AudioDescriptor xsi:type="AudioWaveformType">
                                        <SeriesOfScalar hopSize="PT10N1000F" totalNumOfSamples="1545">
                                                <Min>
-6.10352e-05 -0.000274658 -0.0010376 -0.00161743 -0.00210571 -0.00216675 -0.00259399 -0.00473022 -0.00
1 -0.00299072 -0.0043335 -0.00762939 -0.00497437 -0.00683594 -0.00408936 -0.0071106 -0.00296021 -0.010
-0.0118713 -0.00765991 -0.00363159 -0.0100403 -0.0131226 -0.0138245 -0.0131226 -0.00772095 -0.00854492
-0.0020752 -0.00265503 -0.00158691 -0.00140381 -0.0010376 -0.000518799 -0.000610352 -0.000183105
                                                </Min>
                                                <Max>
0.000335693 0.000671387 0.00152588 0.000396729 0.00177002 0.00125122 0.00286865 0.00119019 0.00280762
74658 0.0032959 0.0027771 0.00552368 0.00445557 0.00588989 0.00494385 0.00674438 0.00964355 0.0007324
0.00210571 0.00167847 0.000671387 0.00109863 0.00112915 0.000488281 0.000274658
                                                </Max>
                                        </SeriesOfScalar>
                                </AudioDescriptor>
                        </Audio>
                </MultimediaContent>
        </Description>
</Mpeg7>
```

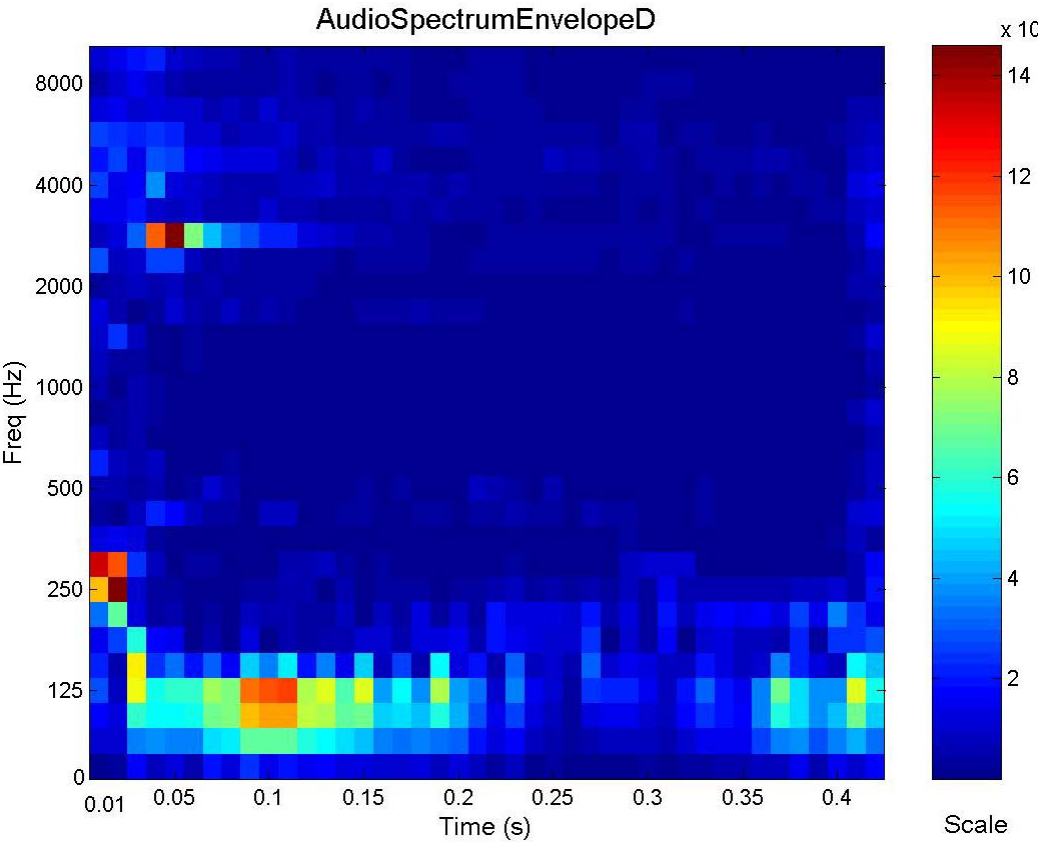
Descriptor

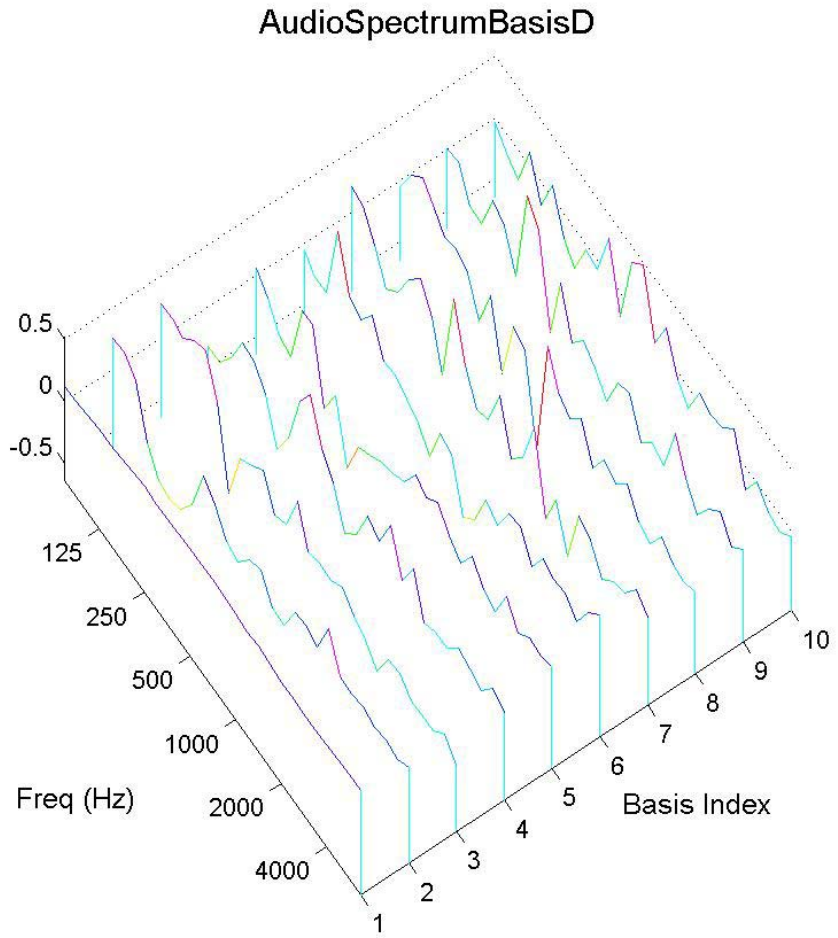
Some Useful Descriptors for Music Processing

- AudioSpectrumEnvelopeD
- AudioSpectrumBasisD
- AudioSpectrumProjectionD
- SoundModelDS
- SoundModelStatePathD
