DAY 3

Intelligent Audio Systems:
A review of the foundations and applications of semantic audio analysis and music information retrieval

Jay LeBoeuf
Imagine Research
jay{at}imagine-research.com

Kyogu Lee
Gracenote
Kglee{at}ccrma.stanford.edu

June 2009

These lecture notes contain hyperlinks to the CCRMA Wiki.

On these pages, you can find supplemental material for lectures - providing extra tutorials, support, references for further reading, or demonstration code snippets for those interested in a given topic .

Click on the symbol on the lower-left corner of a slide to access additional resources.

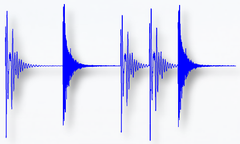
WIKI REFERENCES…

Review from Day 2

- BBQ Today
- Correction on knn formatting
- Name some spectral features
- What are the 3 major components of a MIR system?
- Why do we have to scale our extracted features?
- Which of these did we really not do at all in Lab 2? And, do you think this was a problem?

- How did the lab go?
- Let's dig into some interesting observations from the lab
- Did you try other audio files – other instrument recognizers?

Basic system overview



Segmentation

(Frames, Onsets, Beats, Bars, Chord Changes, etc)

→

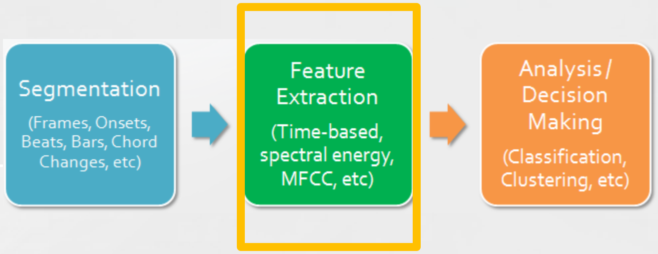
Feature Extraction

(Time-based, spectral energy, MFCC, etc)

→

Analysis / Decision Making

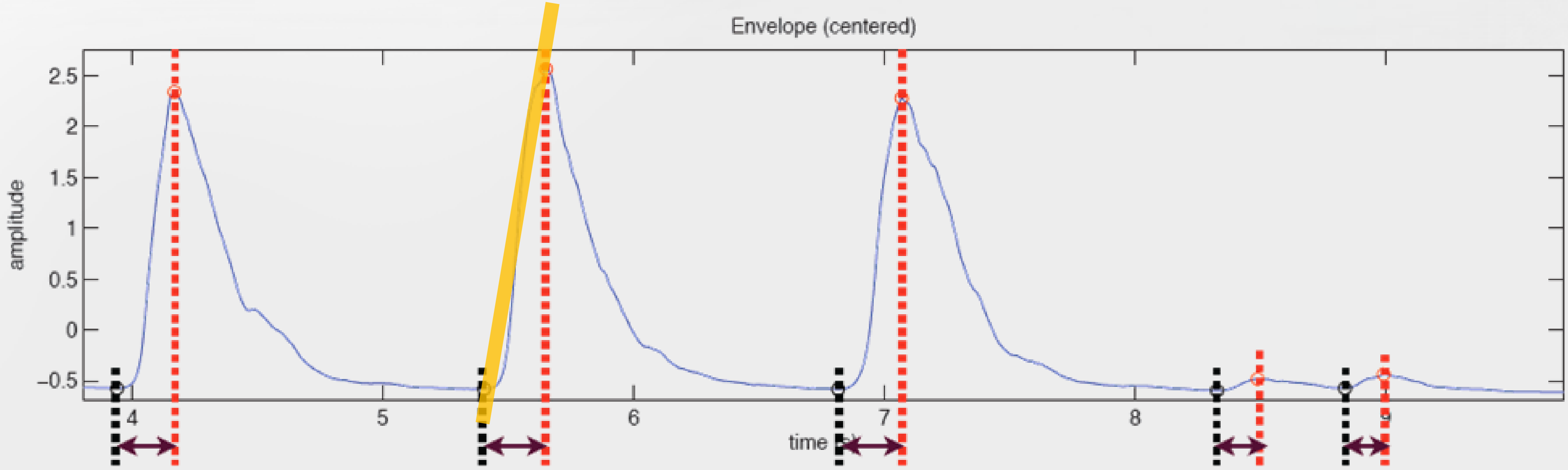
(Classification, Clustering, etc)



FEATURE EXTRACTION

Temporal Information

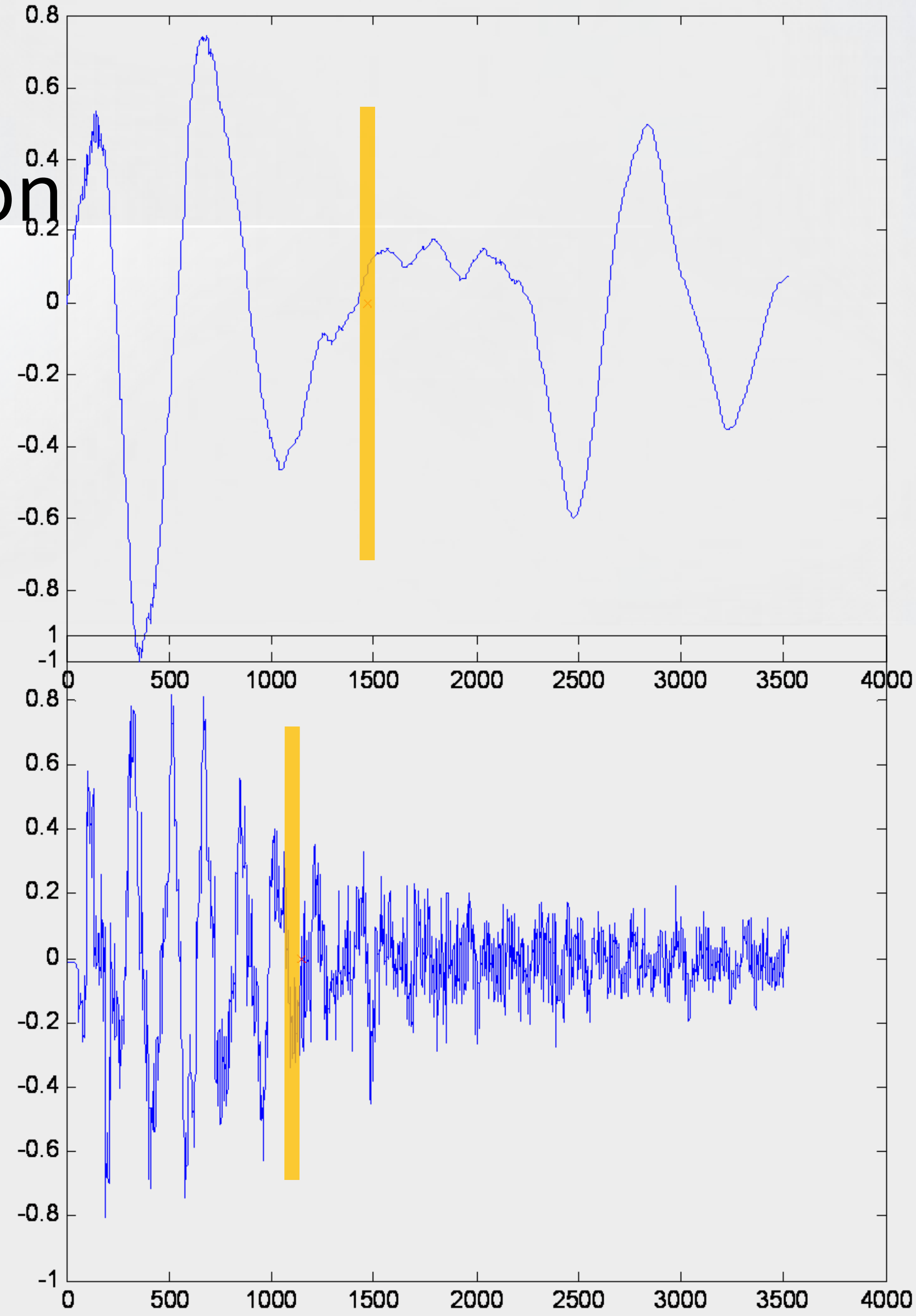
- Rise time or Attack time- time interval between the onset and instant of maximal amplitude
- Attack slope



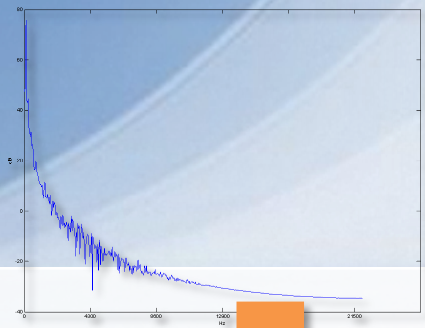
Picture courtesy: Olivier Lartillot

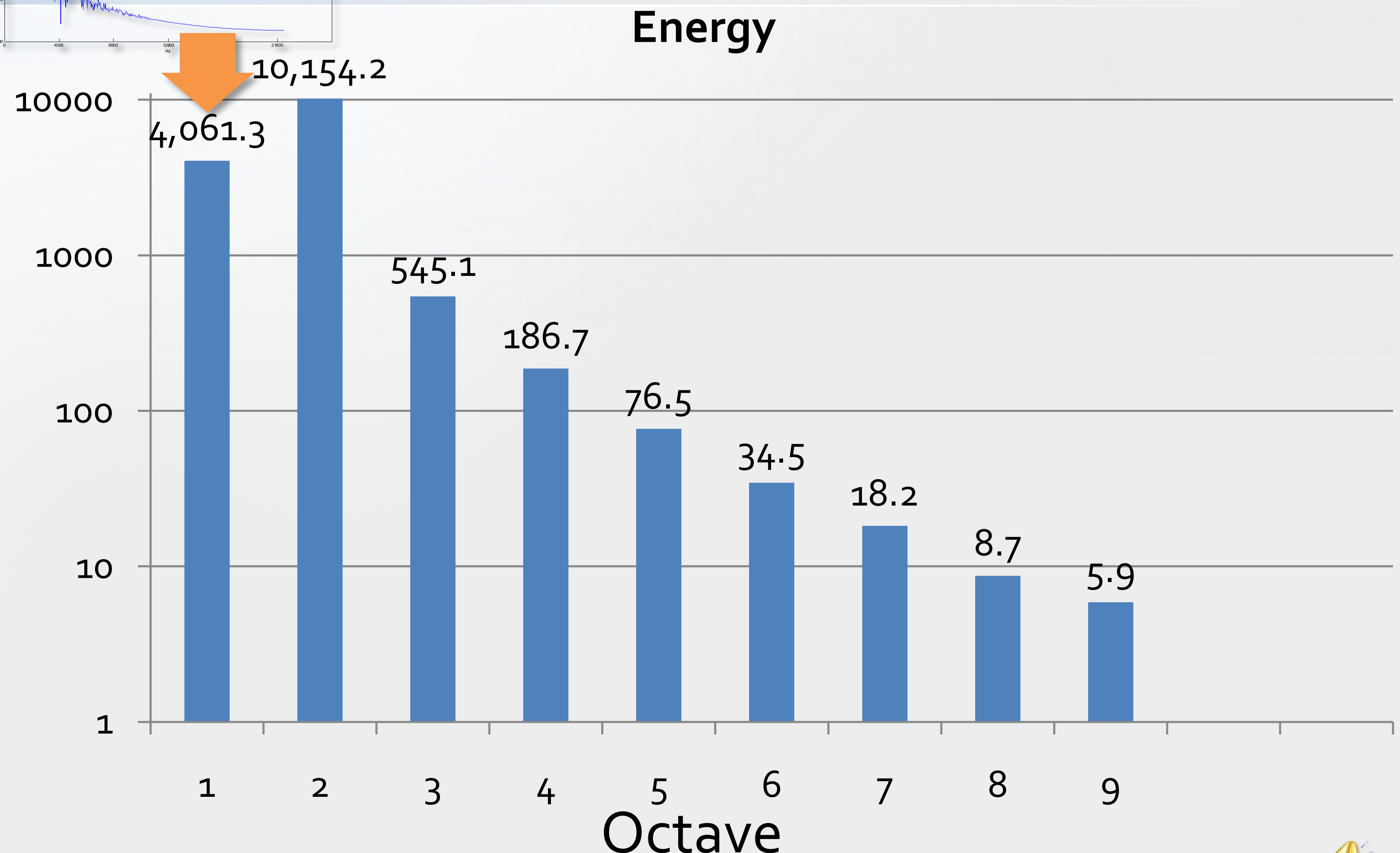
Temporal Information

- Temporal Centroid



Frame 1

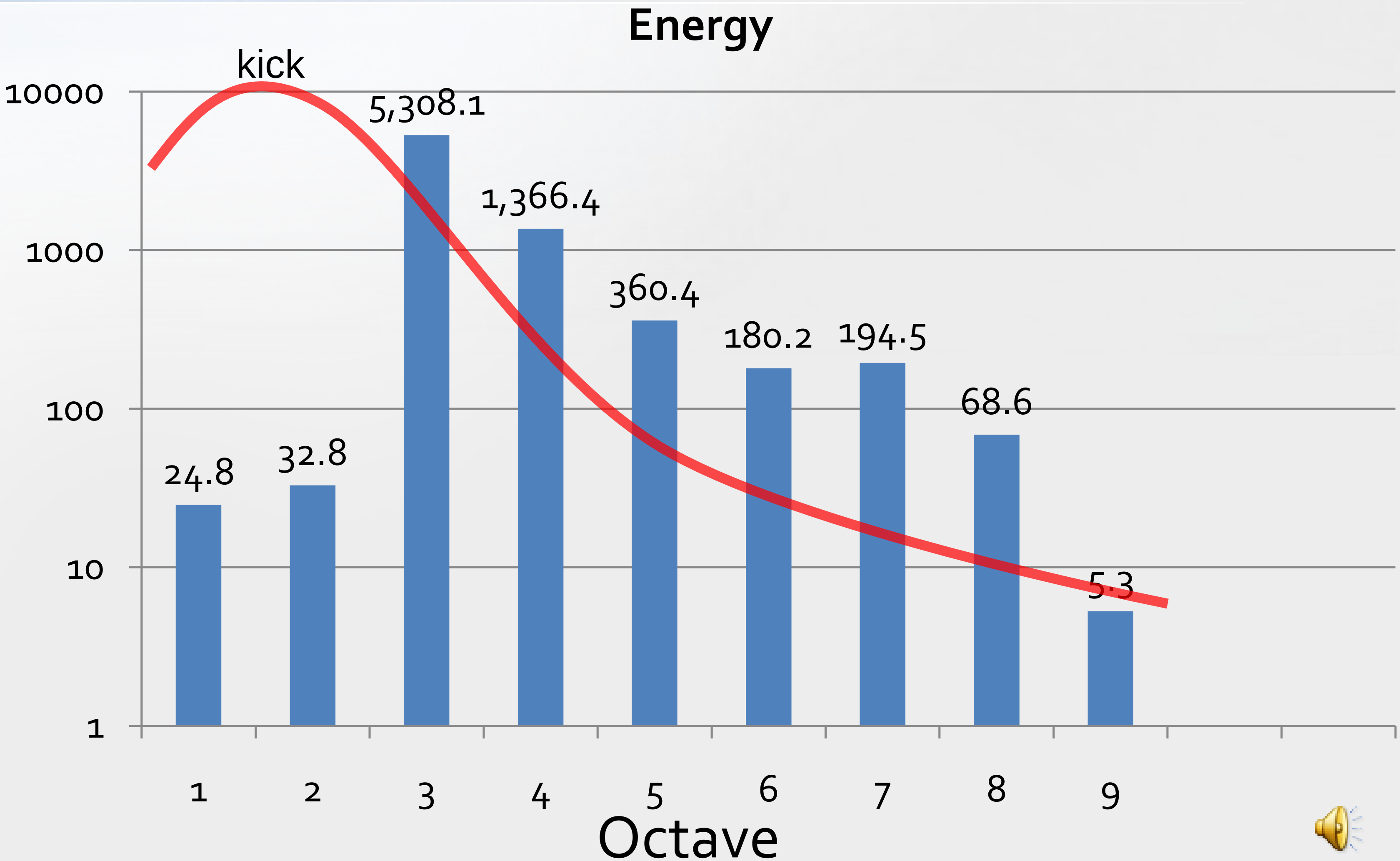



Energy
10,154.2
4,061.3
545.1
186.7
76.5
34.5
18.2
8.7
5.9
10000
1000
100
10
1
1
2
3
4
5
6
7
8
9
Octave

Features – Frame 1

Frame	ZCR	Centroid	BW	Skew	Kurtosis	E1	E2	E3	E4	E5	E6	E7	E8	E9
1	9	2.8kHz	5kHz	2.2	6.7	4000	10100	545	187	77	35	18	9	6

Frame 2


Energy
kick
10000
1000
100
10
1
5,308.1
1,366.4
360.4
180.2
194.5
68.6
24.8
32.8
5.3
1
2
3
4
5
6
7
8
9
Octave

Features : SimpleLoop.wav

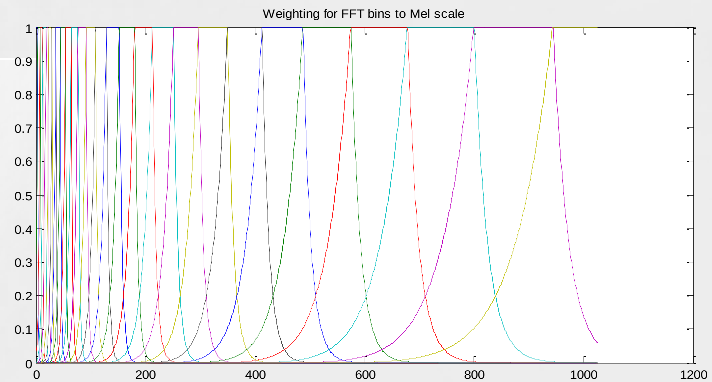
Frame	ZC R	Centroid	BW	Skew	Kurtosis	E1	E2	E3	E4	E5	E6	E7	E8
1	9	2.8kHz	5kHz	2.2	6.7	4000	10100	545	187	77	35	18	9
2	423	3.1kHz	4kHz	2	7.2	24	33	5300	1366	360	180	194	68

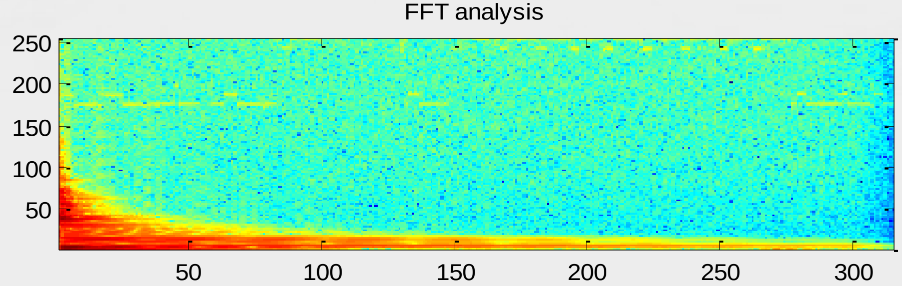
MFCCs

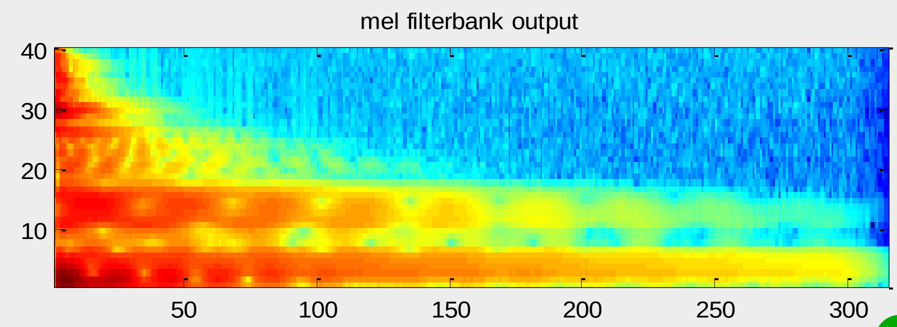
The idea of MFCCs is to capture spectrum in accordance with human perception.

1. STFT
2. log(STFT)
3. Perform mel-scaling to group and smooth coefficients. (perceptual weighting)
4. Decorrelate with DCT

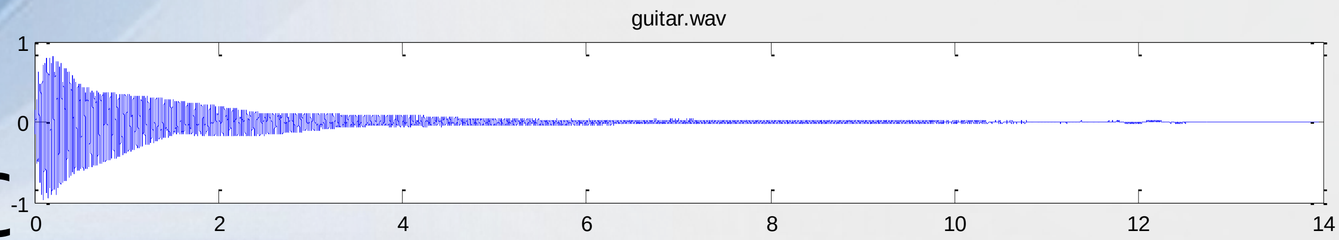
[...continued...]