- SoundModelStateHistogramD

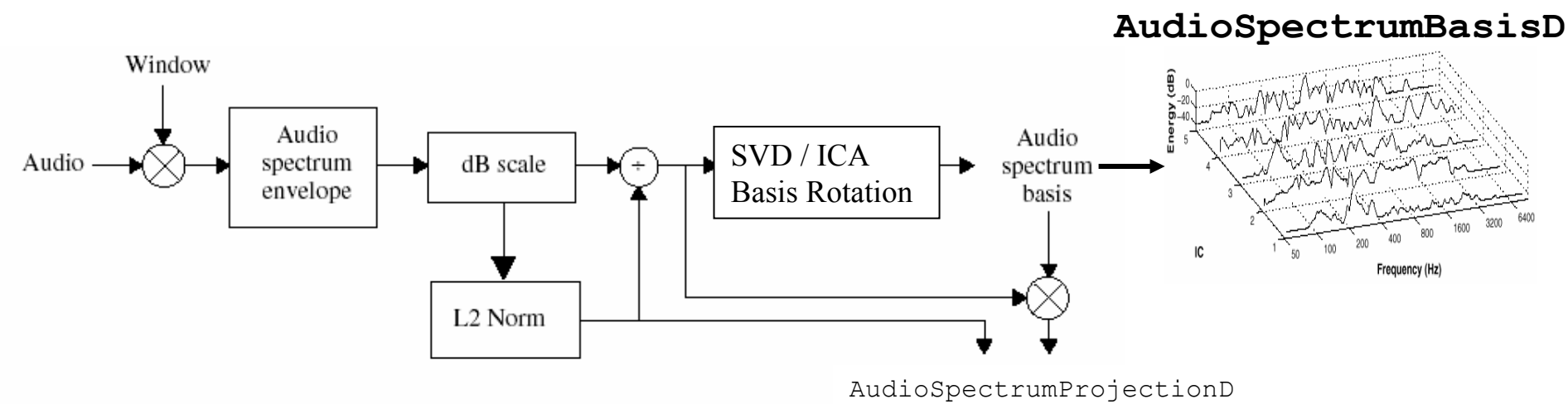
EXAMPLE 1
MUSIC UNMIXING

AudioSpectrumBasisD





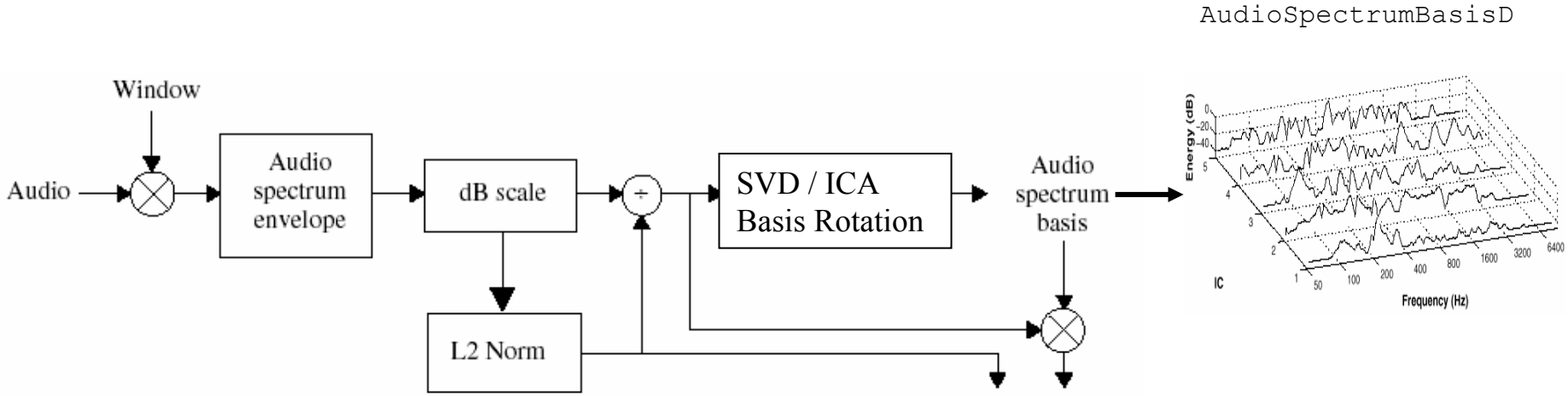
AudioSpectrumBasisD



AudioSpectrumBasisD

```
<?xml version="1.0" encoding="iso-8859-1"?>

<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
         xmlns:xml="http://www.w3.org/XML/1998/namespace"
         xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
        <DescriptionUnit xsi:type="mpeg7:AudioSpectrumBasisType"
                loEdge="62.5"
                hiEdge="8000"
                octaveResolution="1/4">
   <SeriesOfVector totalNumOfSamples="10" vectorSize="29">
                        <Raw mpeg7:dim="10 29">
0.1201    0.2615    0.2824   -0.3166    0.0543   -0.0488    0.2032   -0.0454    0.0059   -0.0380
0.1332    0.2787    0.2871   -0.3218   -0.0599   -0.1399    0.1797    0.1743    0.0297   -0.1650
0.1495    0.1710    0.2622   -0.1536   -0.2318   -0.1331   -0.0150    0.2822   -0.2078   -0.2520
0.1666   -0.2016    0.3710    0.0918   -0.2735    0.4871   -0.2337    0.1640   -0.2134    0.1076
0.1673   -0.3793    0.4110    0.0740    0.2243    0.0979   -0.1349    0.0556    0.1018   -0.1950
0.1795   -0.4073    0.1210   -0.0697    0.2445    0.0275    0.0970    0.0920    0.0264    0.0961
0.1977   -0.3536   -0.4789   -0.3835   -0.3068    0.1915    0.2389    0.0287   -0.2537   -0.2085
0.2142   -0.1924   -0.0700   -0.1612   -0.0765   -0.0466    0.0432   -0.2389    0.5219   -0.3181
0.1961    0.1725    0.0022    0.2620   -0.5313   -0.0297   -0.2916    0.1007    0.3674   -0.0392