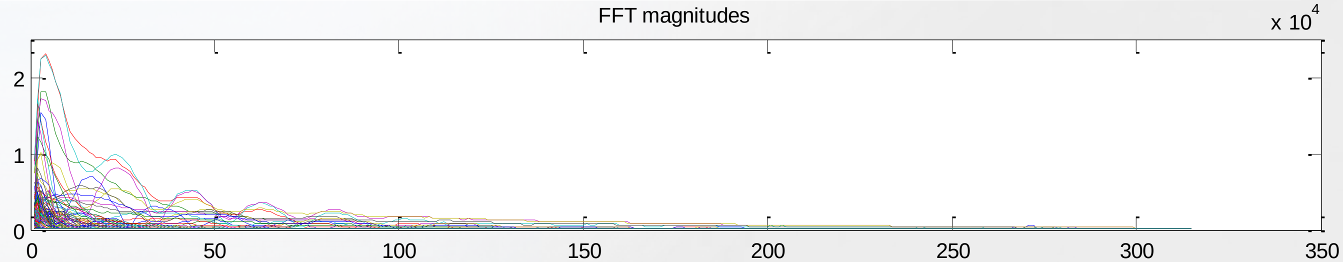




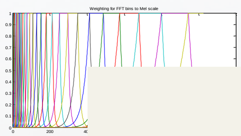
MFCC


guitar.wav
1
0
-1
0
2
4
6
8
10
12
14
x 10^4

1


FFT magnitudes
2
1
0
0
50
100
150
200
250
300
350

2

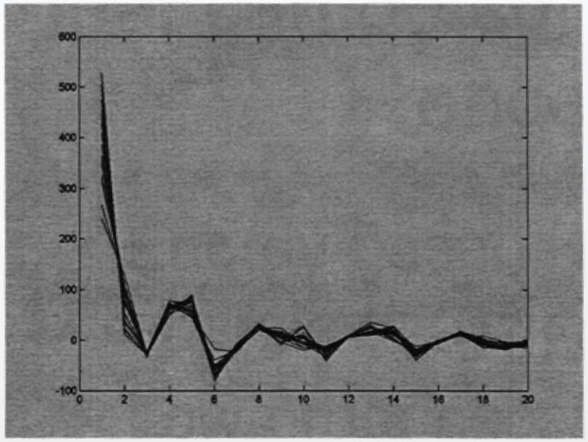


3

4

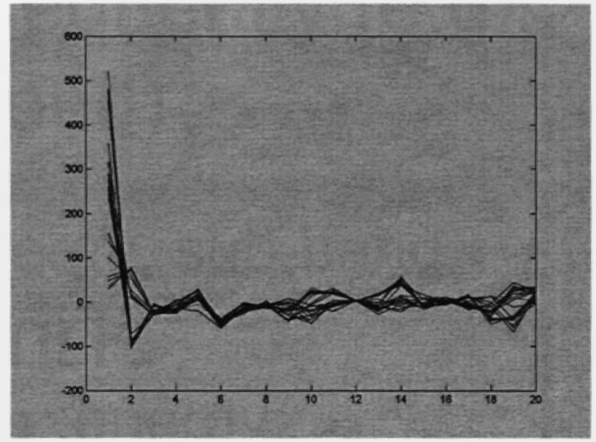
MFCC of Music

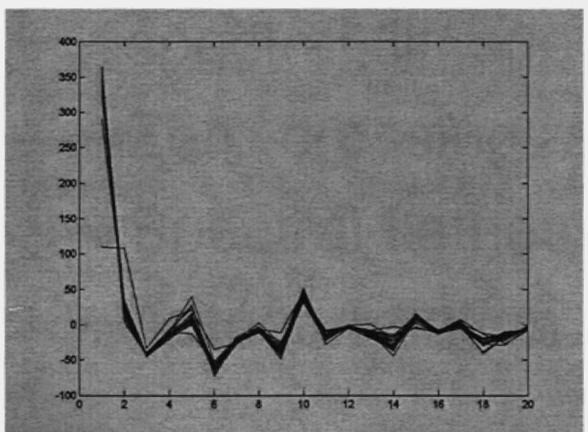
(Petruncio, 2003)



Piano

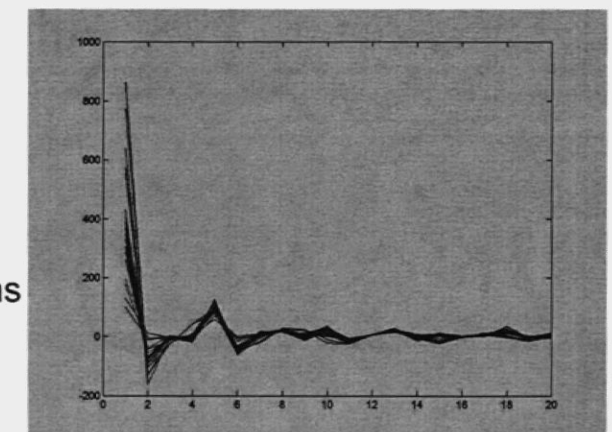
Saxophone





Tenor
Opera
Singer

Drums



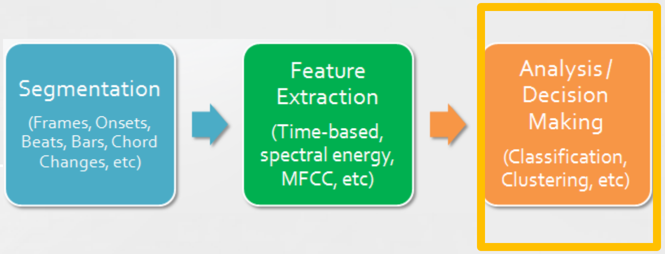
Spectral Energy vs. MFCC

Features: Measuring changes

- Δ and Δ Δ
  - Change between frames
  - How quickly the change is occurring
- Spectral flux is the distance between the spectrum of successive frames

Feature extraction

- Feature design and creation uses one’s domain knowledge.
- Choosing discriminating features is critical
- Smaller feature space yields smaller, simpler models, faster training, often less training data needed



ANALYSIS AND DECISION MAKING

Supervised vs. Unsupervised

- Unsupervised - “clustering”
- Supervised – binary classifiers (2 classes)
- Multiclass is derived from binary

Clustering

- Unsupervised learning – find pockets of data to group together
- Statistical analysis techniques

Clustering

- K = # of clusters

- Choosing the number of clusters – note that choosing the “best” number of clusters according to minimizing total squared distance will always result in same # of clusters as data points.

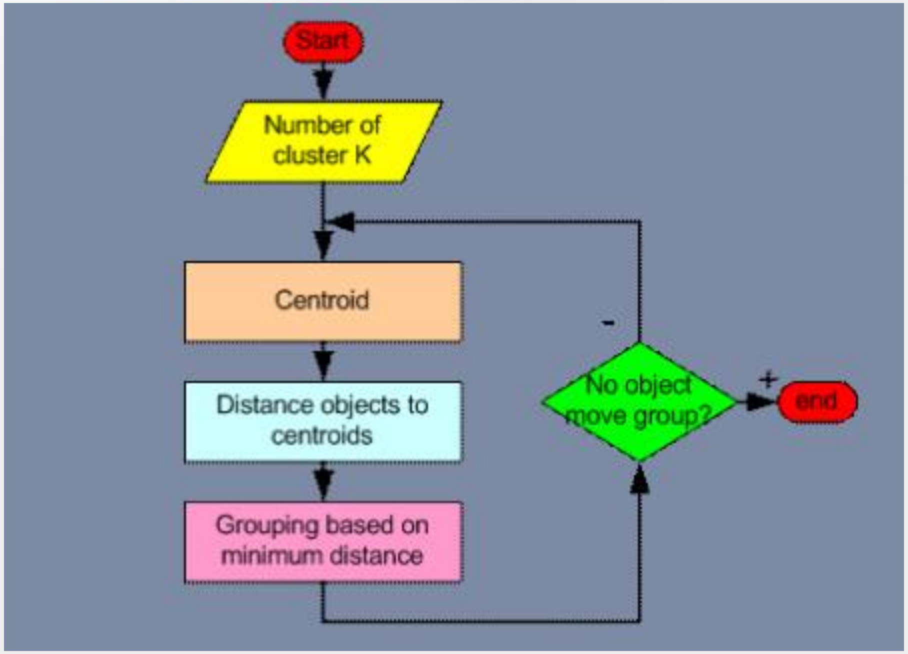
Clustering

The basic goal of clustering is to divide the data into groups such that the points within a group are close to each other, but far from items in other groups.

Hard clustering – each point is assigned to one and only one cluster.

Demo

- http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html



K-Means

The key points relating to *k-means clustering are:*

- k-means is an automatic procedure for clustering unlabelled data;
- it requires a pre-specified number of clusters;
- Clustering algorithm chooses a set of clusters with the minimum within-cluster variance
- Guaranteed to converge (eventually)
- Clustering solution is dependent on the initialization

(You get different results with each running)

K-Means

The initialization method needs to be further specified. There are several possible ways to initialize the cluster centers:

- *Choose random data points as cluster centers*
- *Randomly assign data points to K clusters and compute means as initial centers*
- *Choose data points with extreme values*
- *Find the mean for the whole data set then perturb into k means*
- *Find ground-truth for data*