0.2026    0.0719    0.0929    0.4462   -0.2368   -0.1012    0.4502   -0.4018   -0.3263   -0.0673
0.2026    0.0719    0.0929    0.4462   -0.2368   -0.1012    0.4502   -0.4018   -0.3263   -0.0673
0.2026    0.0719    0.0929    0.4462   -0.2368   -0.1012    0.4502   -0.4018   -0.3263   -0.0673
                                <!-- More Values Here . . . -->
                        </Raw>
                </SeriesOfVector>
        </DescriptionUnit>
</Mpeg7>
```

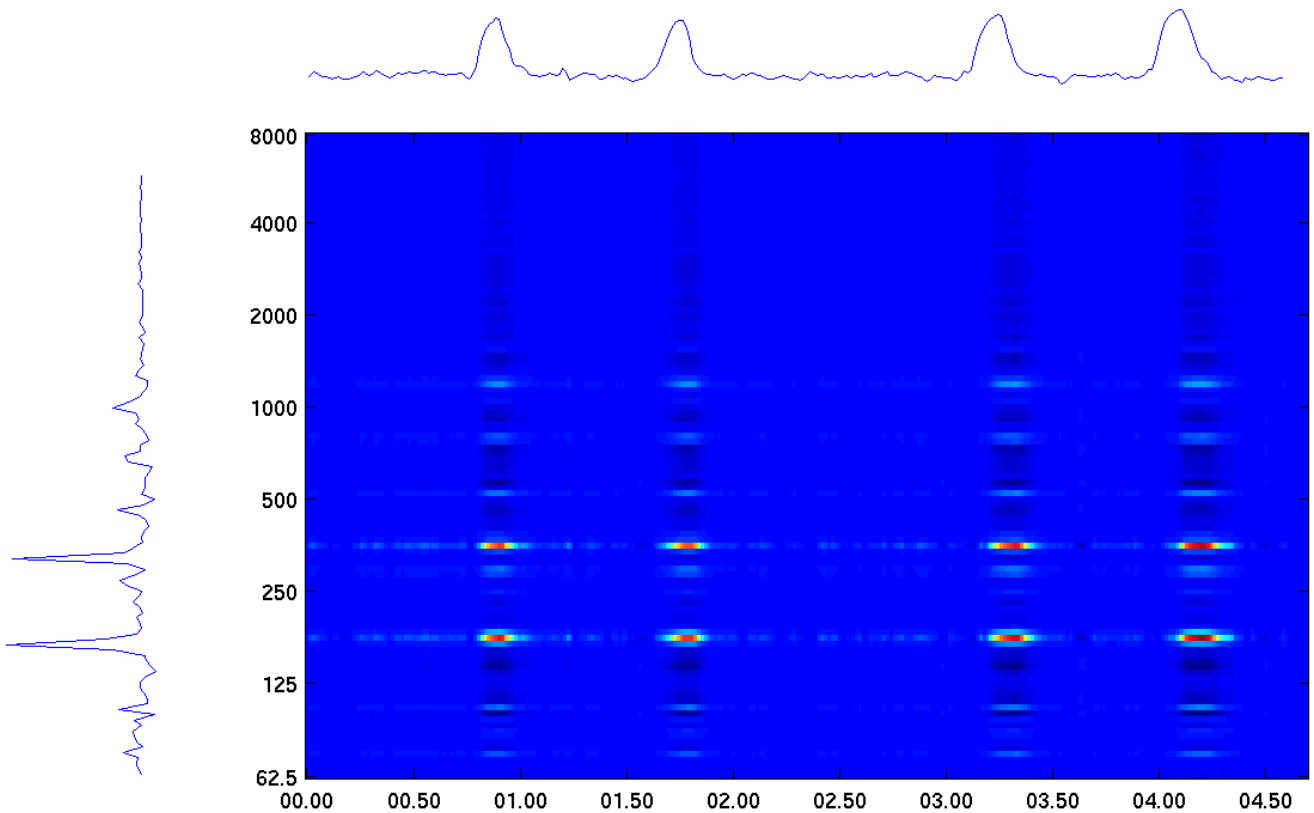
AudioSpectrumProjectionD



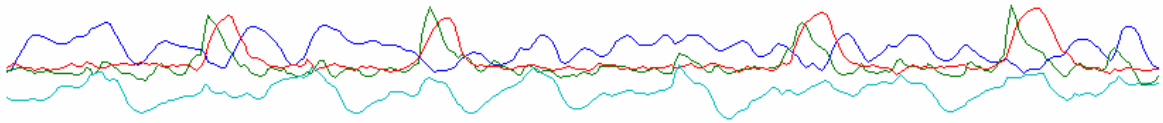
AudioSpectrumProjectionD

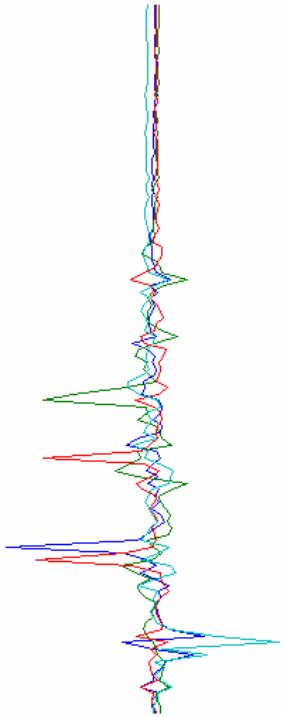
```
<?xml version="1.0" encoding="iso-8859-1"?>

<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
         xmlns:xml="http://www.w3.org/XML/1998/namespace"
         xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
        <DescriptionUnit xsi:type="mpeg7:AudioSpectrumProjectionType"
                loEdge="62.5"
                hiEdge="8000"
                octaveResolution="1/4">
   <SeriesOfVector totalNumOfSamples="42" vectorSize="11">
                        <Raw mpeg7:dim="42 11">
150.6518   -0.9894   -0.1050   -0.0792    0.0041    0.0033   -0.0168    0.0323    0.0300    0.0058    0.0136
153.5409   -0.9874   -0.1281   -0.0711   -0.0015    0.0405    0.0182    0.0044   -0.0207    0.0073    0.0081
151.8271   -0.9965   -0.0049   -0.0263   -0.0457   -0.0245    0.0159   -0.0218    0.0094   -0.0009    0.0061
159.2136   -0.9925    0.0722   -0.0613    0.0020   -0.0259    0.0551   -0.0069   -0.0194    0.0158    0.0294
166.5635   -0.9924    0.0904   -0.0299    0.0211    0.0127    0.0277   -0.0483    0.0134   -0.0231    0.0126
174.5434   -0.9921    0.0905   -0.0718    0.0175    0.0102   -0.0248    0.0015    0.0219   -0.0078    0.0064
174.9529   -0.9950    0.0730   -0.0370    0.0084   -0.0130    0.0017    0.0150    0.0029   -0.0199   -0.0158
180.2206   -0.9923    0.0924   -0.0395    0.0143    0.0284   -0.0486    0.0106   -0.0265    0.0124   -0.0013
182.9182   -0.9949    0.0643    0.0390   -0.0075    0.0476    0.0036   -0.0282    0.0119    0.0095   -0.0055
176.4337   -0.9964    0.0655   -0.0029   -0.0153    0.0180   -0.0219   -0.0049   -0.0061    0.0230   -0.0094
173.3415   -0.9978    0.0418   -0.0088   -0.0376   -0.0150    0.0044    0.0047    0.0031    0.0083   -0.0097
182.4516   -0.9967    0.0279   -0.0220   -0.0535    0.0242   -0.0084    0.0126   -0.0049   -0.0174   -0.0045
                                  <!-- More Values Here . . . -->
                        </Raw>
                </SeriesOfVector>
        </DescriptionUnit>
</Mpeg7>
```

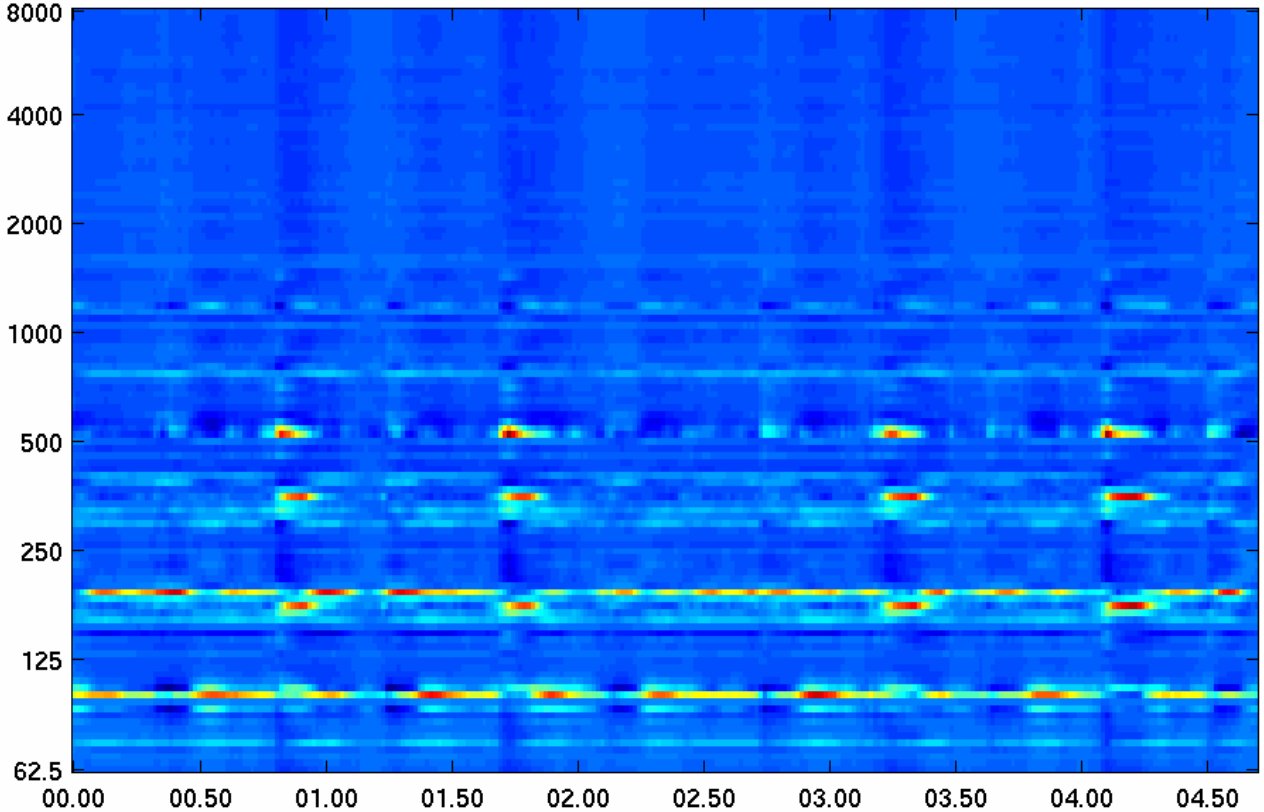
Outer Product Spectrum Reconstruction Individual Basis Component



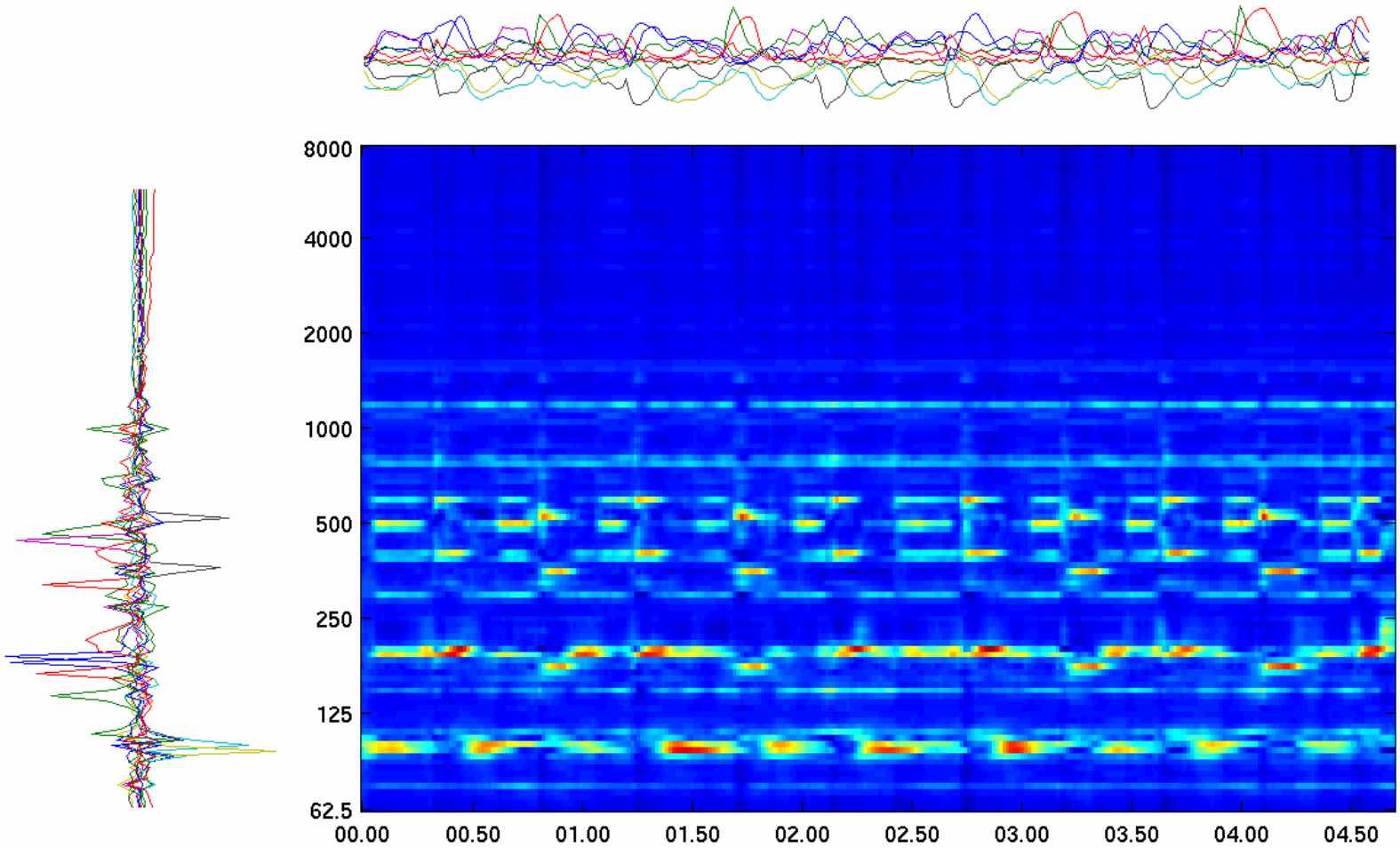
4 Component Reconstruction







10 Component Reconstruction



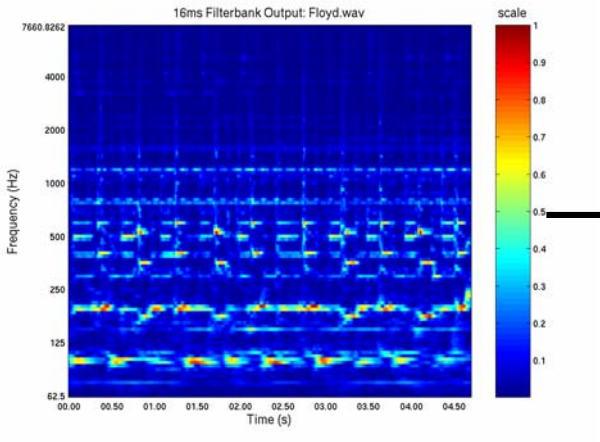
Music Unmixing

- Linear basis projection using SVD and ICA
  - spectrum subspace separation
  - fast computation of subspace ICA
  - full-rate filterbank masking