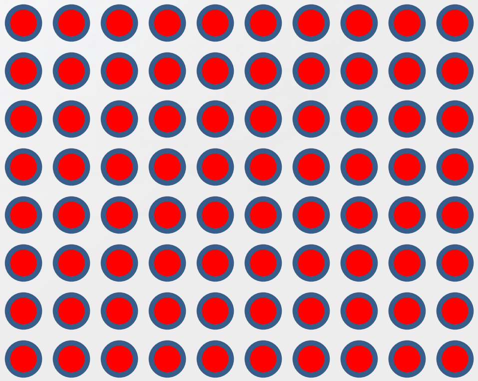
EVALUATION

Our classifier accuracy is 83.4%

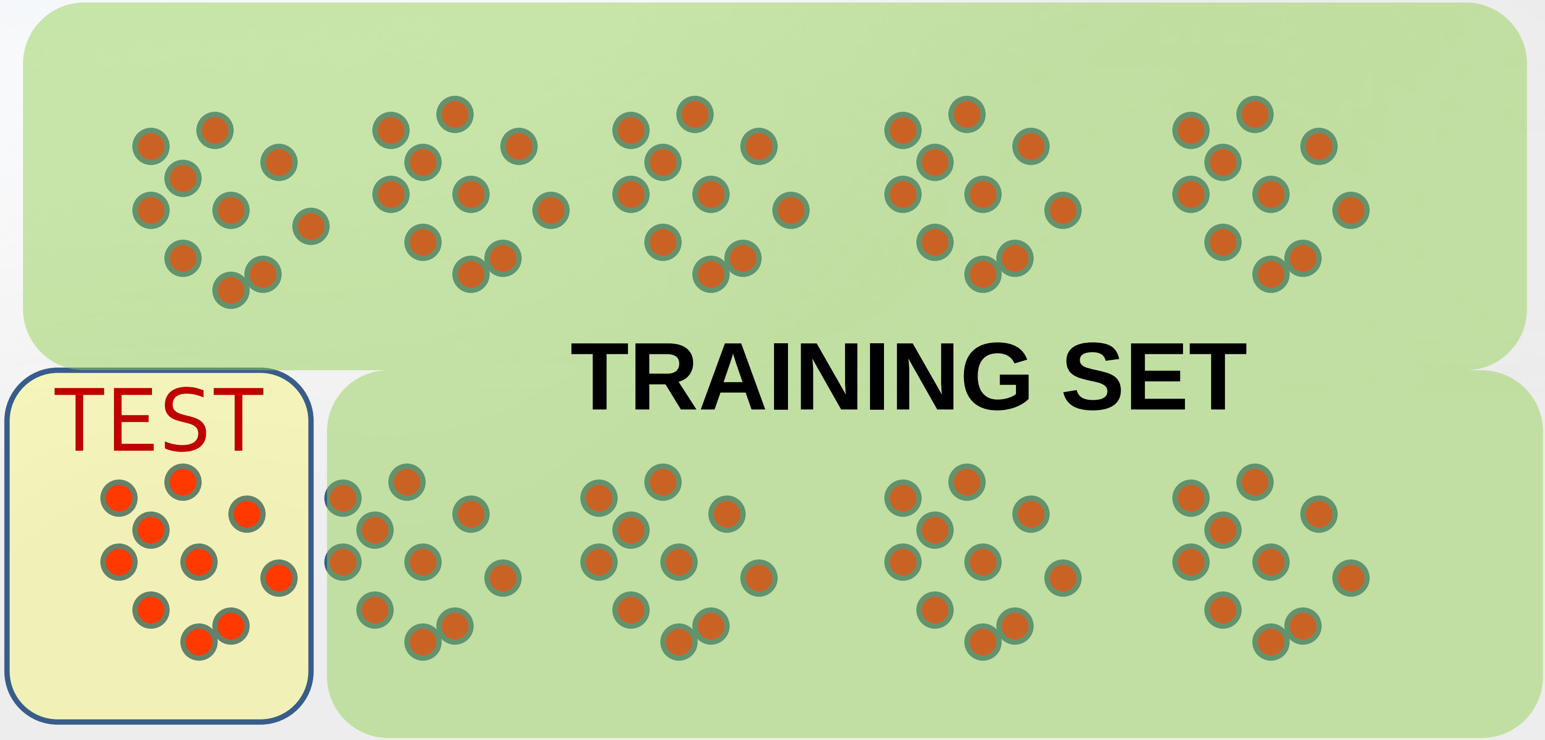
Cross-validation

- Say, 10-fold cross validation
- Divide test set into 10 random subsets.
- 1 test set is tested using the classifier trained on the remaining 9.
- We then do test/train on all of the other sets and average the percentages. Helps prevent over fitting.
- Do not optimize too much on cross validation – you can severely overfit. Sanity check with a test set.

Cross-validation



Cross-validation



Fold 1: 70%

Cross-validation

Fold 1: 70%
Fold 2: 80%

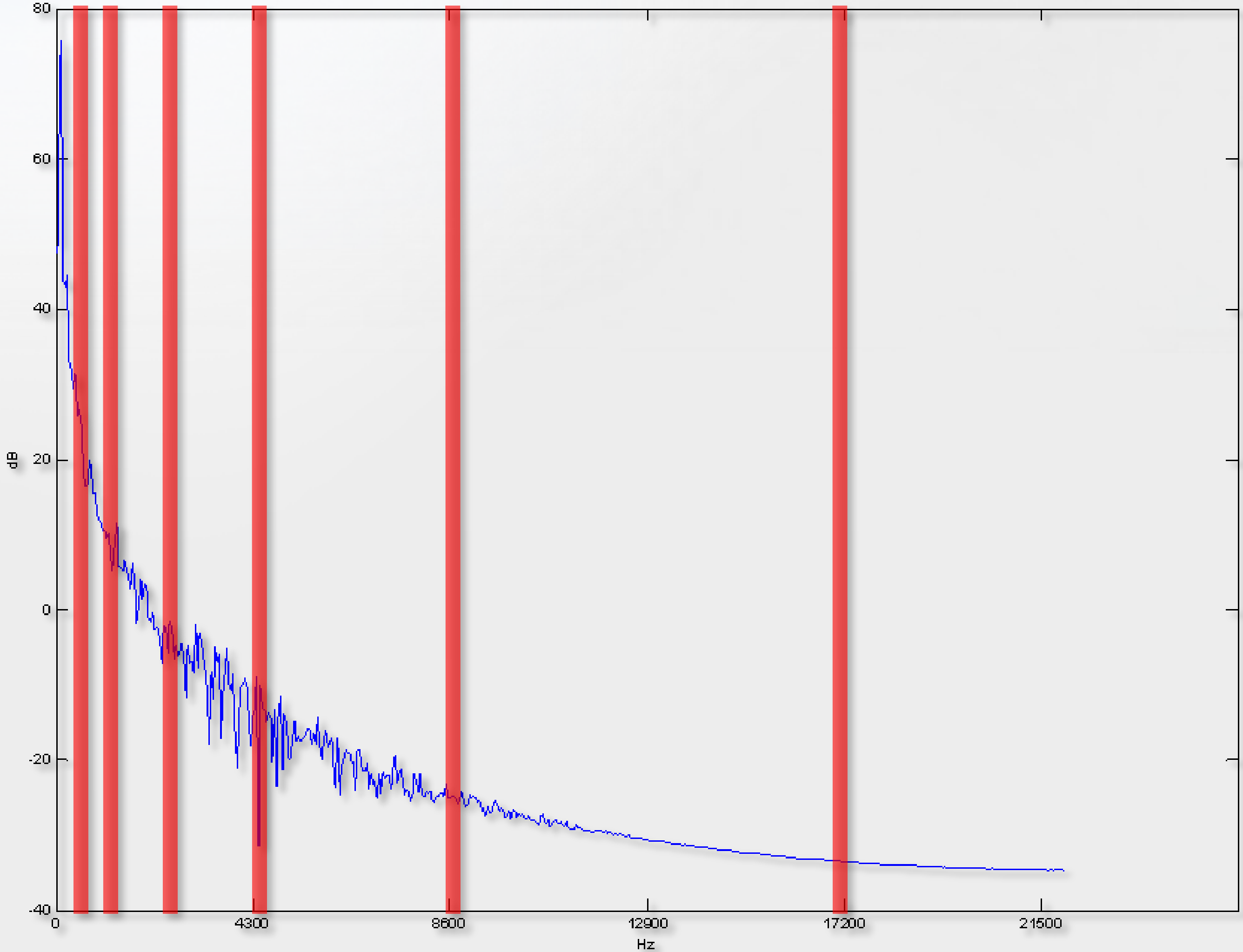
Cross-validation

Fold 1: 76%
Fold 2: 80%
Fold 3: 77%
Fold 4: 83%
Fold 5: 72%
Fold 6: 82%
Fold 7: 81%
Fold 8: 71%
Fold 9: 90%
Fold 10: 82%
Mean = 79.4%

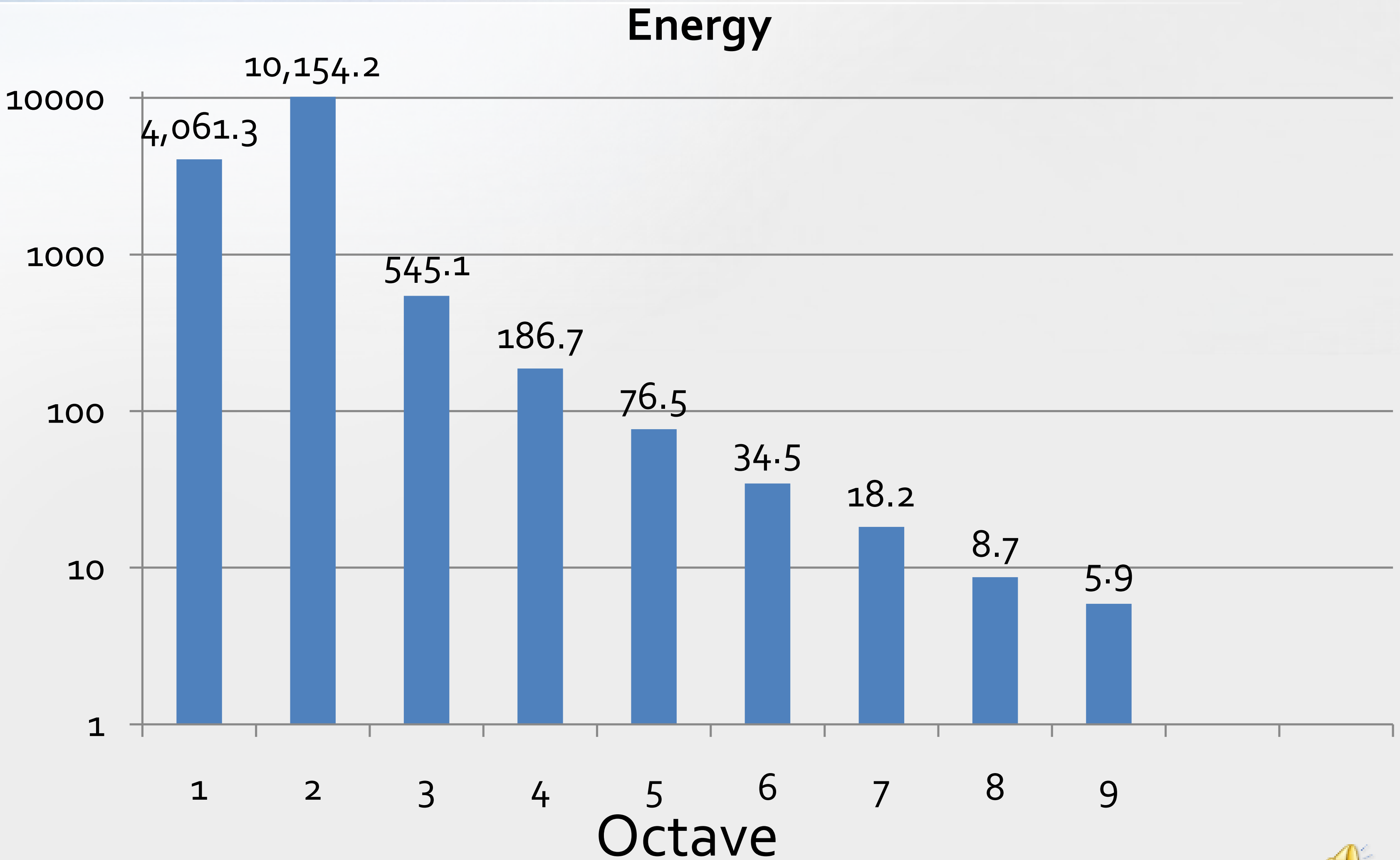
Stratified Cross-Validation

- Same as cross-validation, except that the folds are chosen so that they contain equal proportions of labels.

Spectral Bands



Frame 1

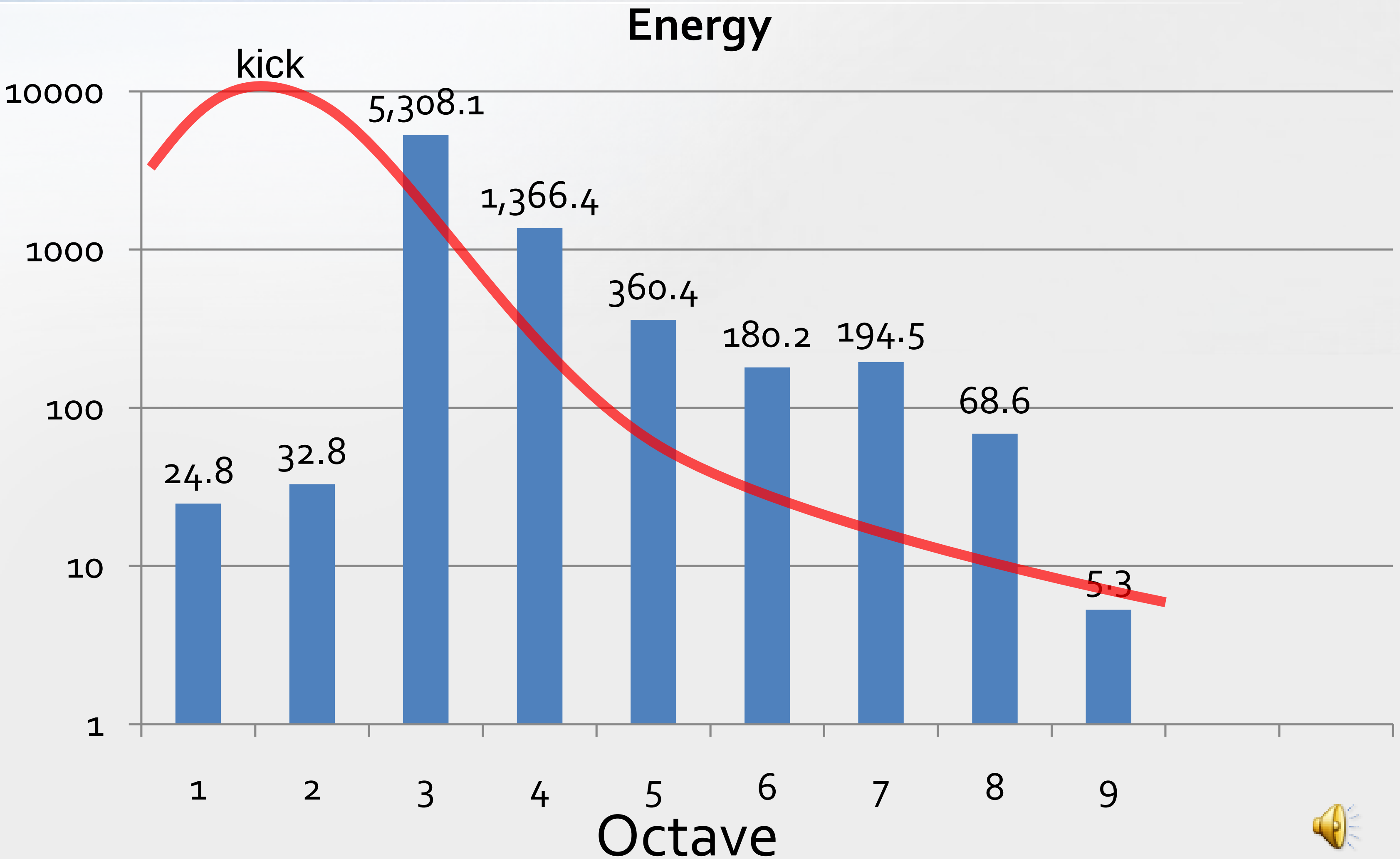


Energy
10,154.2
4,061.3
545.1
186.7
76.5
34.5
18.2
8.7
5.9
10000
1000
100
10
1
1
2
3
4
5
6
7
8
9
Octave

Features – Frame 1

Frame	ZCR	Centroid	BW	Skew	Kurtosis	E1	E2	E3	E4	E5	E6	E7	E8	E9
1	9	2.8kHz	5kHz	2.2	6.7	4000	10100	545	187	77	35	18	9	6

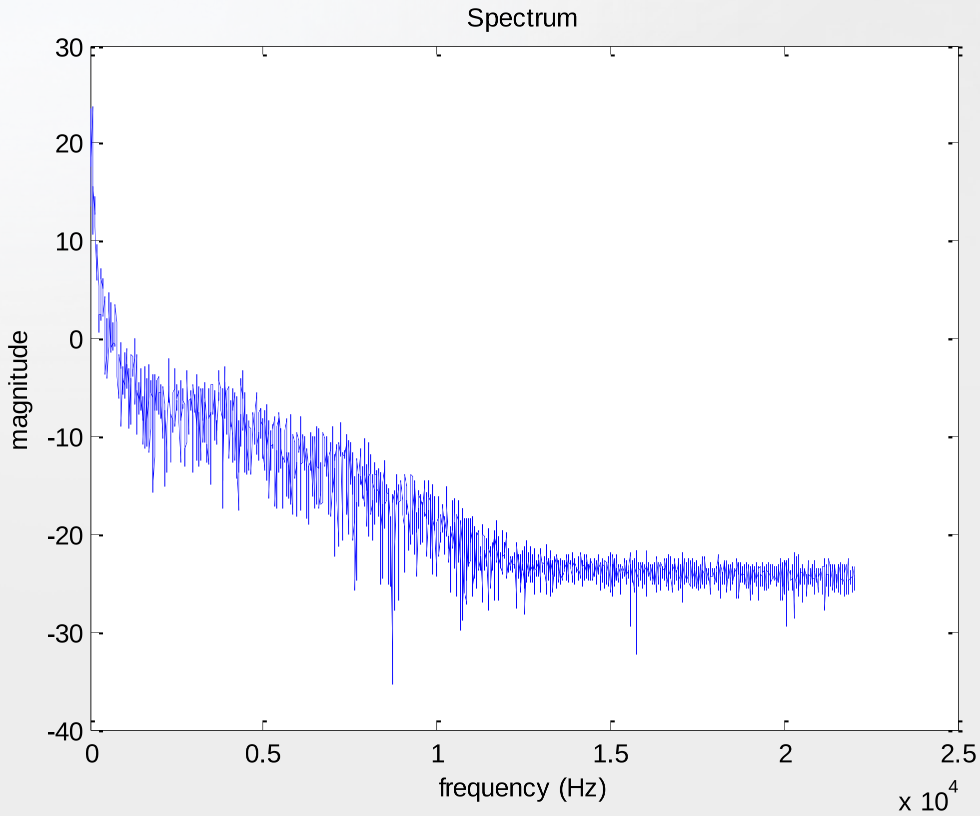
Frame 2

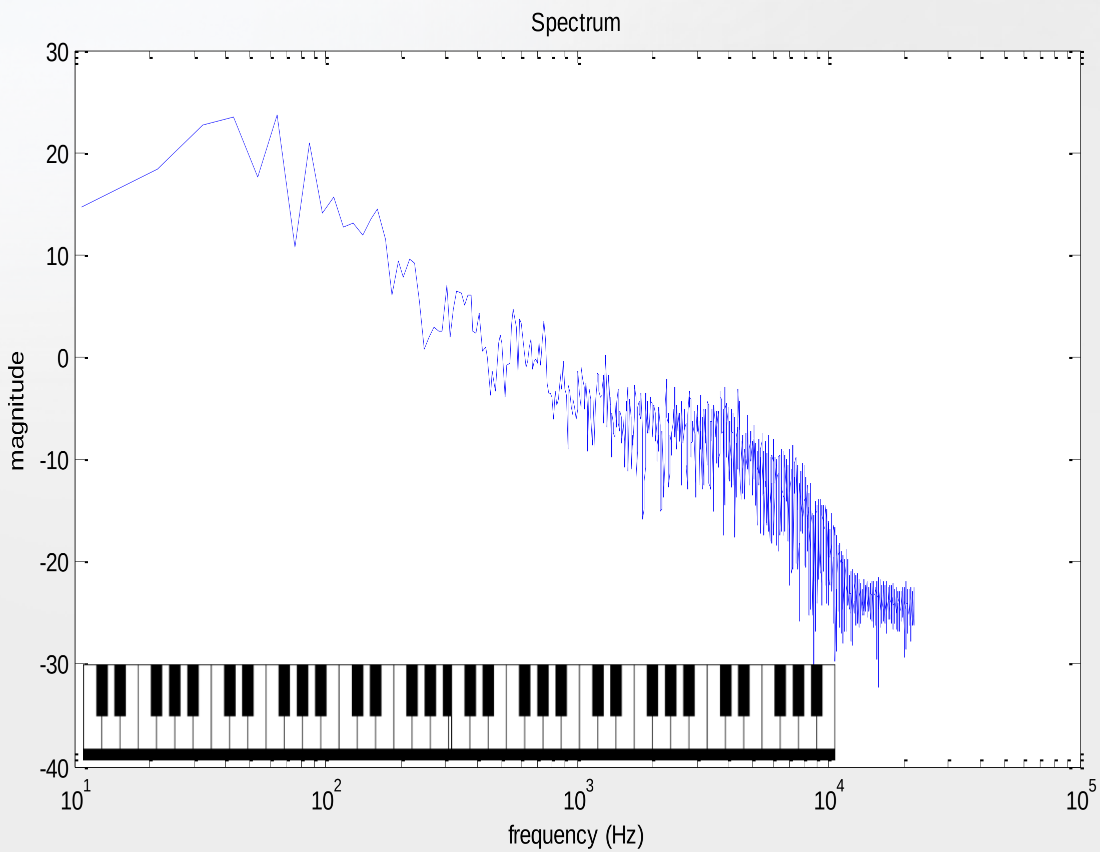

Energy
kick
10000
1000
100
10
1
24.8
32.8
5,308.1
1,366.4
360.4
180.2
194.5
68.6
5.3
1
2
3
4
5
6
7
8
9
Octave

Features : SimpleLoop.wav

Frame	ZCR	Centroid	BW	Skew	Kurtosis	E1	E2	E3	E4	E5	E6	E7	E8
1	9	2.8kHz	5kHz	2.2	6.7	4000	10100	545	187	77	35	18	9
2	423	3.1kHz	4kHz	2	7.2	24	33	5300	1366	360	180	194	68