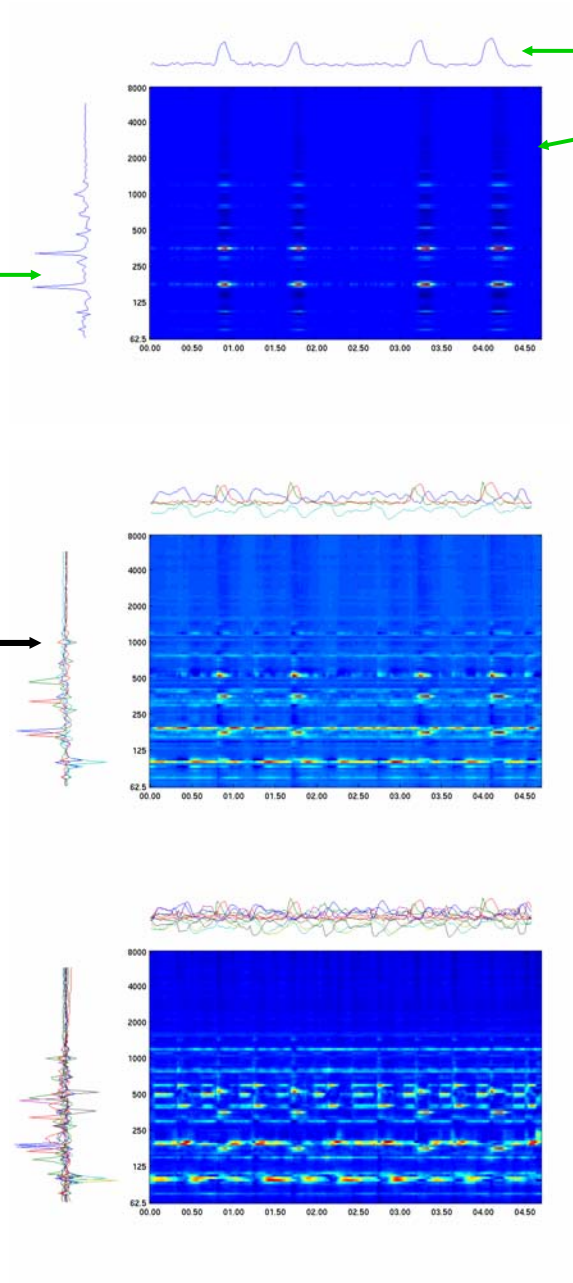
- Blocked ICA functions
  - subspace reconstruction $Y_j = XV_jV_j^+$
  - cluster subspaces to identify “tracks”
  - sum masked filterbank output to create audio

Independent Spectrogram Subspace Layers

Mixture Spectrogram

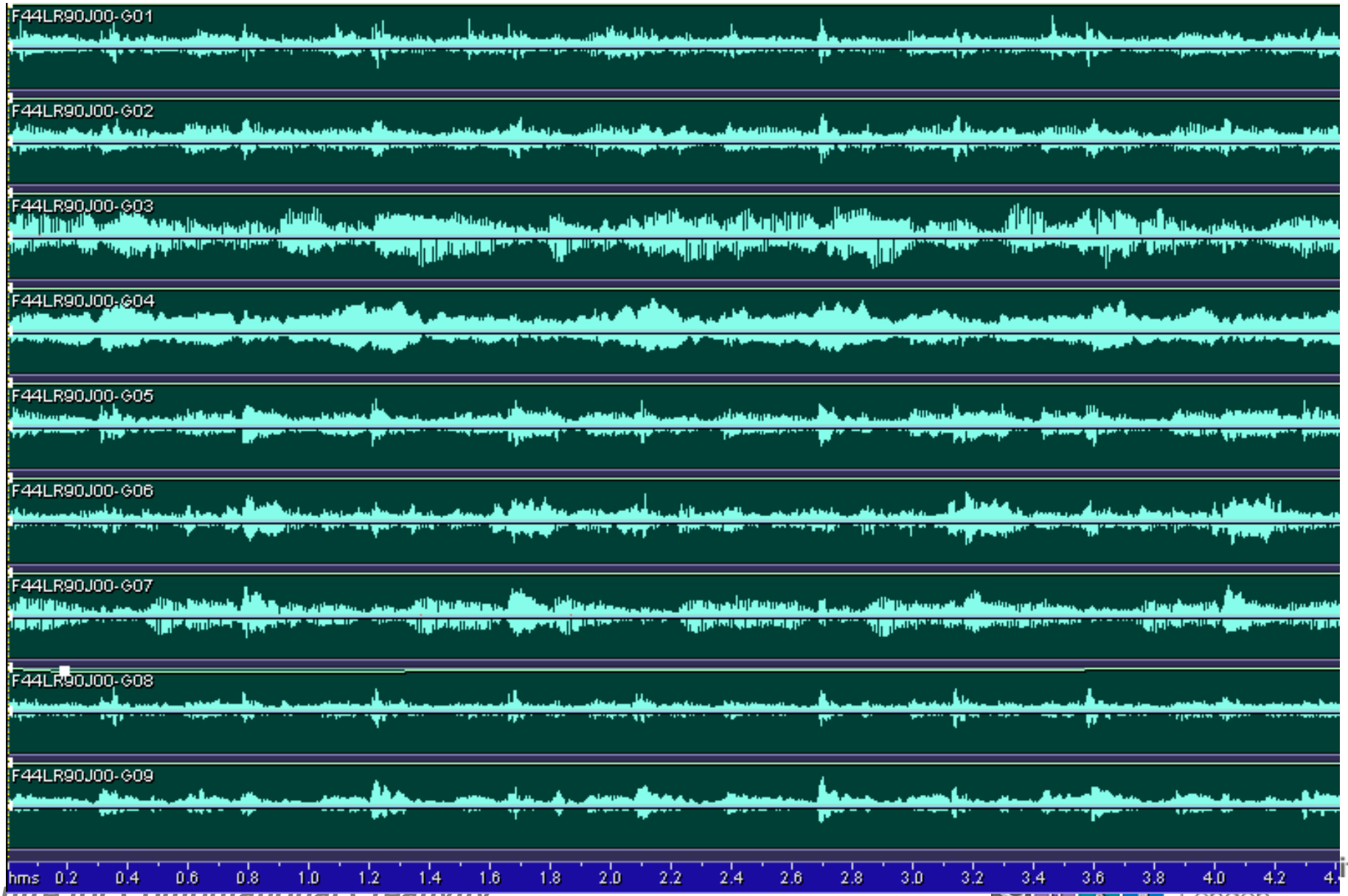


Subspace Extraction



Music Unmixing Example

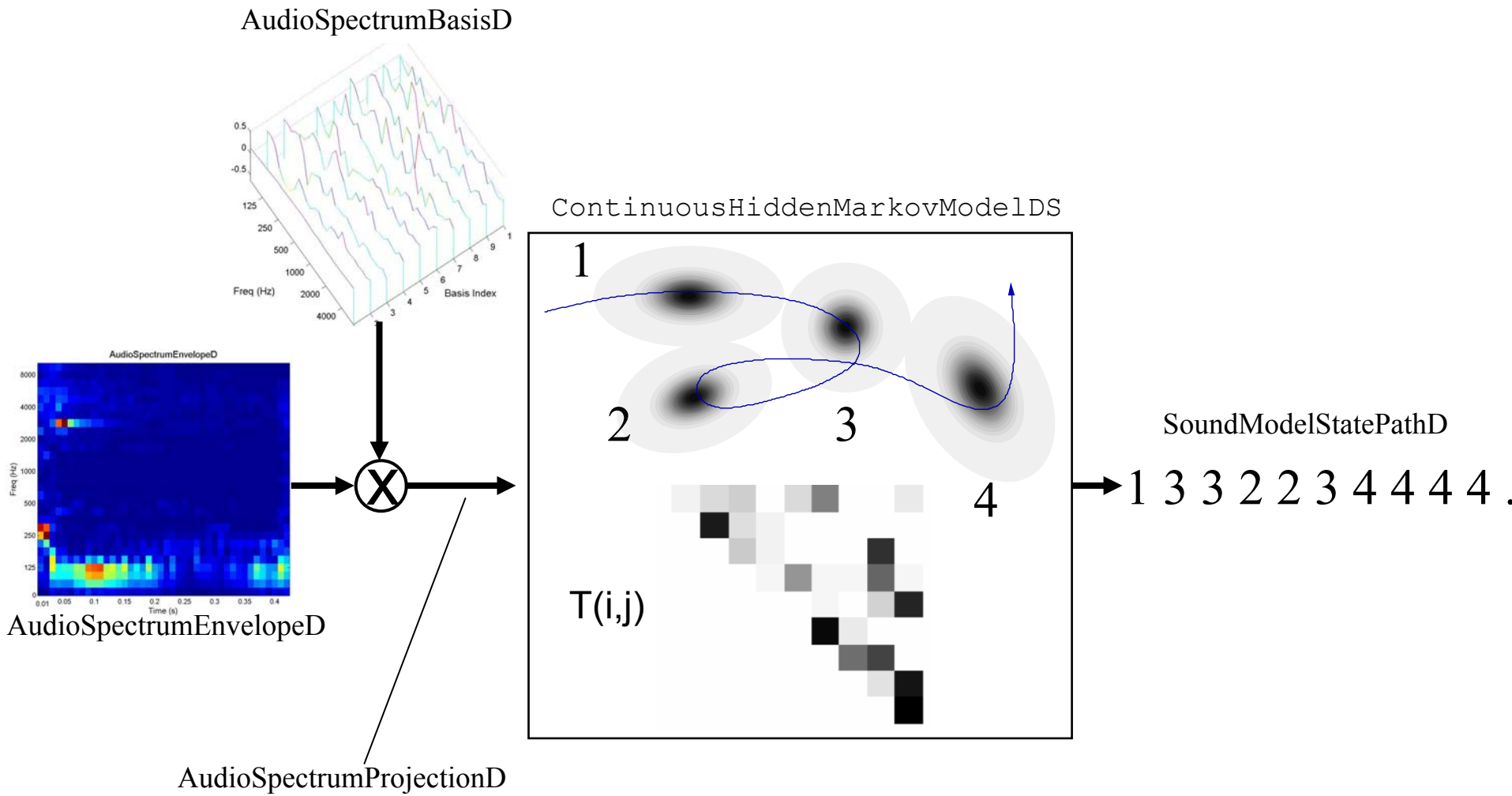
(Pink Floyd: mono -> 9 subspace tracks)



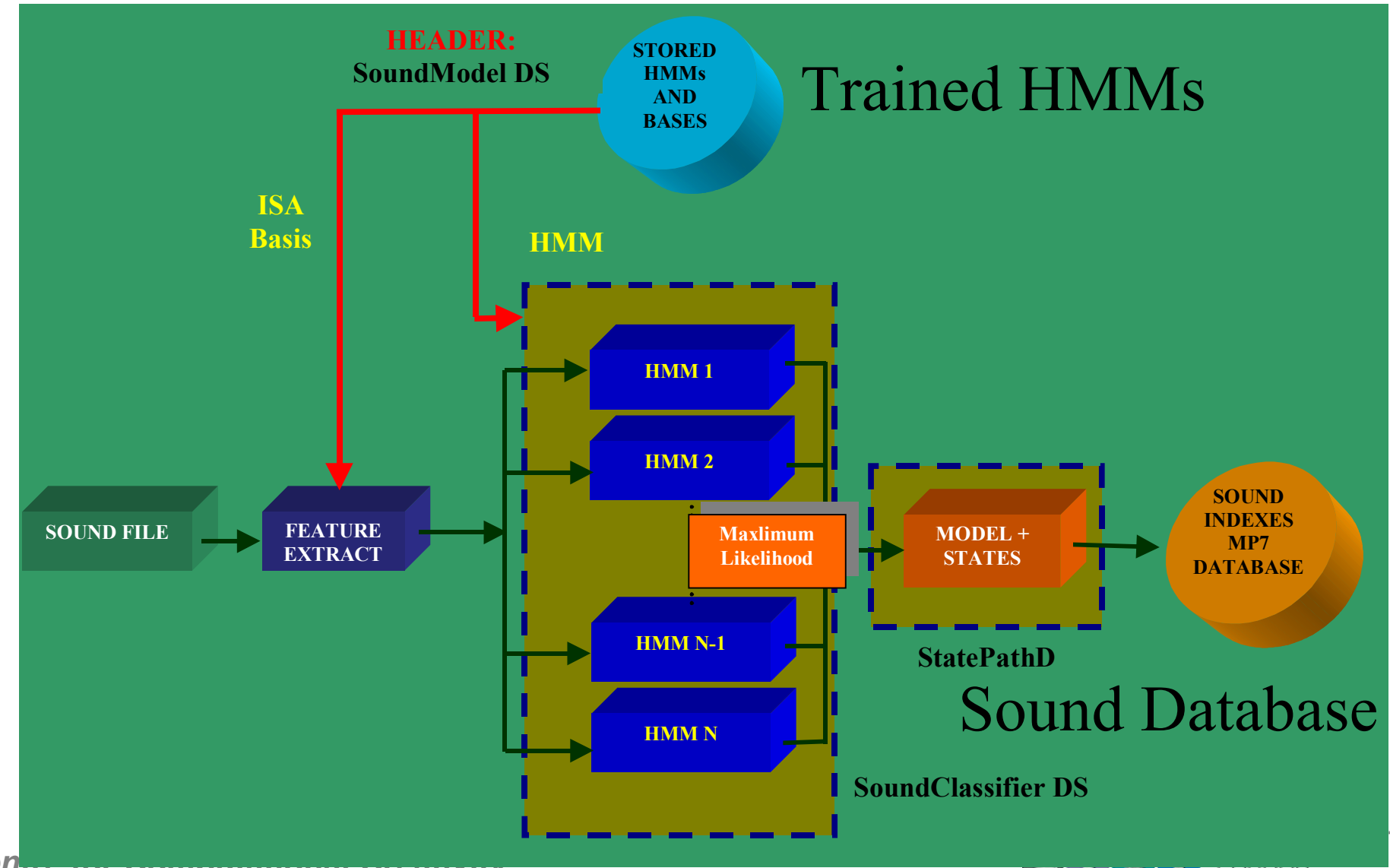
EXAMPLE 2

AUTOMATIC AUDIO CLASSIFICATION

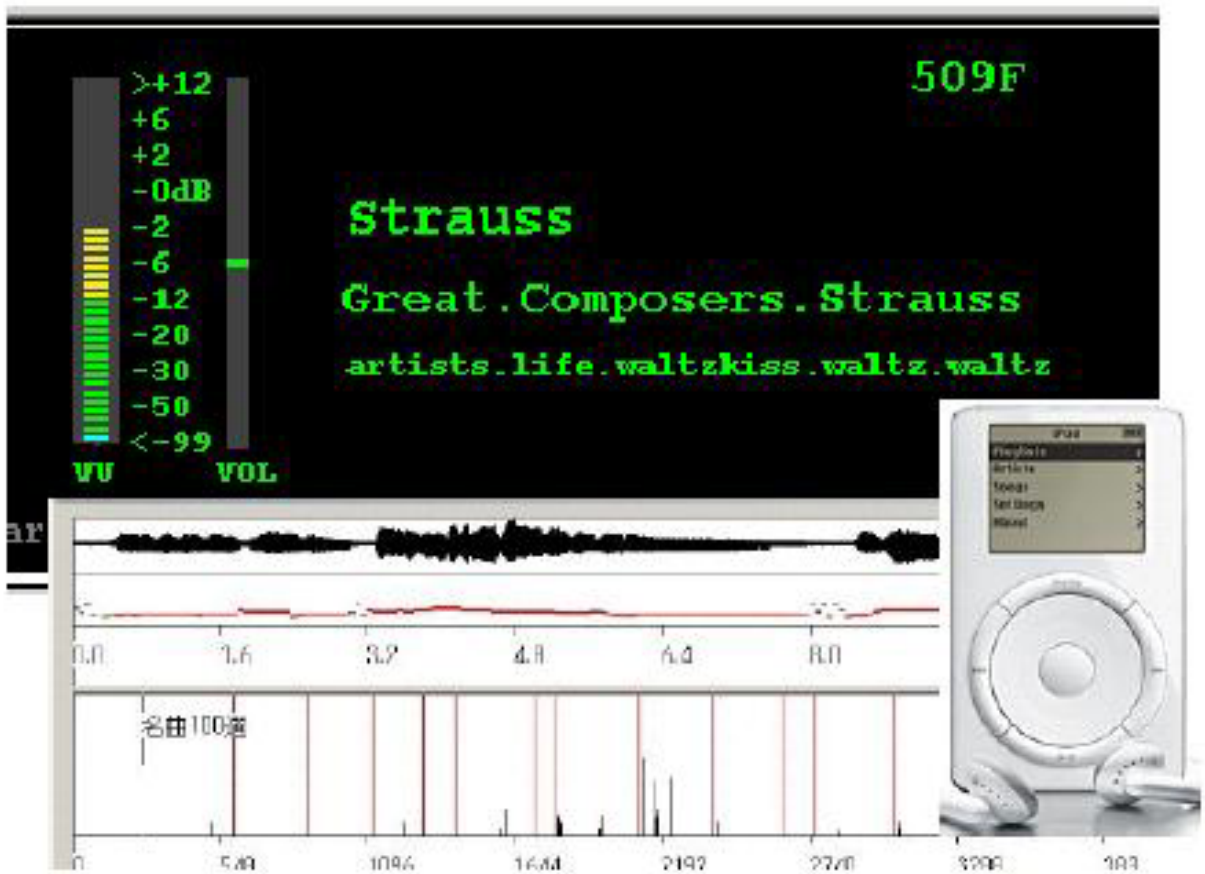
Sound Model DS and related descriptors



Sound Recognition using HMMs



MPEG-7:
Intelligent Music Browsing



Music Genre Classification:

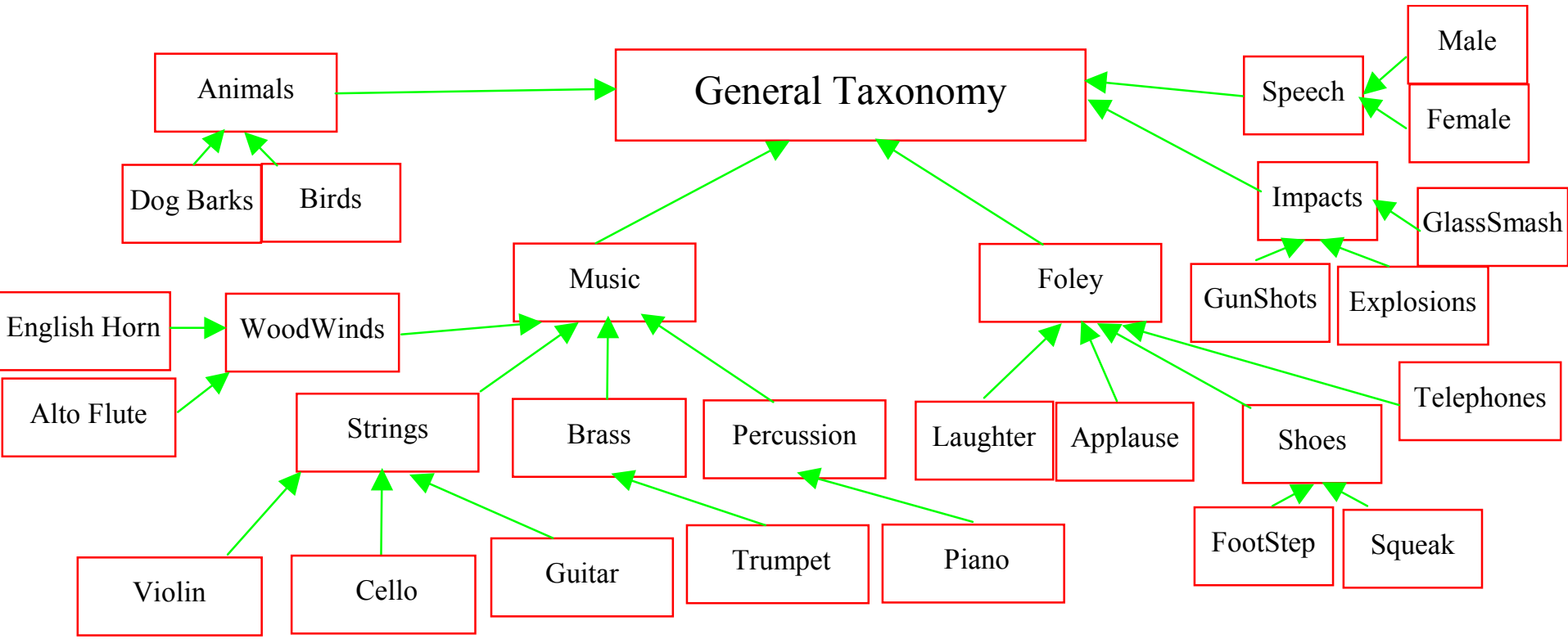
Class Name		Num of Files	Num Segments
1)	Blues	79	86
2)	hiphop	15	129
3)	Gospel	23	25
4)	Country	27	28
5)	DrumNBass	26	275
6)	Classical	8	156
7)	2Step	39	311
8)	Merengue	34	304
9)	Reggae	80	398
10)	Salsa	39	425
Totals		370	2137

Music Genre Classification

Sound / Class		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Blues	[1]	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Salsa	[2]	0.000	**0.833**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**0.167**
Gospel	[3]	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Country	[4]	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
HipHop	[5]	0.000	0.000	0.000	0.000	**0.933**	0.000	**0.067**	0.000	0.000	0.000
Reggae	[6]	0.000	**0.042**	0.000	0.000	0.000	**0.958**	0.000	0.000	0.000	0.000