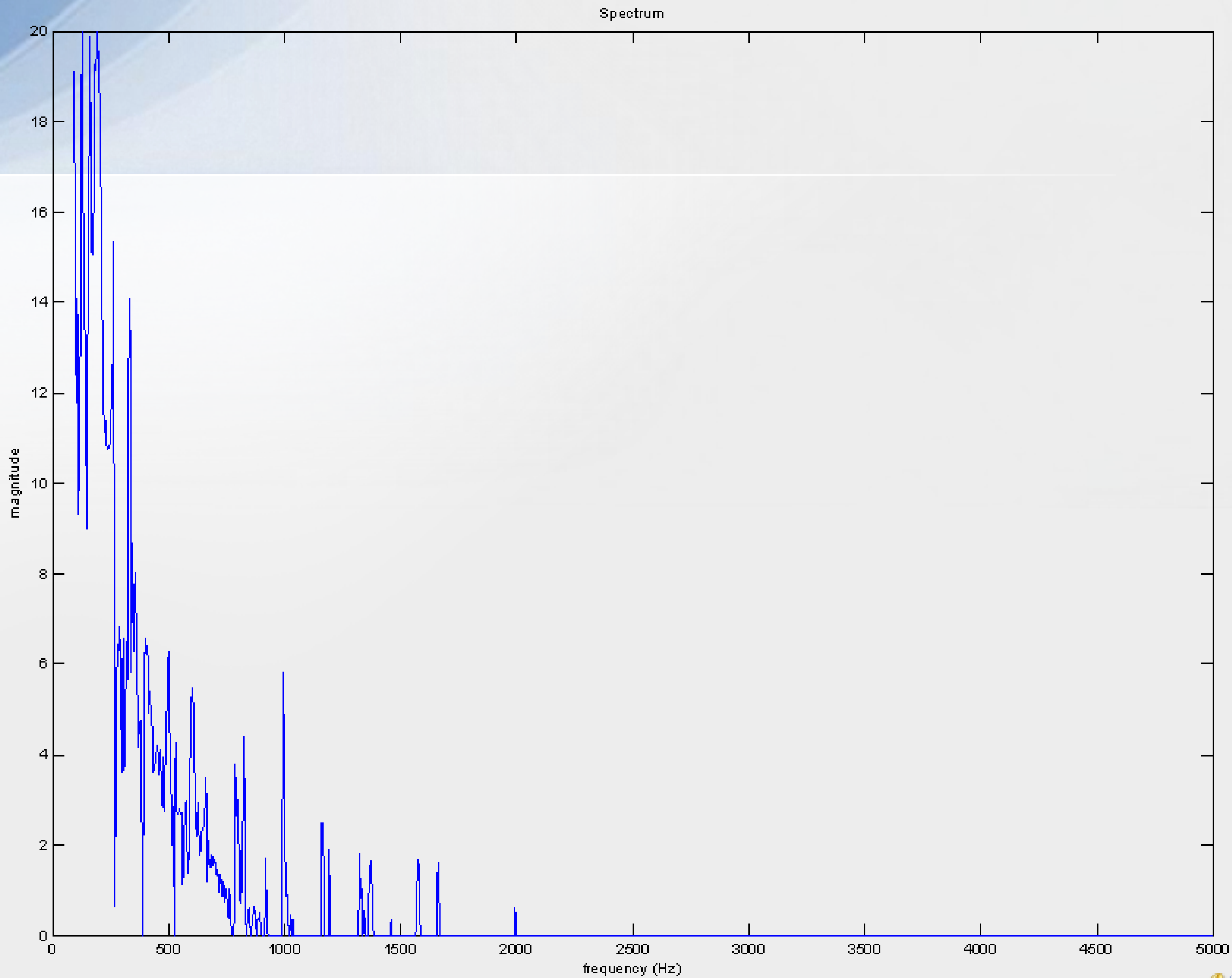
Log Spectrogram



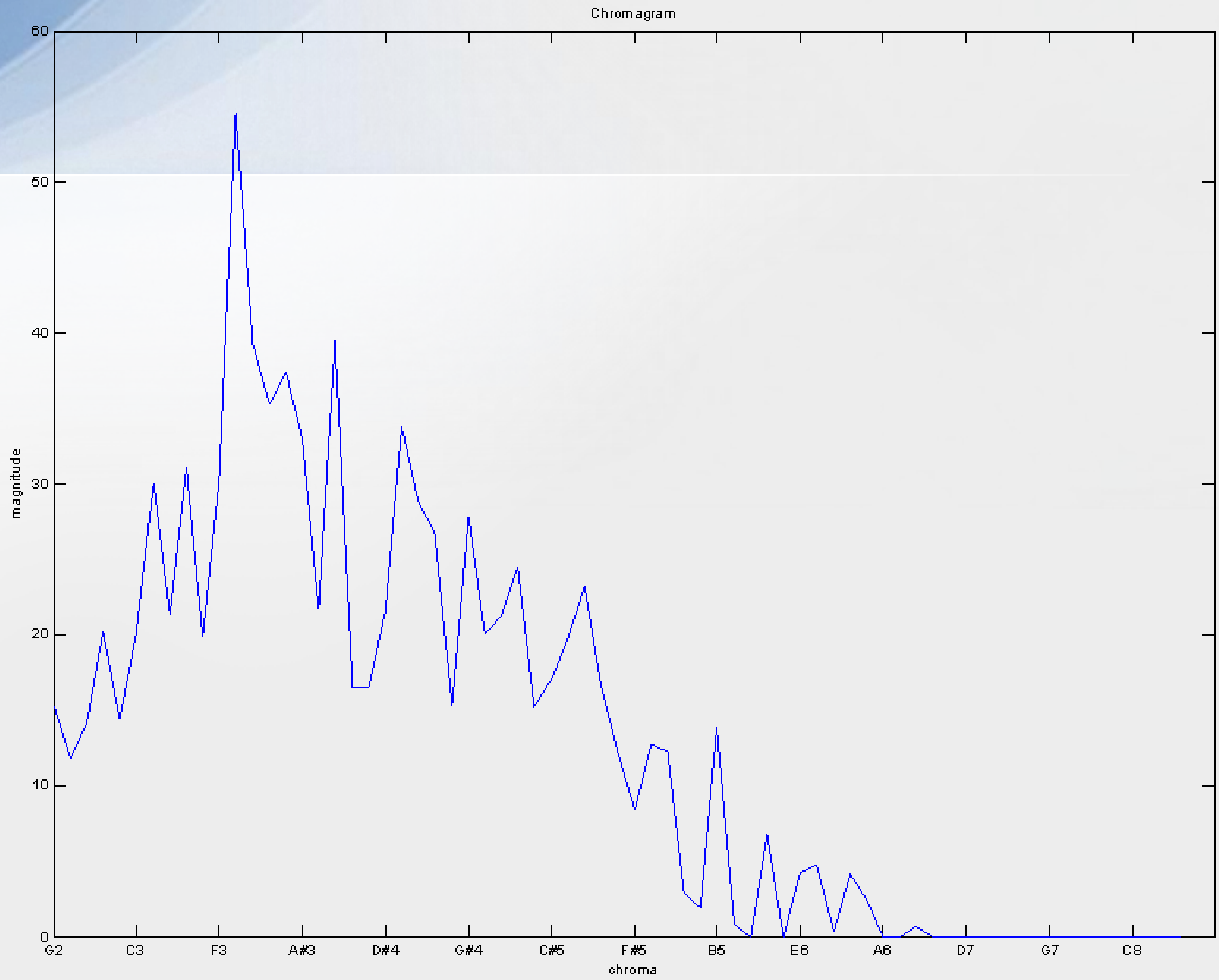


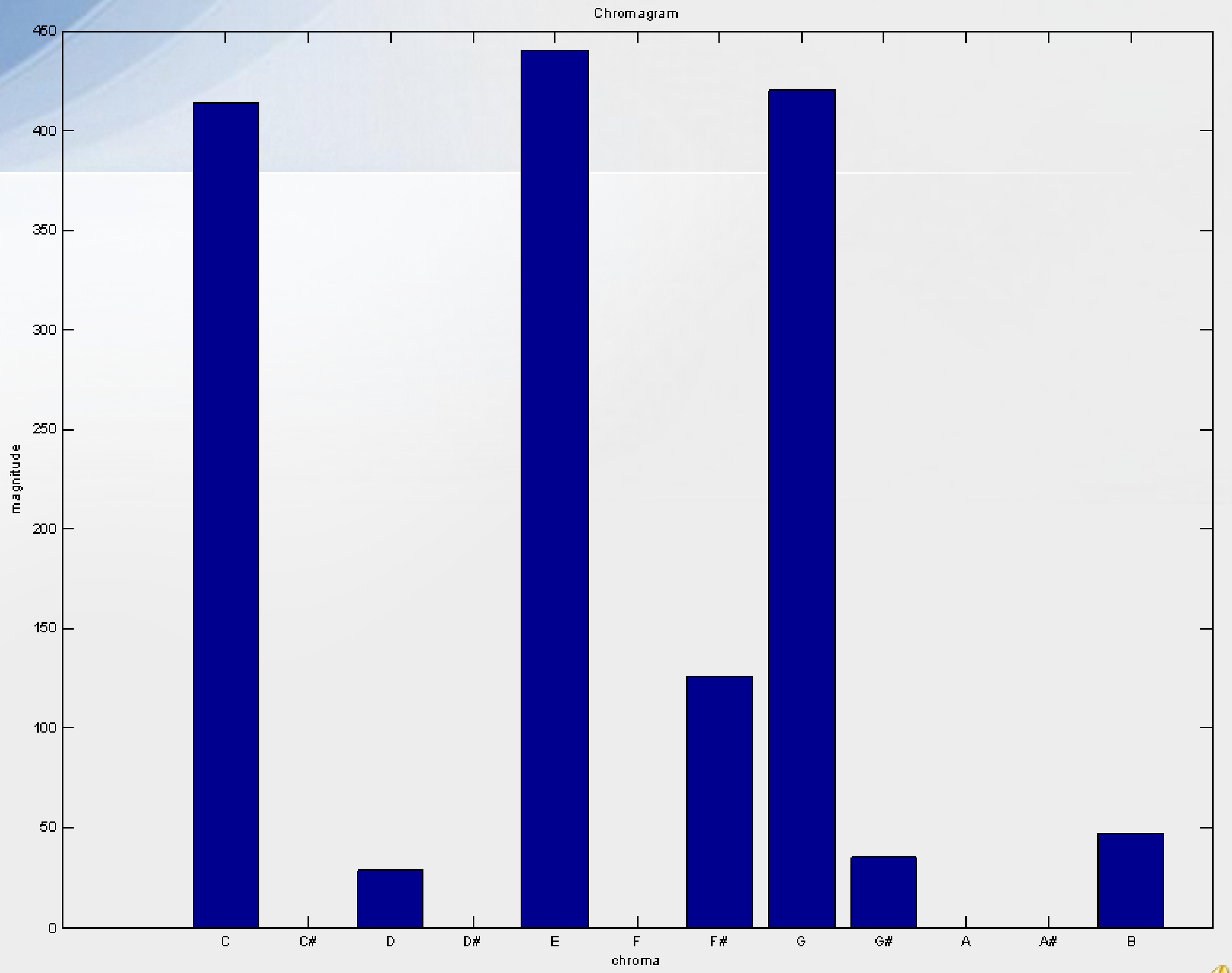
Spectrum
30
20
10
0
-10
-20
-30
-40
magnitude
10^1
10^2
10^3
10^4
10^5
frequency (Hz)

Chroma Bins

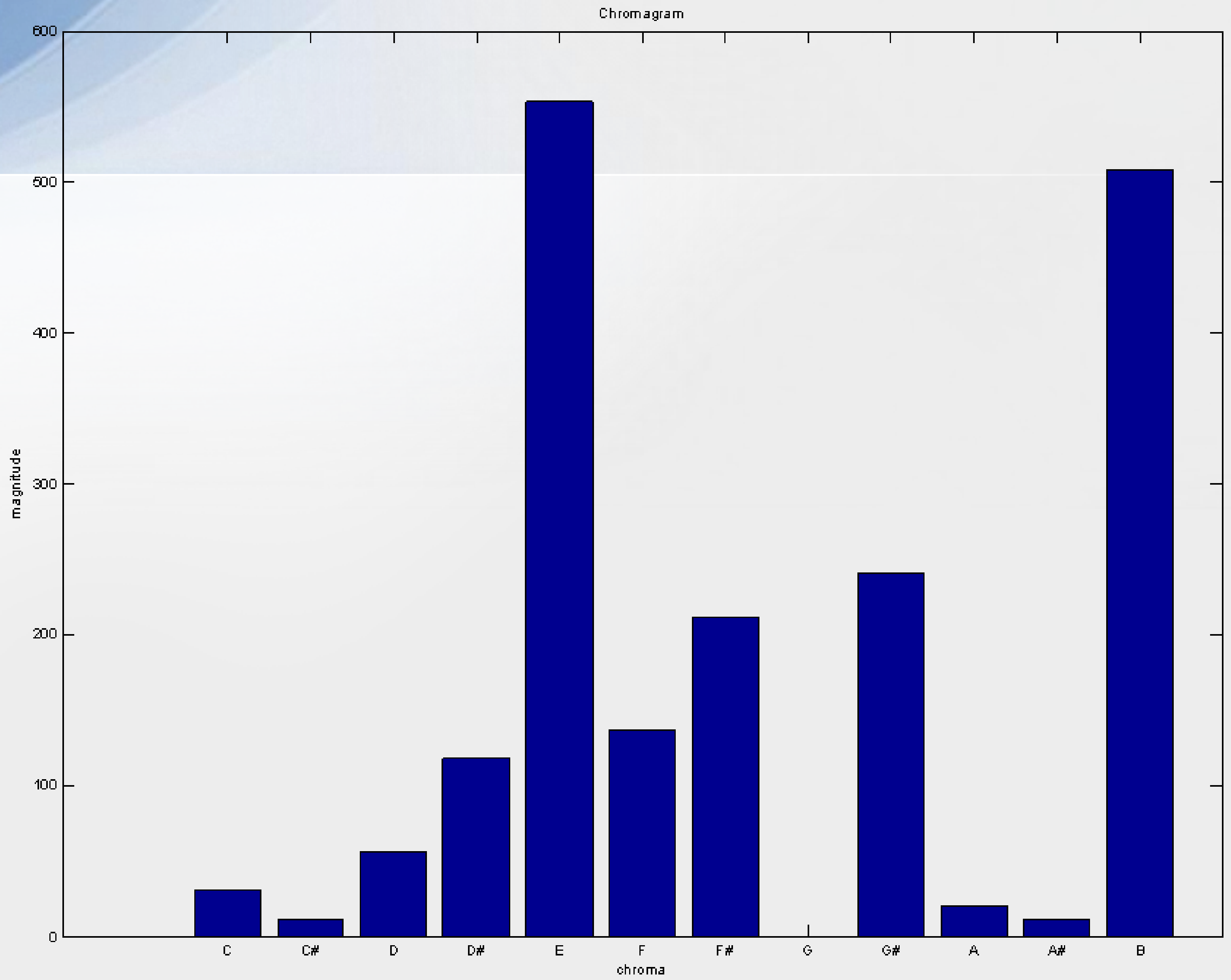


Spectrum
magnitude
frequency (Hz)
0
2
4
6
8
10
12
14
16
18
20
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000


Chromagram
magnitude
chroma
60
50
40
30
20
10
0
G2
C3
F3
A#3
D#4
G#4
C#5
F#5
B5
E6
A6
D7
G7
C8

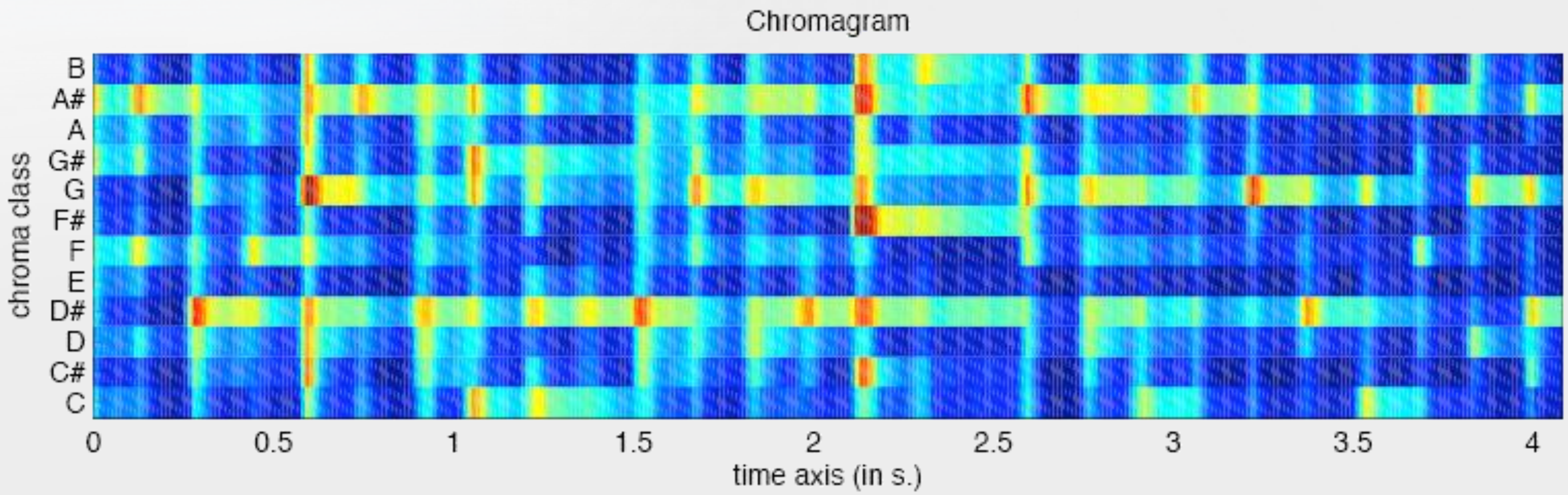


Chromagram
magnitude
0
50
100
150
200
250
300
350
400
450
C
C#
D
D#
E
F
F#
G
G#
A
A#
B
chroma

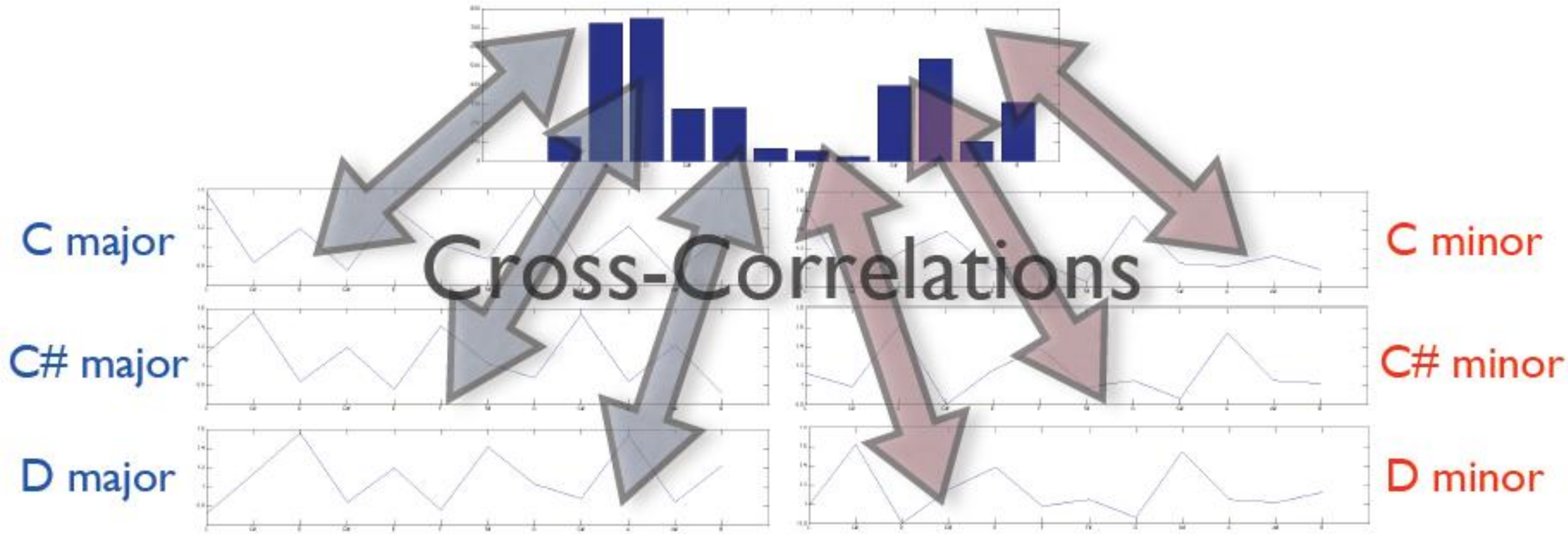


Chromagram
600
500
400
300
200
100
0
magnitude
C
C#
D
D#
E
F
F#
G
G#
A
A#
B
chroma