2Step	[7]	0.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000	0.000
Classical	[8]	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**1.000**	0.000	0.000
DrumNBass	[9]	0.000	0.000	0.000	0.000	0.000	**0.125**	0.000	0.000	**0.625**	**0.250**
Merengue	[10]	0.000	**0.100**	0.000	0.000	**0.100**	0.000	0.000	0.000	0.000	**0.800**

Mean Recognition rate = 91.500

Semantic Audio: General Sound Taxonomy

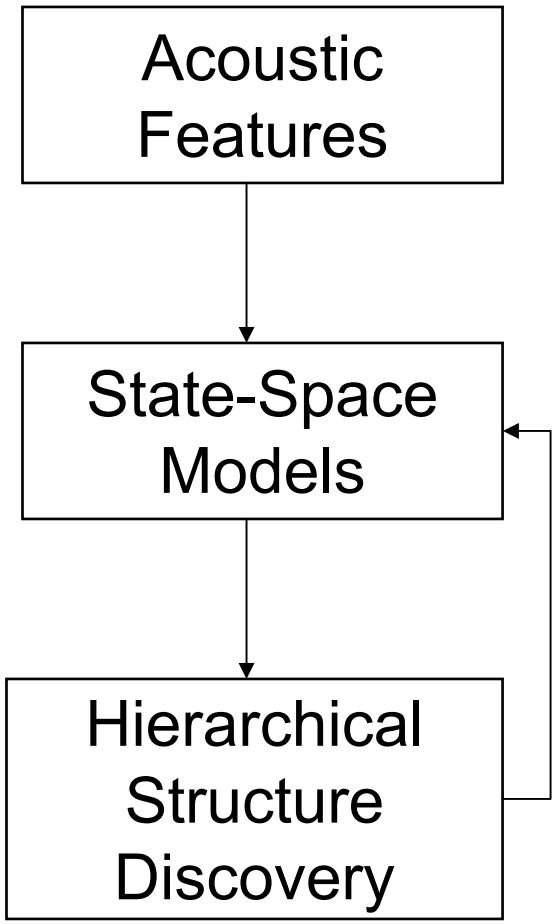


DS: General Audio Classification

	Conventional Training		Entropic Training	
Class	**# Hit**	**# Miss**	**# Hit**	**# Miss**
Speech:Female	39	1	39	1
Speech:Male	69	1	68	2
Music	12	0	12	0
Bird Calls	9	6	12	3
Applause	6	0	5	1
DogBarks	14	1	15	0
Explosions	6	1	7	0
FootSteps	11	0	10	1
Glass Smash	8	5	12	1
Gunshots	13	0	12	1
Shoe Squeaks	2	2	4	0
Laughter	13	5	17	1
Telephones	14	4	12	6
Flute	2	2	4	0
Piano	3	2	5	0
Cello	5	1	6	0
Cor Anglais	2	2	4	0
Guitar	1	2	3	0
Trumpet	3	2	4	1
Violin	6	0	5	1
Totals	**238**	**37**	**256**	**19**
Performance	**86.55%**		**93.09%**	

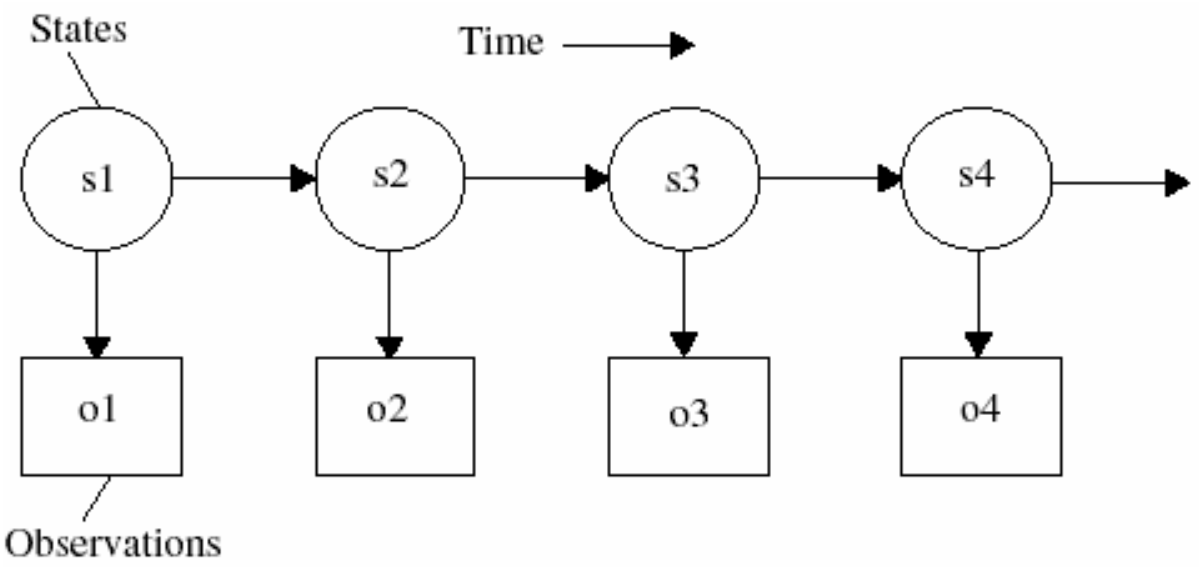
EXAMPLE 3
STRUCTURE EXTRACTION

Structure Discovery

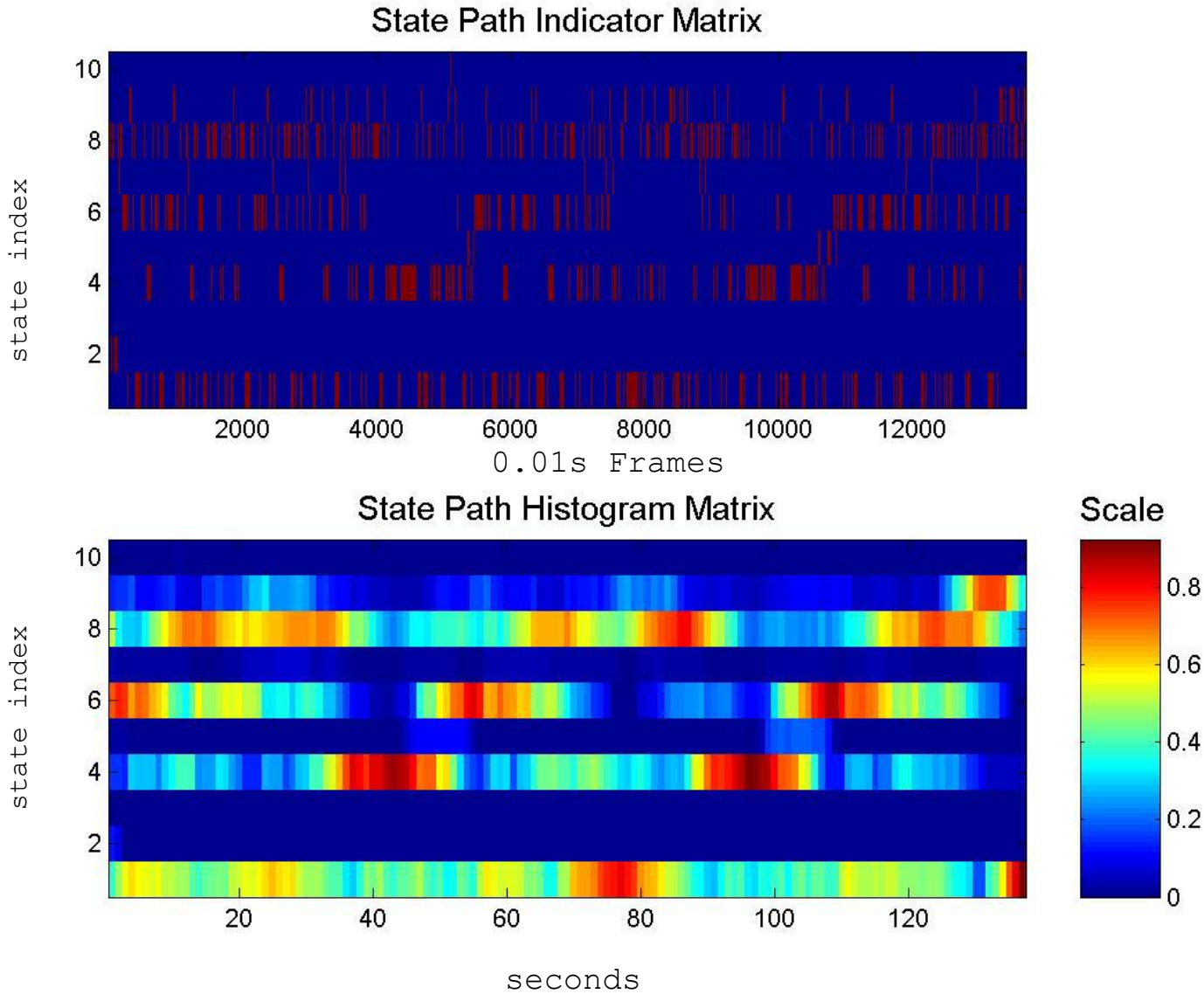


SoundModelStatePathD

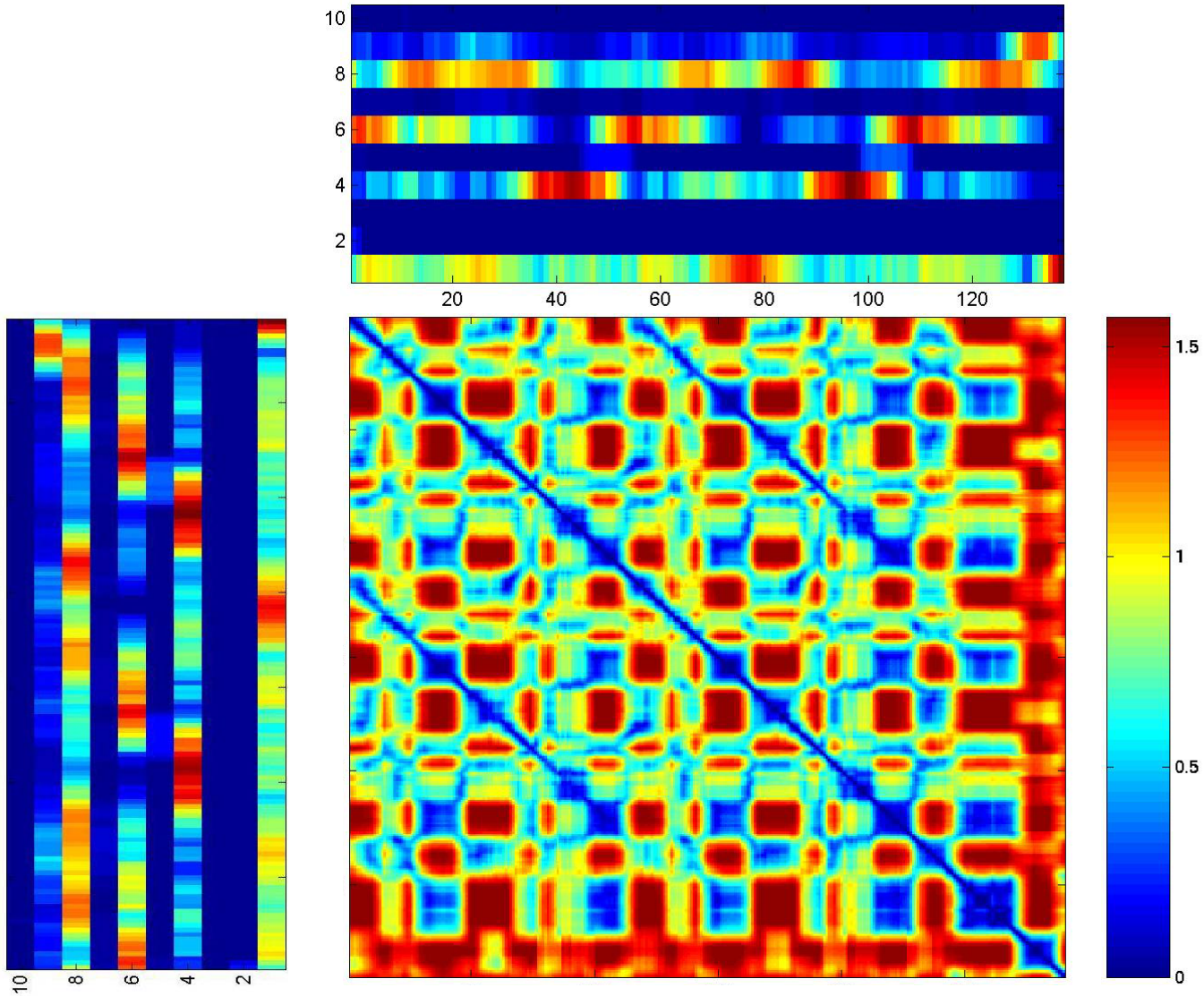
A simplified representation of spectral dynamics



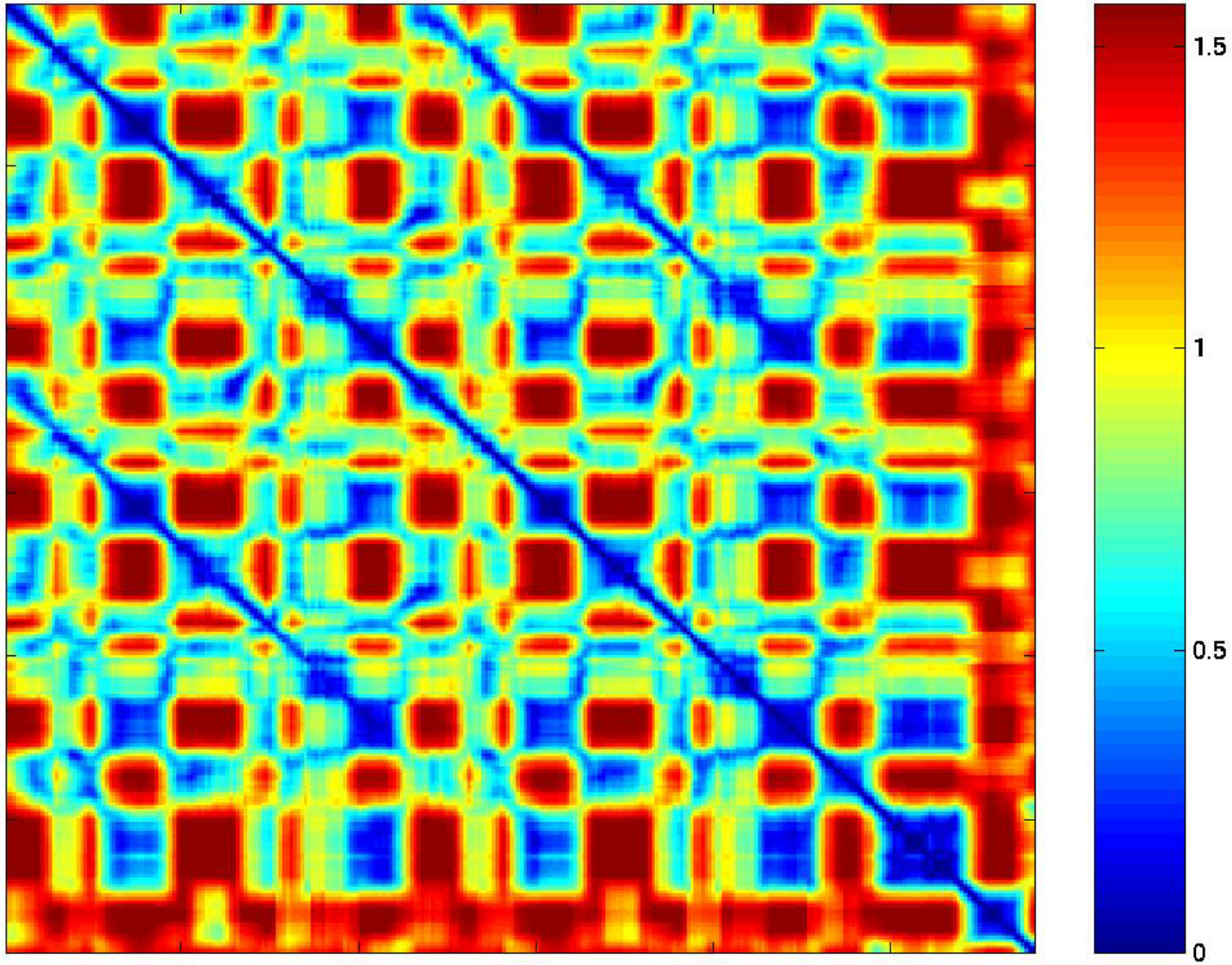
SoundModelStateHistogramD



High-Level Structure Discovery



S-Matrix



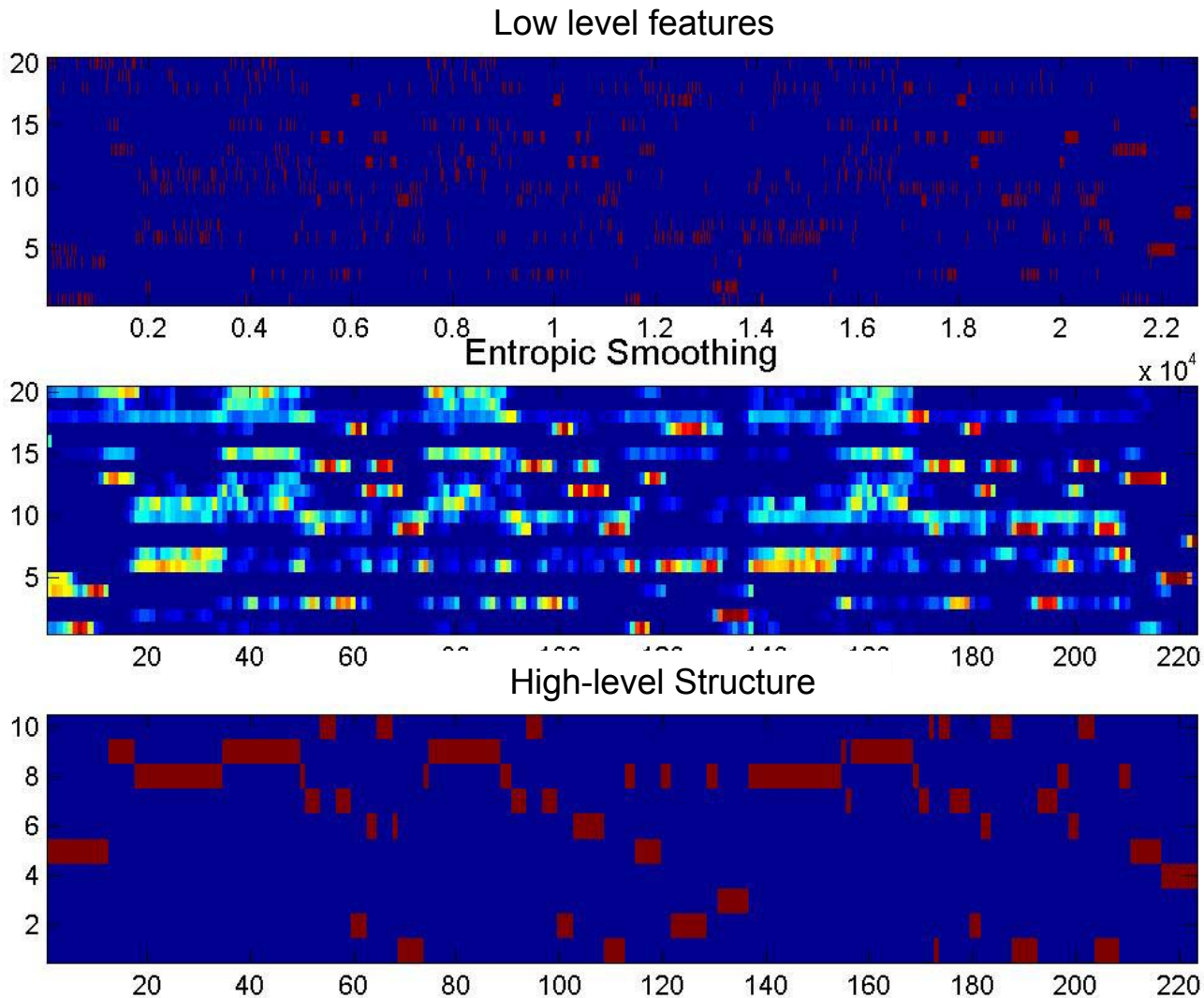
STRUCTURE EXTRACTION == SEGMENTATION

Structure Discovery

Acoustic Features

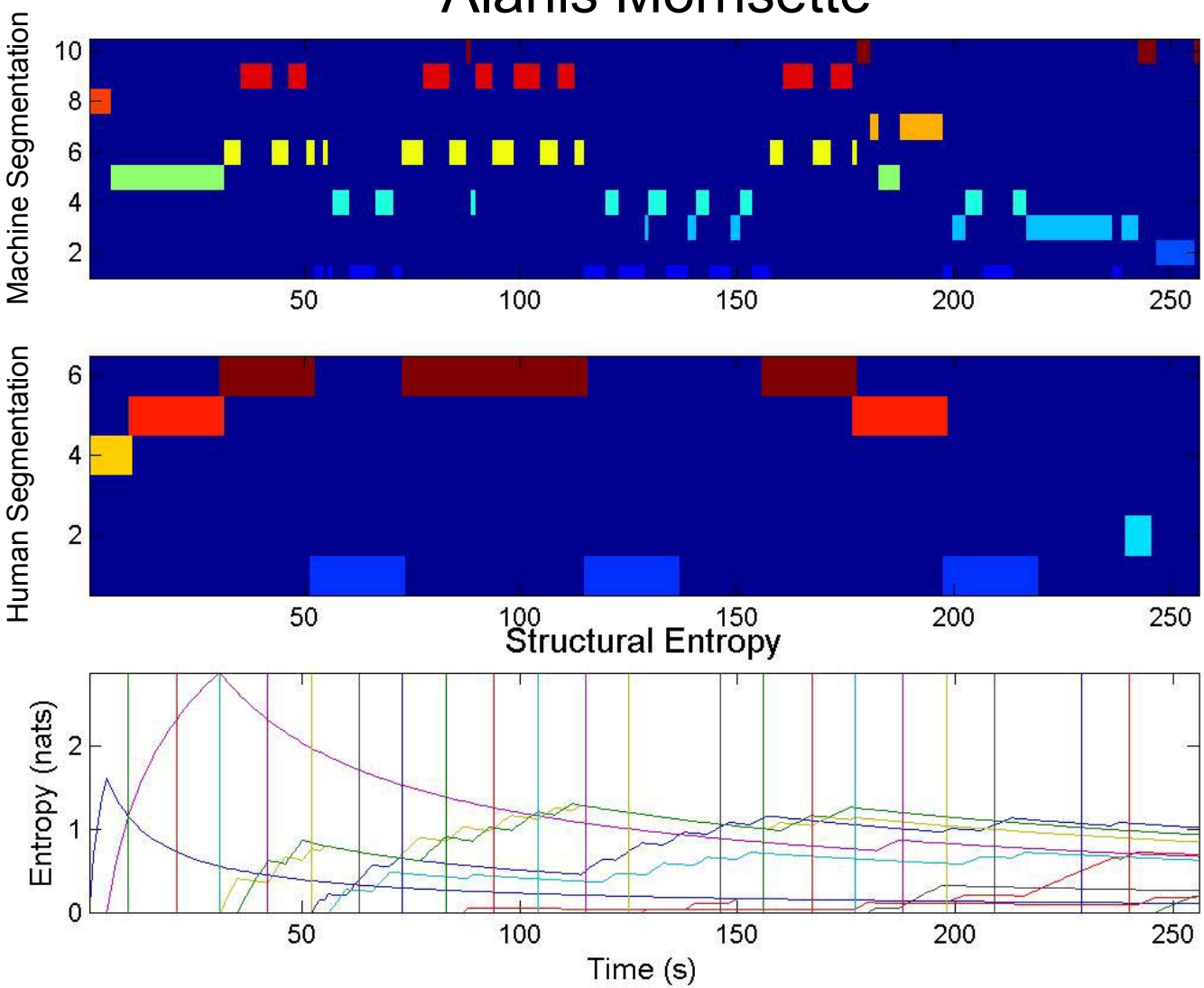
State-Space Models

Hierarchical Structure Discovery



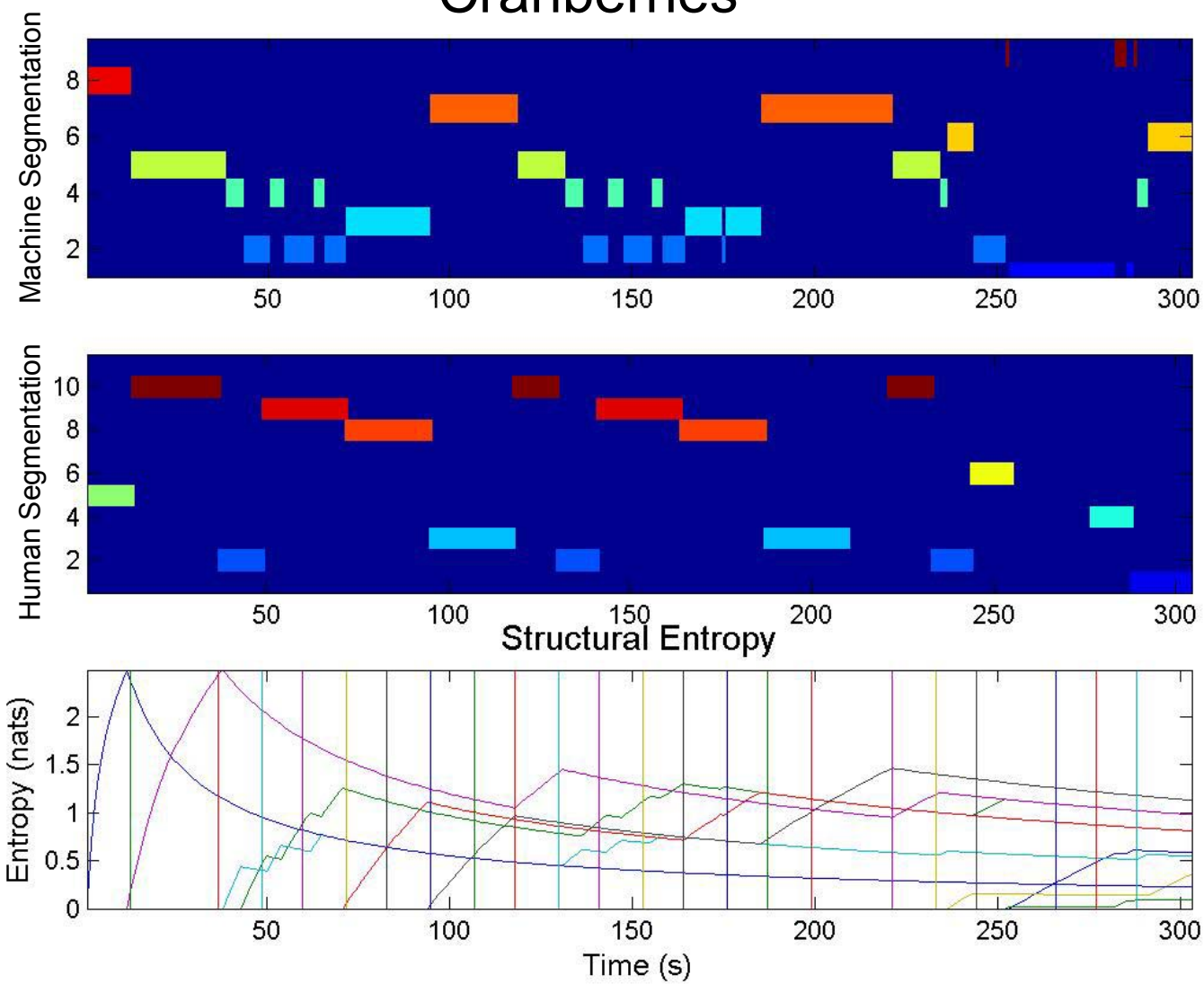
High-Level Structure Discovery

Alanis Morrisette



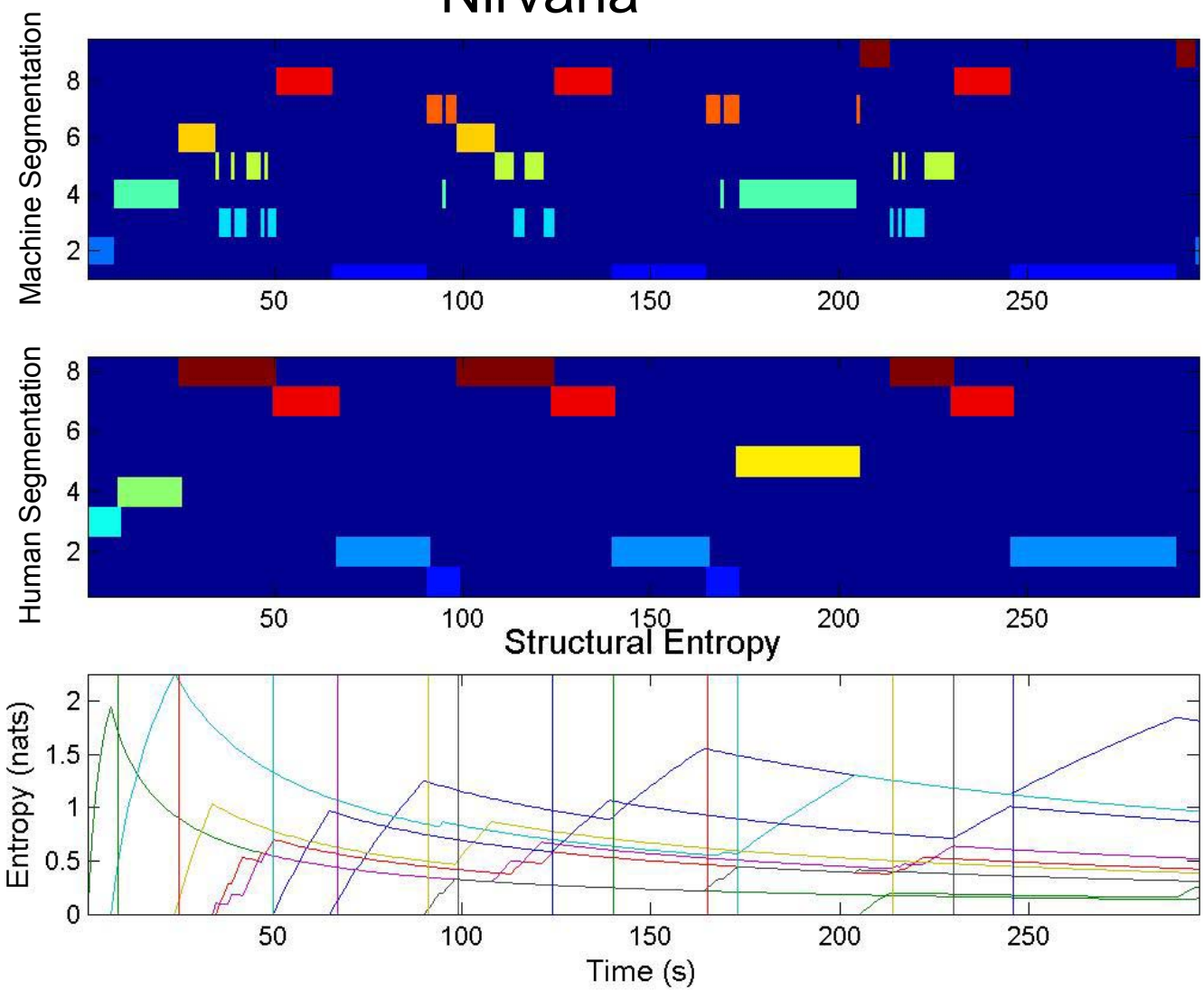
High-Level Structure Discovery

Cranberries

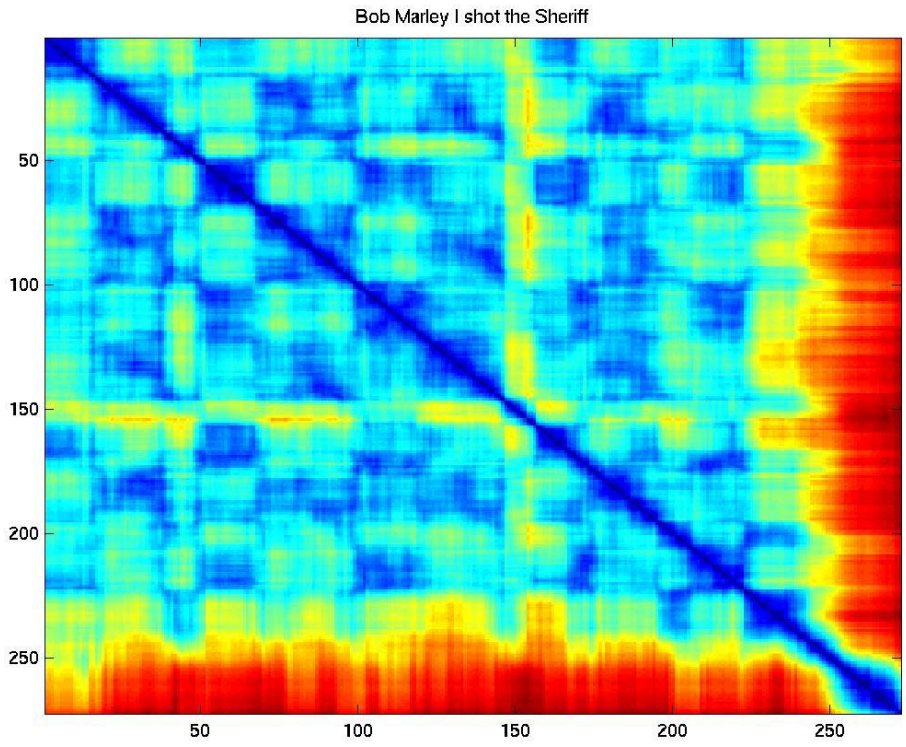


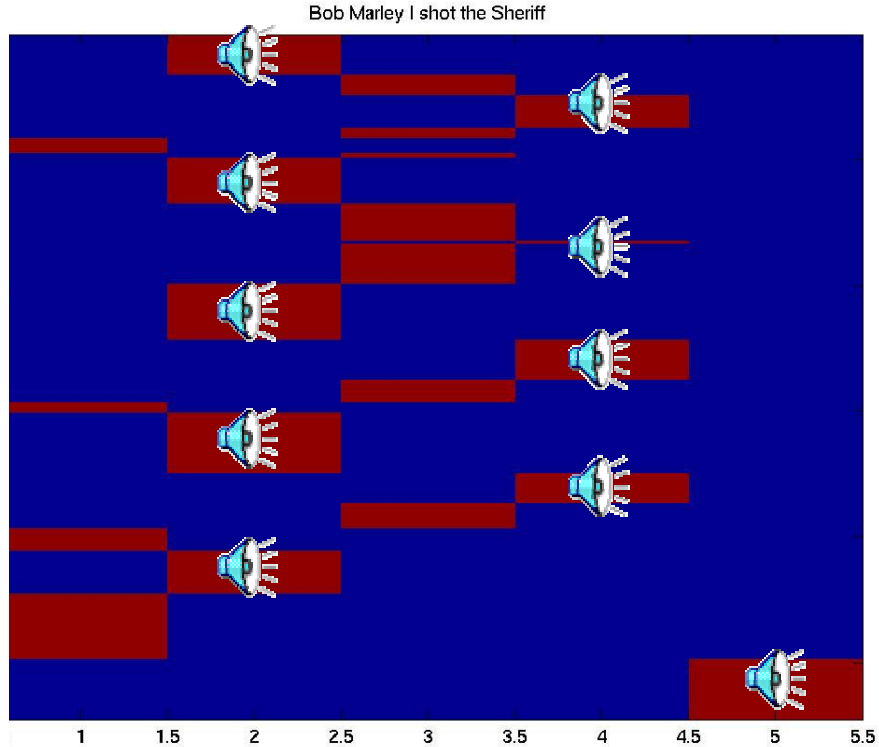
High-Level Structure Discovery

Nirvana



High-Level Structure Discovery





EXAMPLE 4

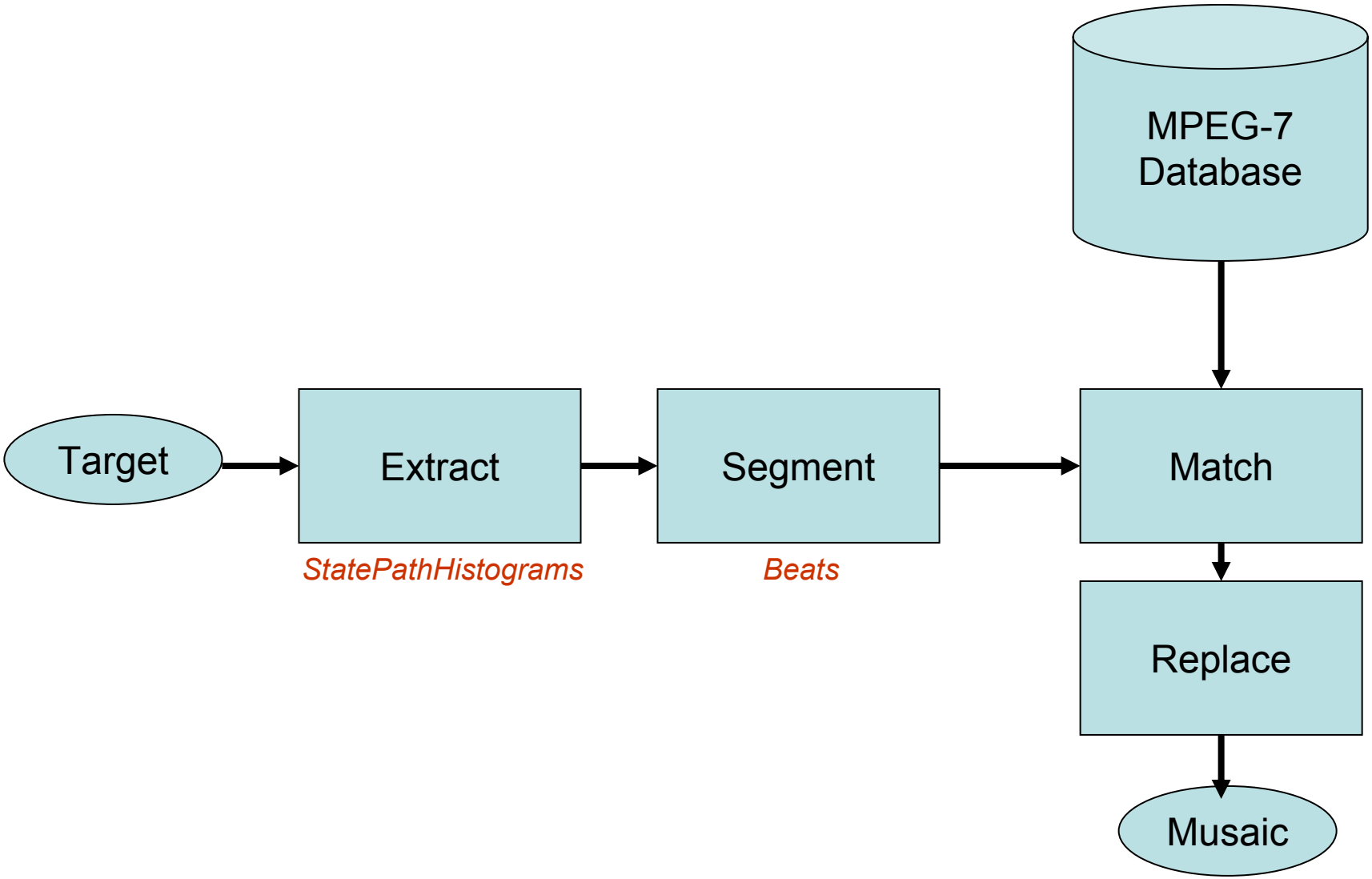
MUSAICS

Musaics

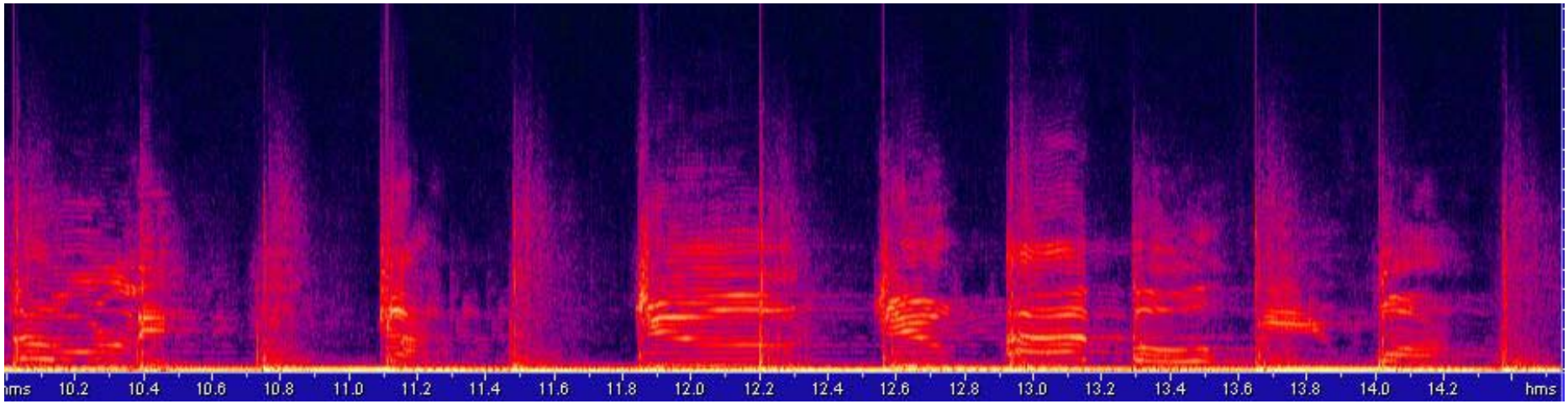
(*Music Mosaics*)

- C-Matrix : Cross-Song Similarity Matrix
  - Outer product of target and source histograms
- Find segments similar to target segment
  - Similarity between all target and database segments
  - SORT columns of similarity matrix
- Replace segments with similar material
  - Segmentation boundaries (beat alignment)
  - Replace with “best fit” using DTW on most similar segments
- EXAMPLES

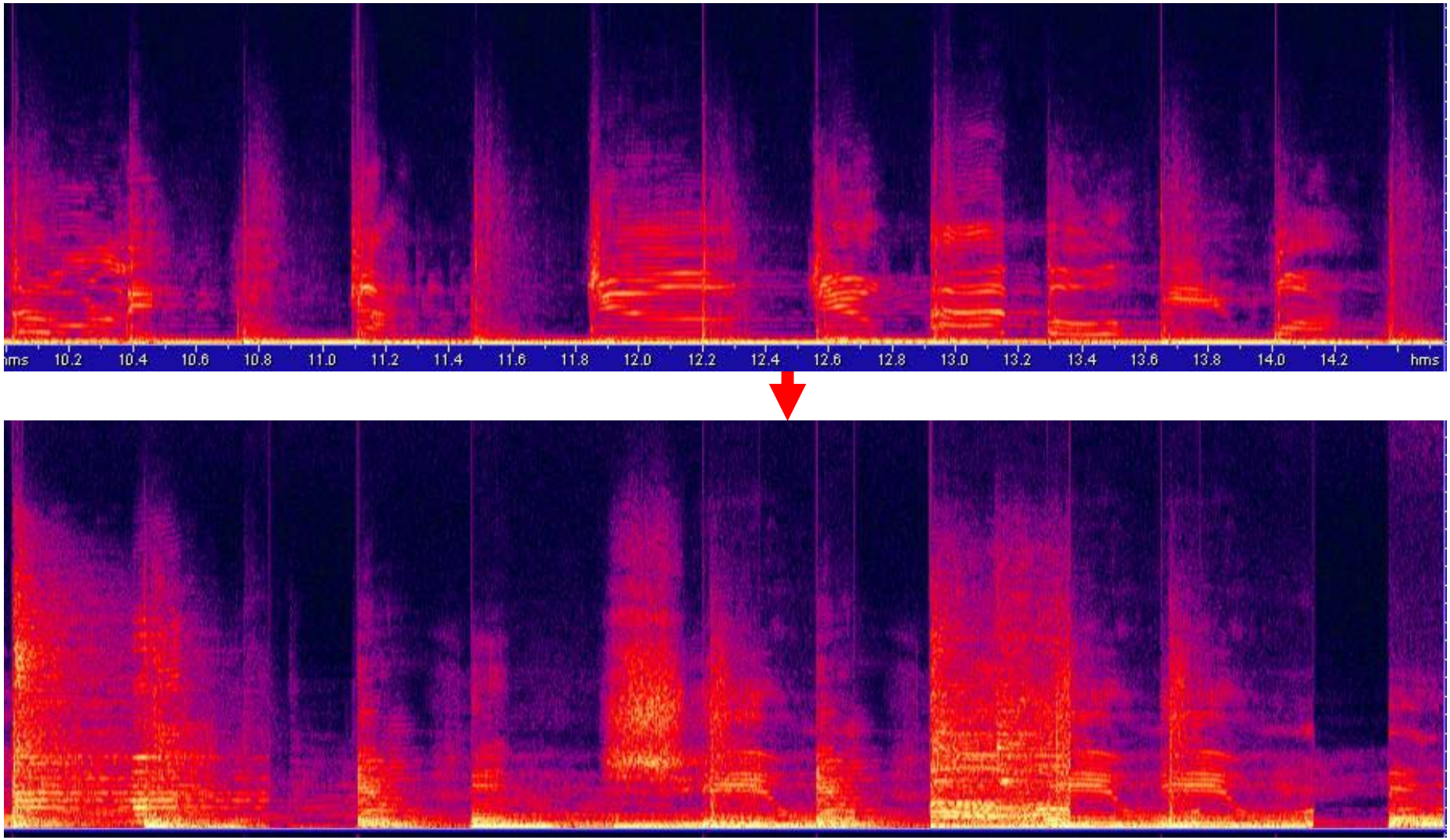
Musaics


MPEG-7
Database
Target
Extract
Segment
Match
StatePathHistograms
Beats
Replace
Musaic

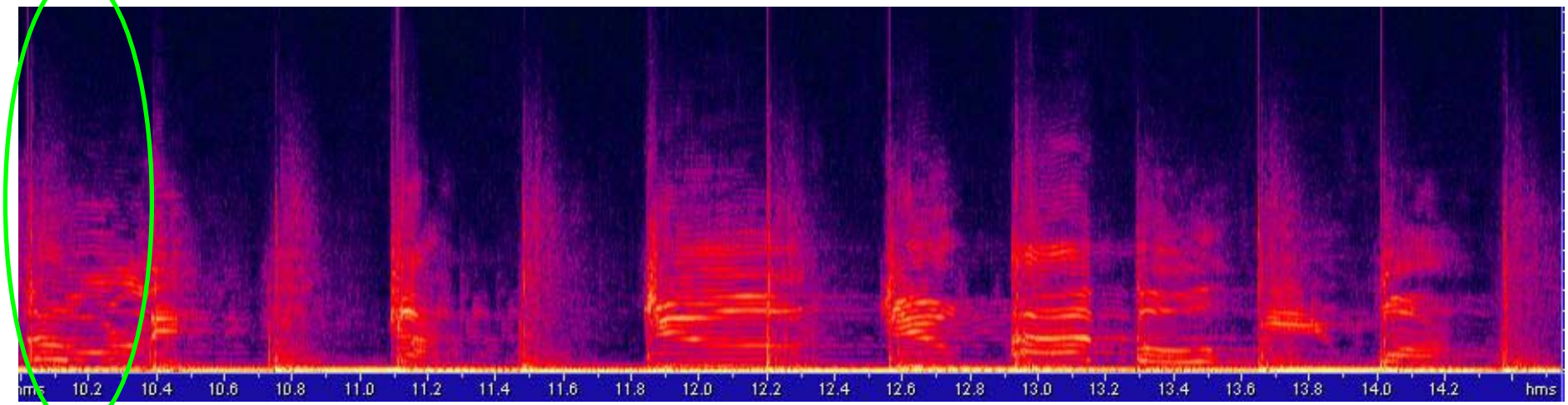
Musaics

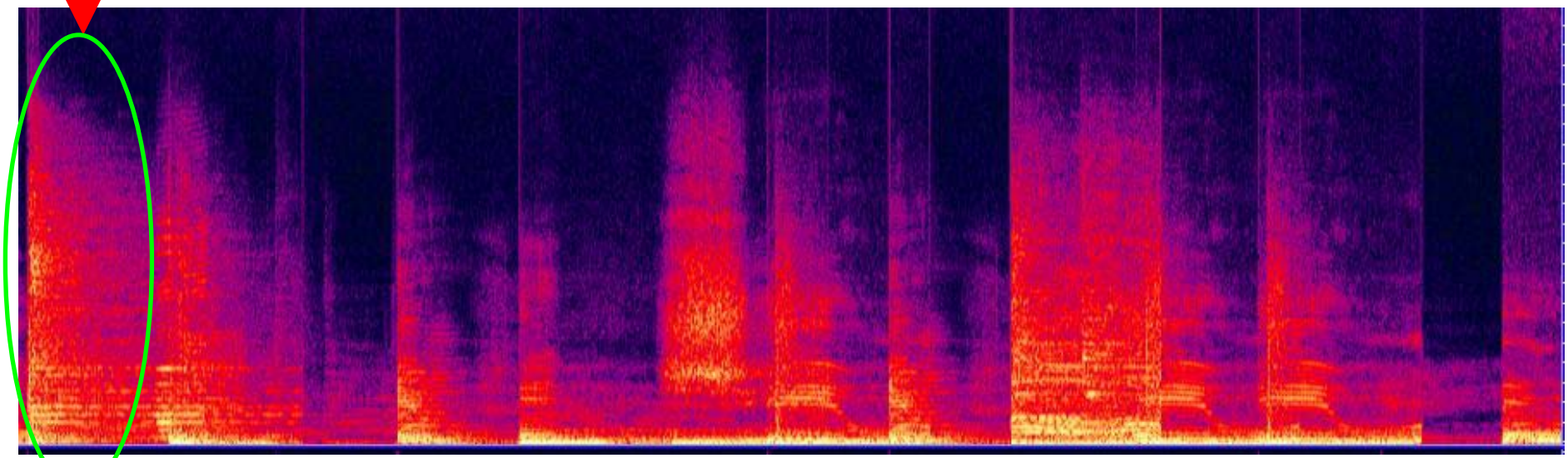

hms
10.2
10.4
10.6
10.8
11.0
11.2
11.4
11.6
11.8
12.0
12.2
12.4
12.6
12.8
13.0
13.2
13.4
13.6
13.8
14.0
14.2
hms

Musaics

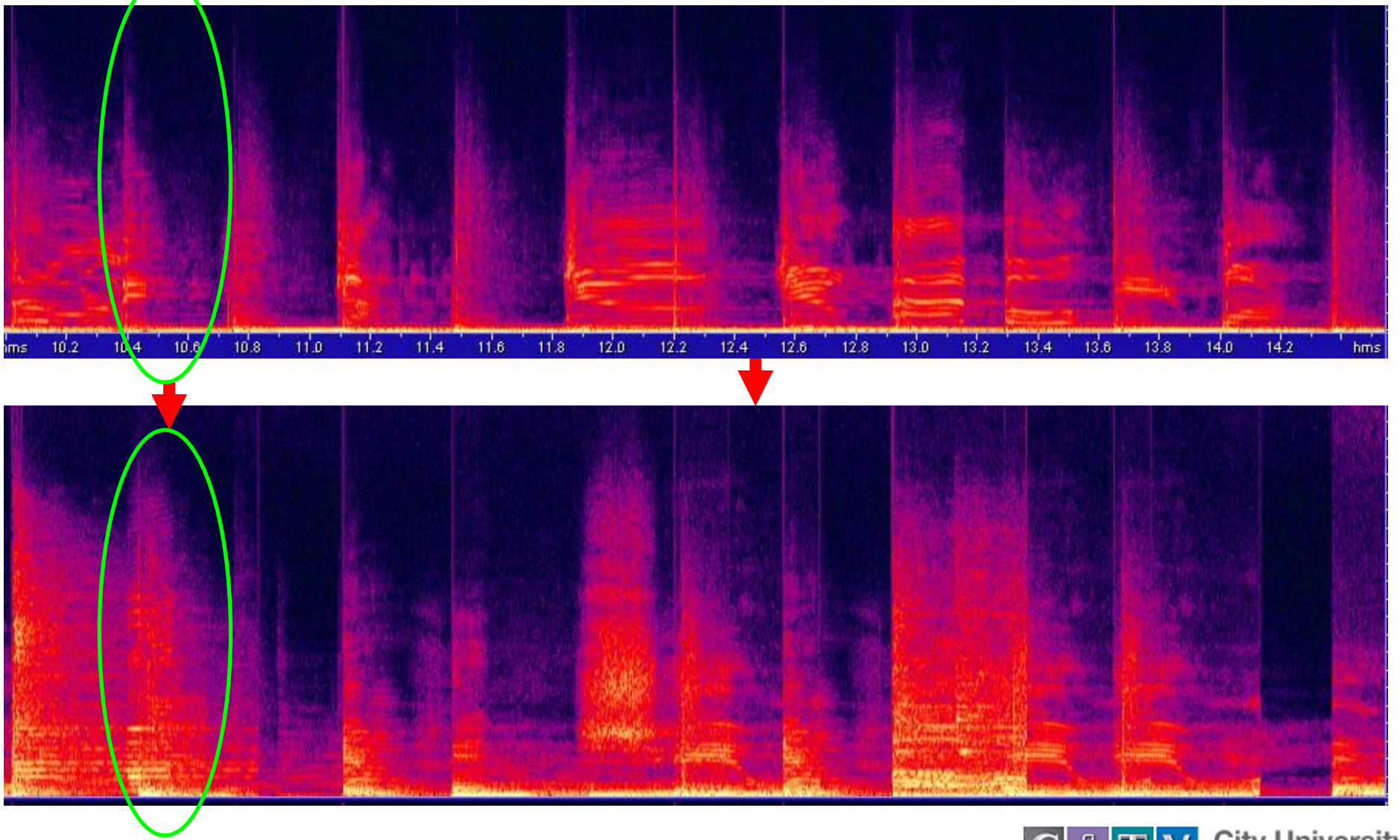

hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

Musaics

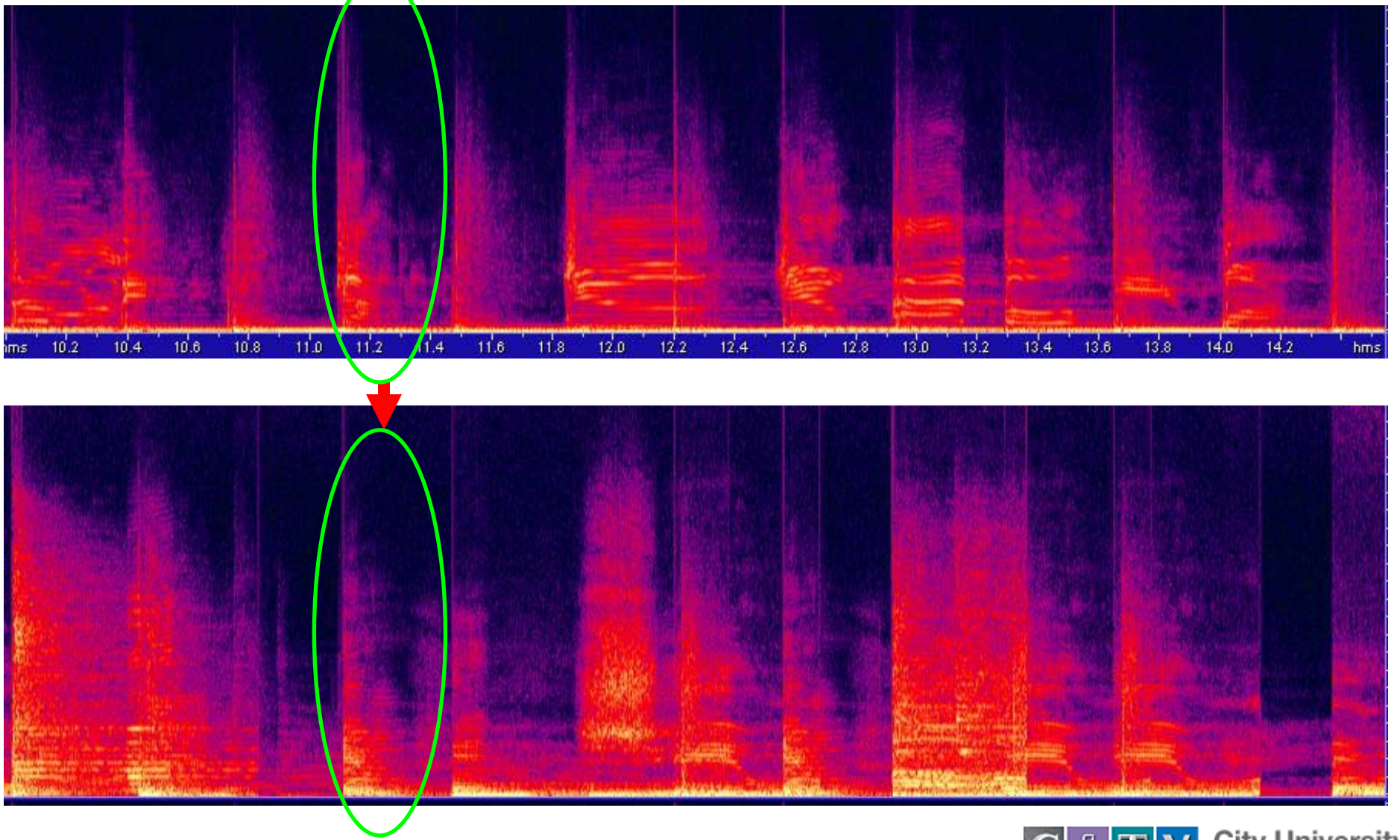

10.2
10.4
10.6
10.8
11.0
11.2
11.4
11.6
11.8
12.0
12.2
12.4
12.6
12.8
13.0
13.2
13.4
13.6
13.8
14.0
14.2
hms



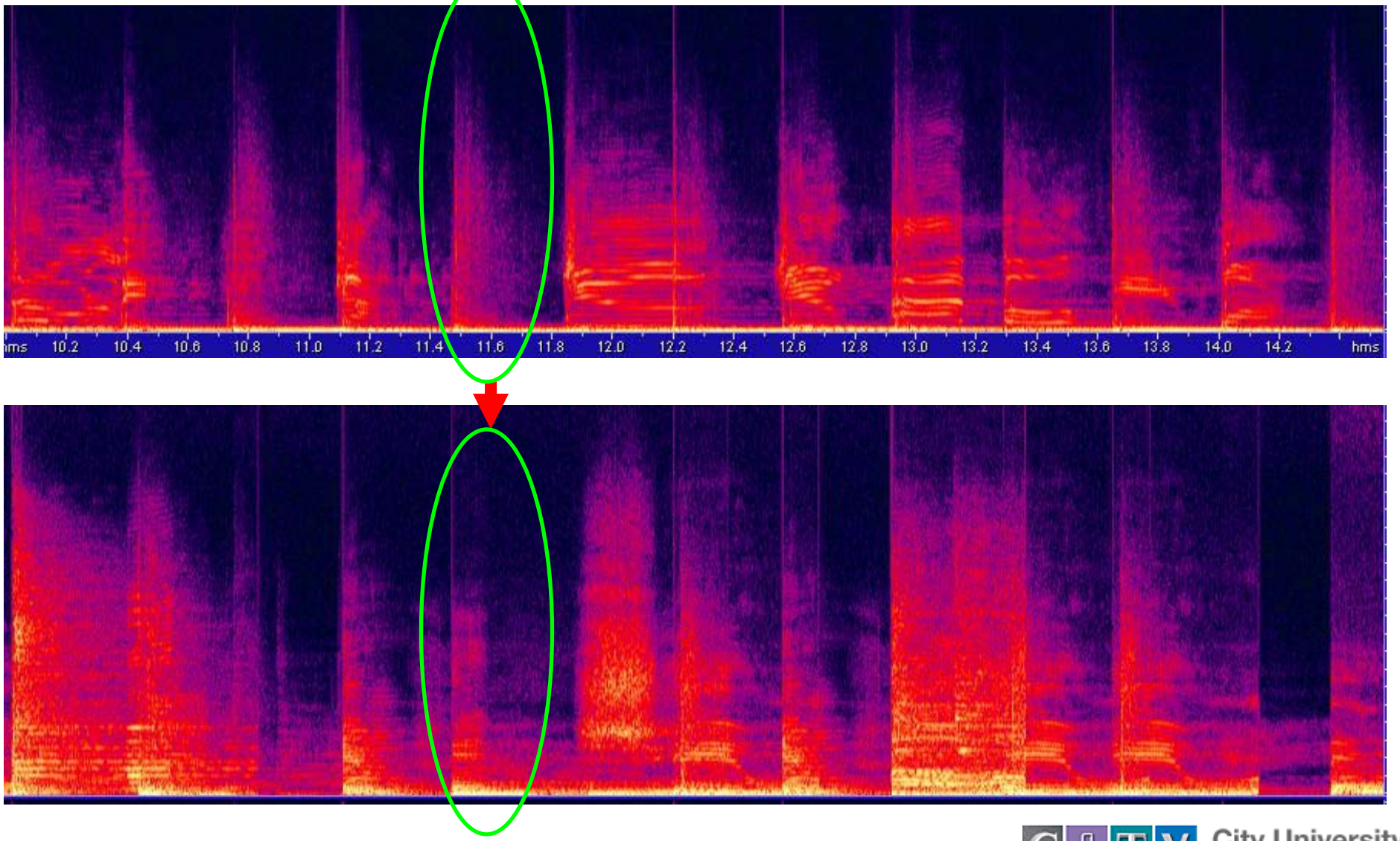
Musaics


hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

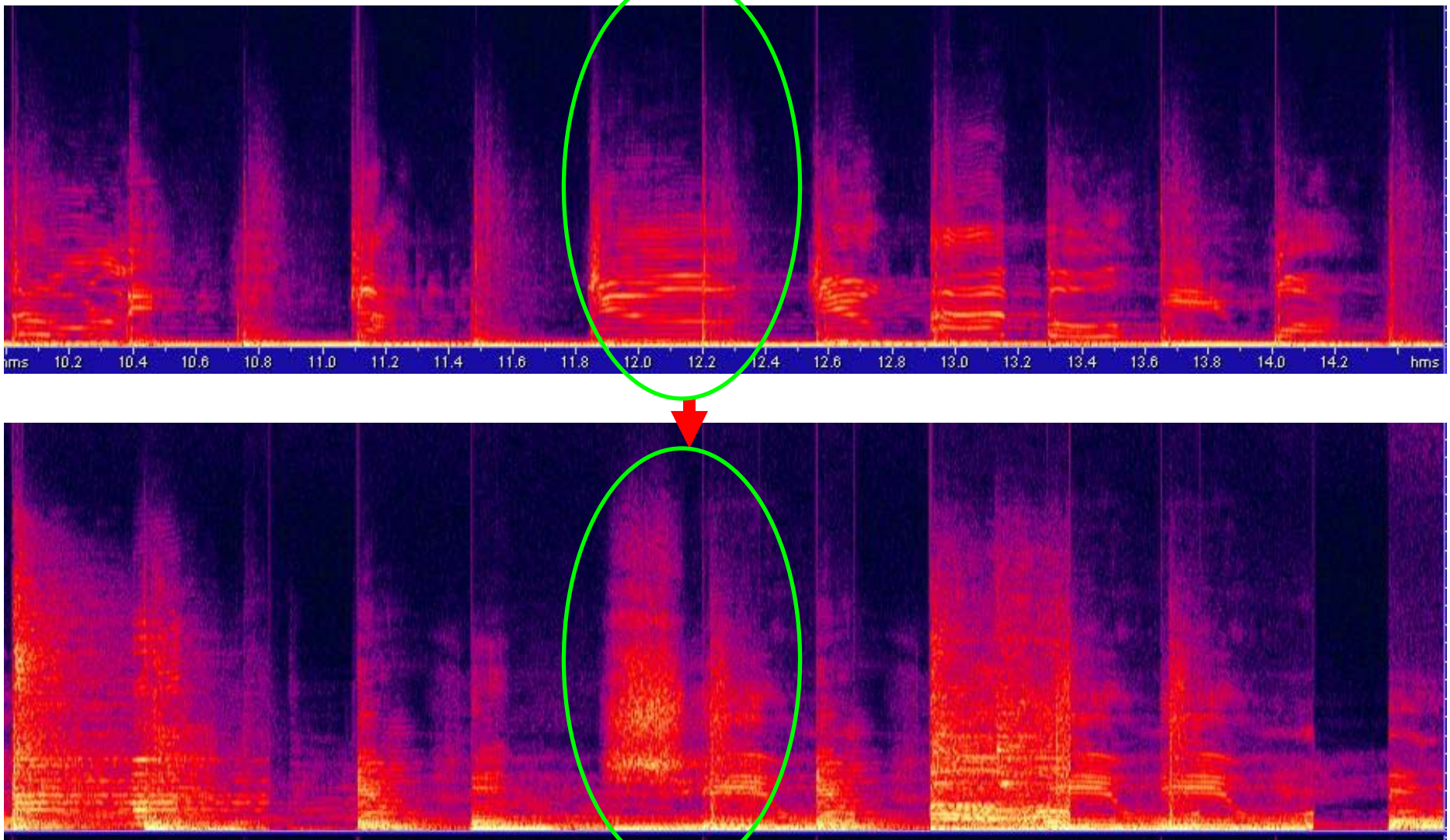
Musaics


hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

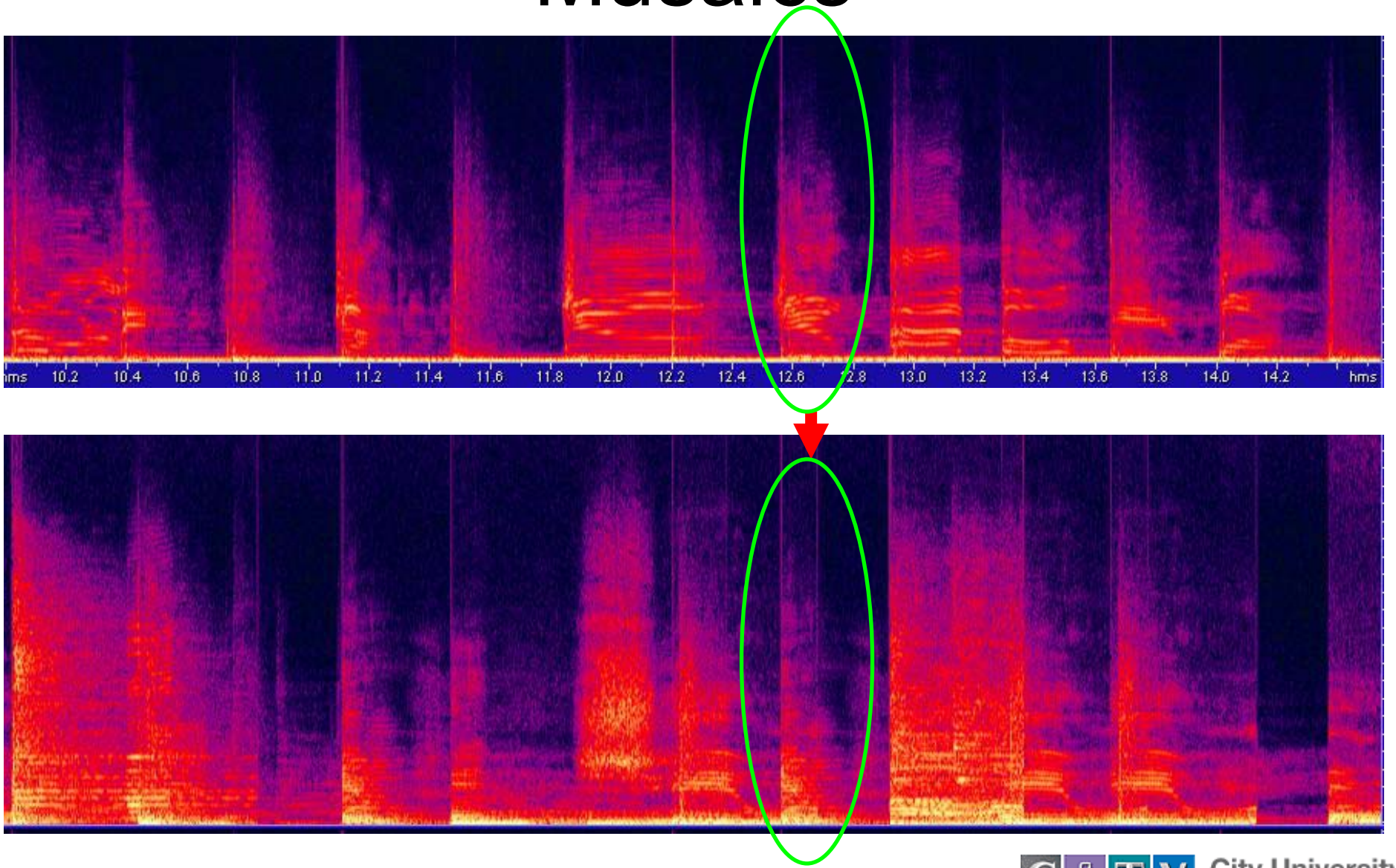
Musaics


hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

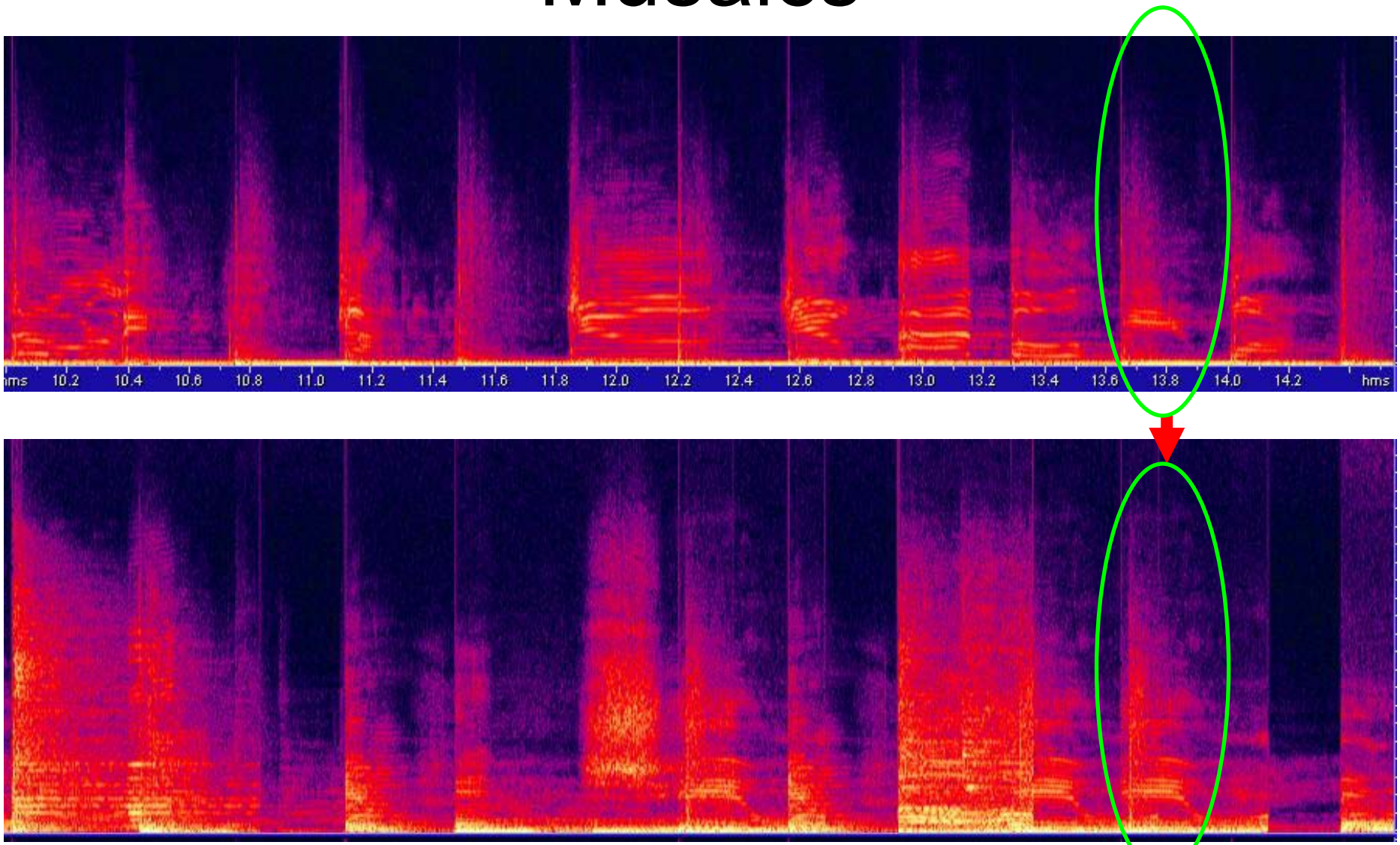
Musaics


hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

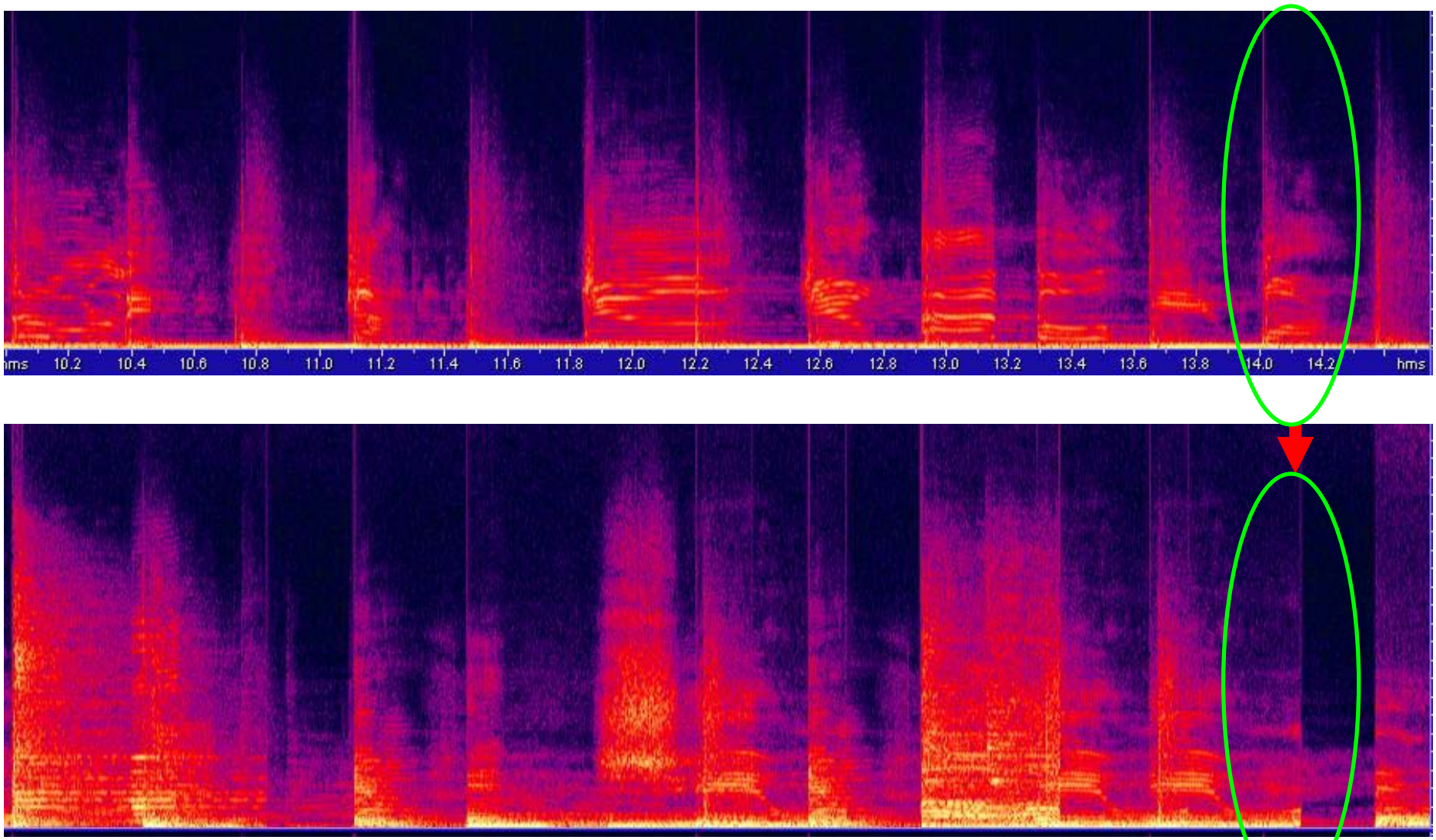
Musaics


hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

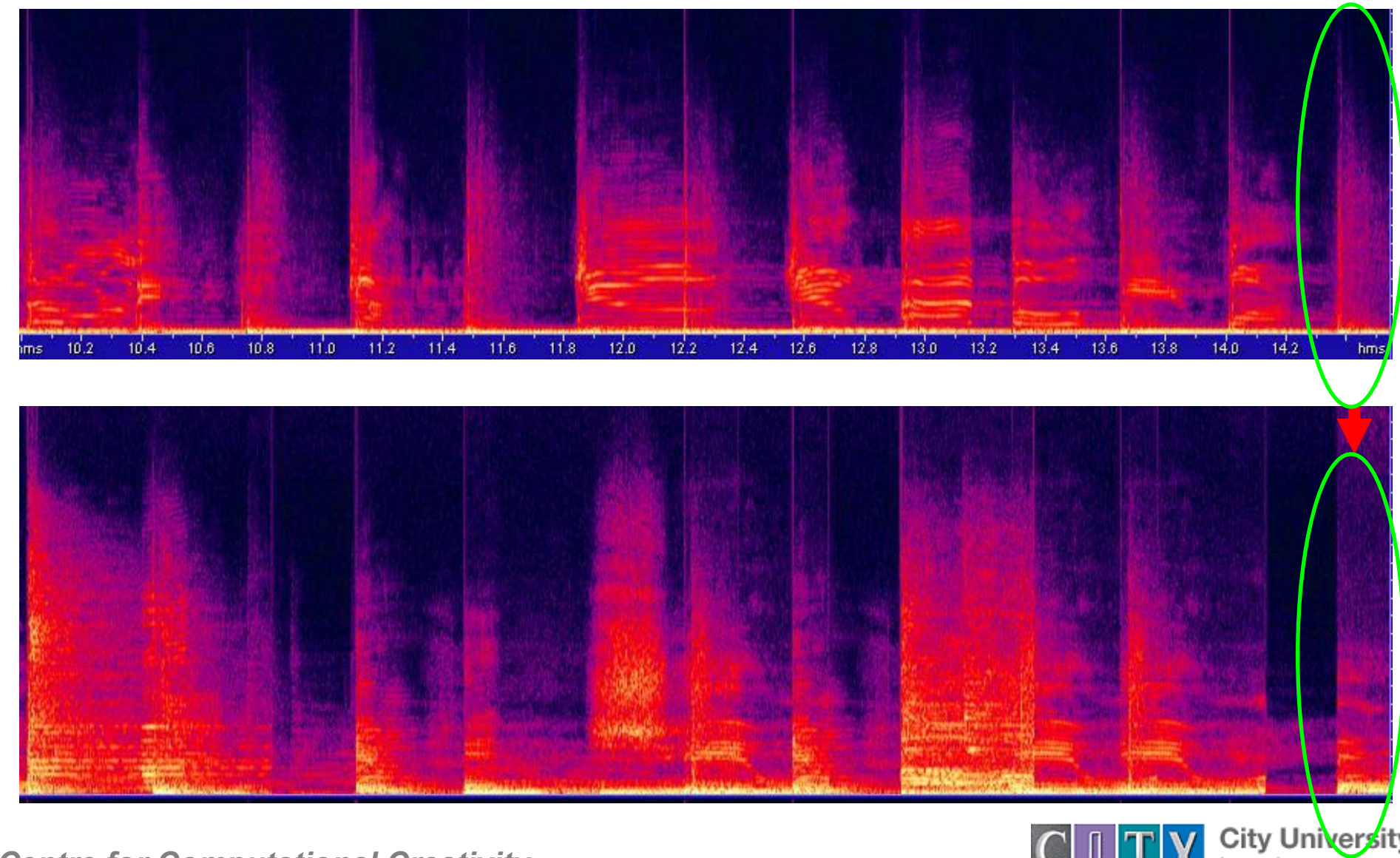
Musaics


hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

Musaics


hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

Musaics


hms 10.2 10.4 10.6 10.8 11.0 11.2 11.4 11.6 11.8 12.0 12.2 12.4 12.6 12.8 13.0 13.2 13.4 13.6 13.8 14.0 14.2 hms

Musaics

- New Content by Similarity Replacement
  - C-Matrix: Cross-Song Similarity Map
  - 1 Target, Many Sources
- Constraints
  - Preserve Rhythm by Beat Tracking
  - Preserve Beats by DTW alignment
- Bigger Source Database == Better
  - Greater Number of Accurate Matches

Acknowledgements

- International Standards Organisation
  - ISO/IEC JTC 1 SC29 WG11 (MPEG)
- Mitsubishi Electric Research Labs
- Massachusetts Institute of Technology
  - Music Mind Machine Group (formerly Machine Listening Group)
  - Paris Smaragdis, Youngmoo Kim, Brian Whitman
  - Iroro Orife, John Hershey, Alex Westner, Kevin Wilson
- City University
  - Department of Computing
  - Centre for Computational Creativity