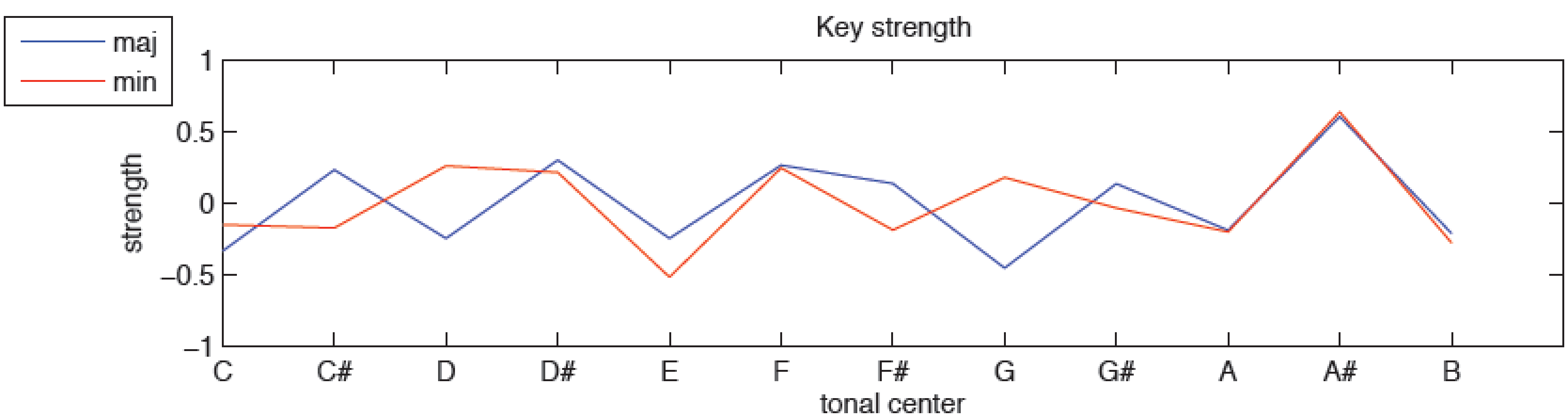
Example



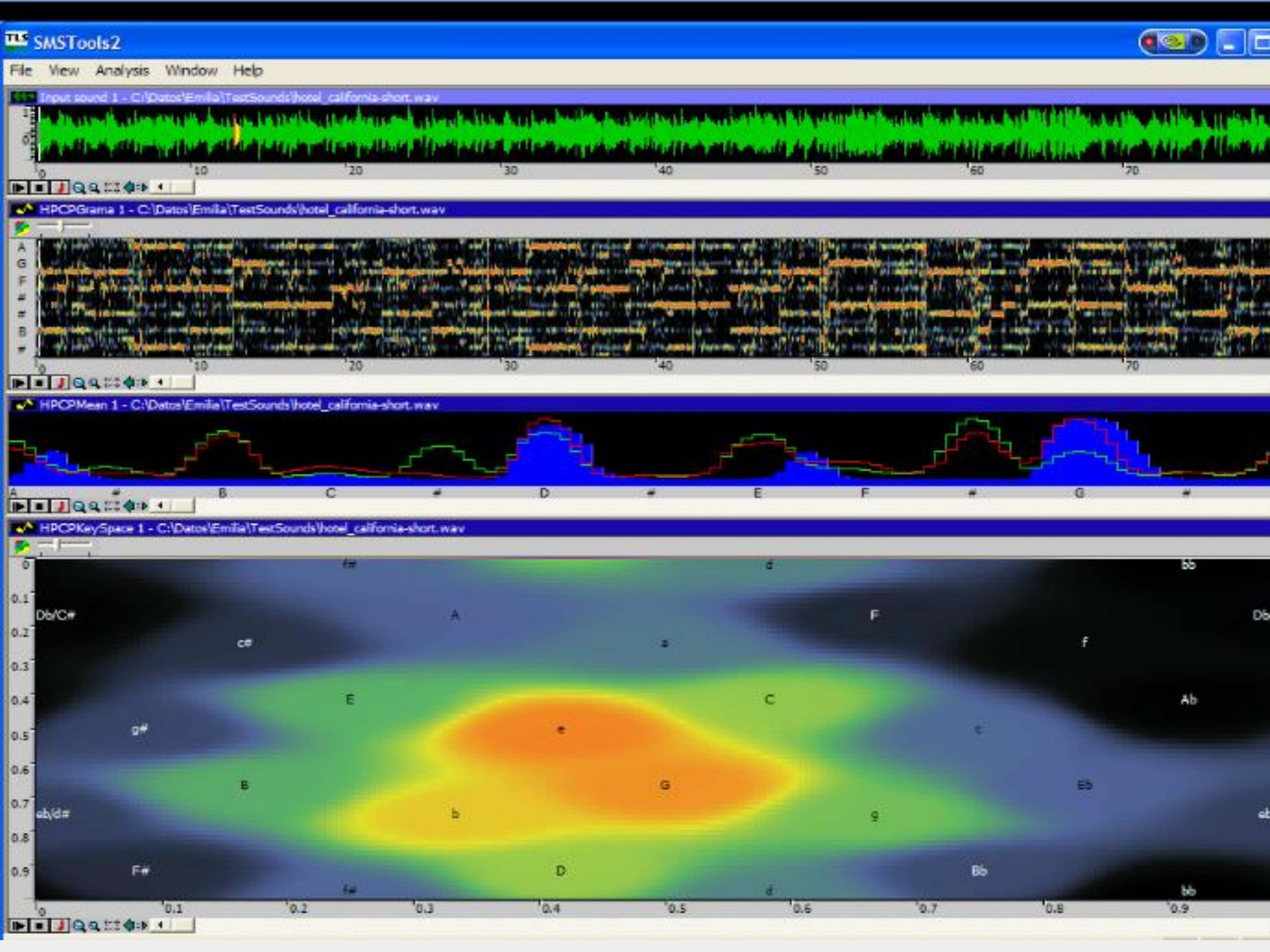
Picture courtesy: Olivier Lartillot



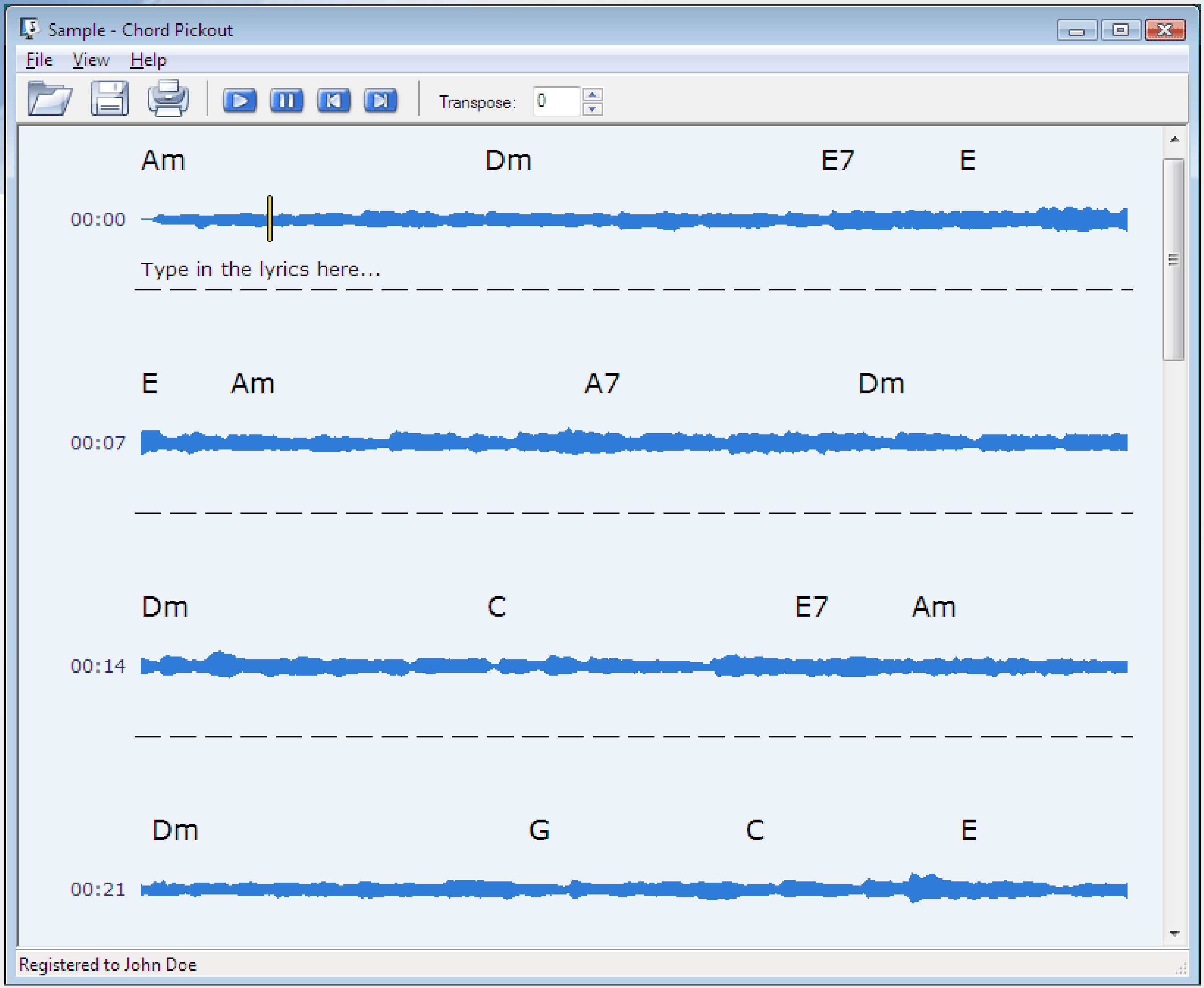
The resulting graph indicate the cross-correlation score for each different tonality candidate.



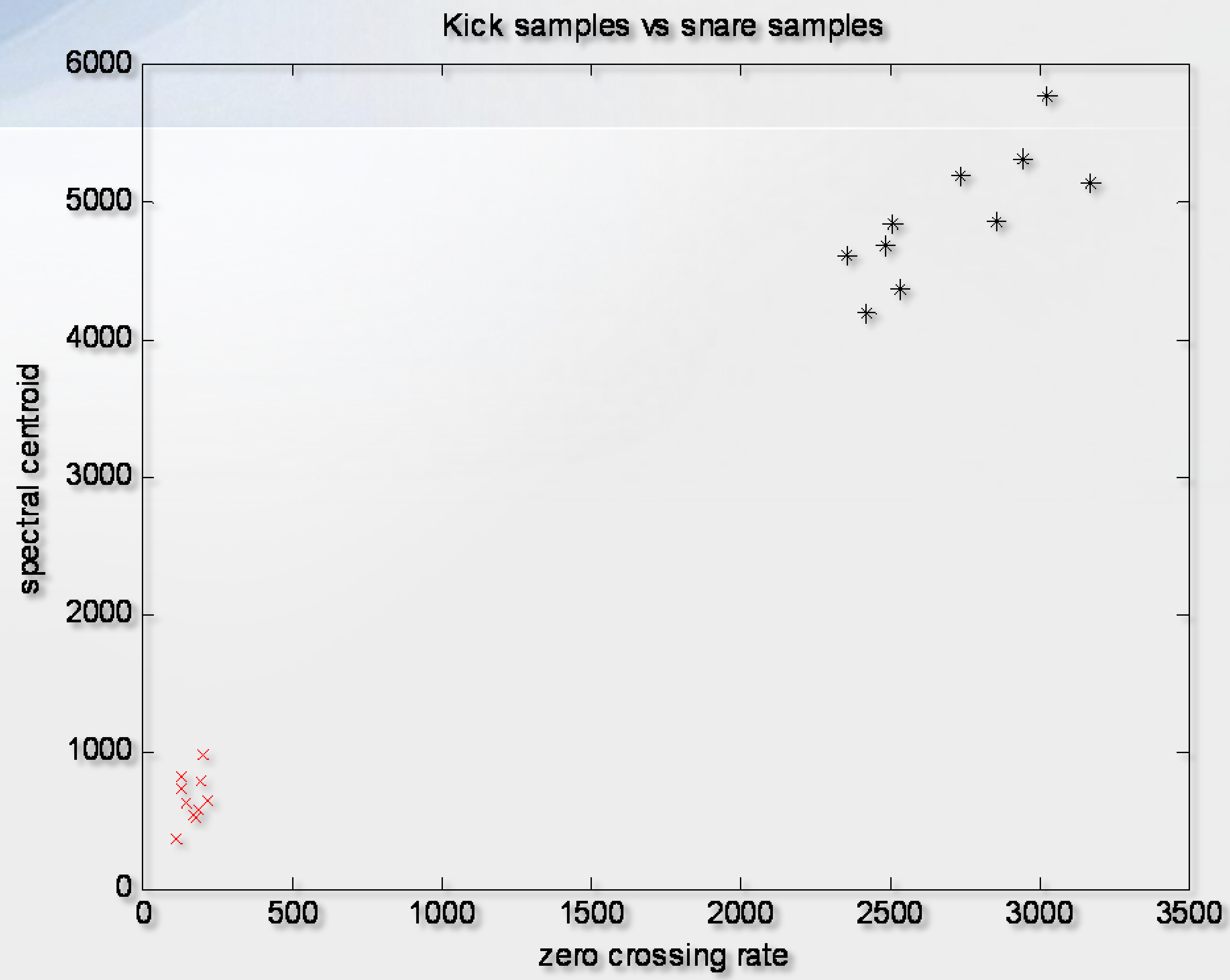
Picture courtesy: Olivier Lartillot

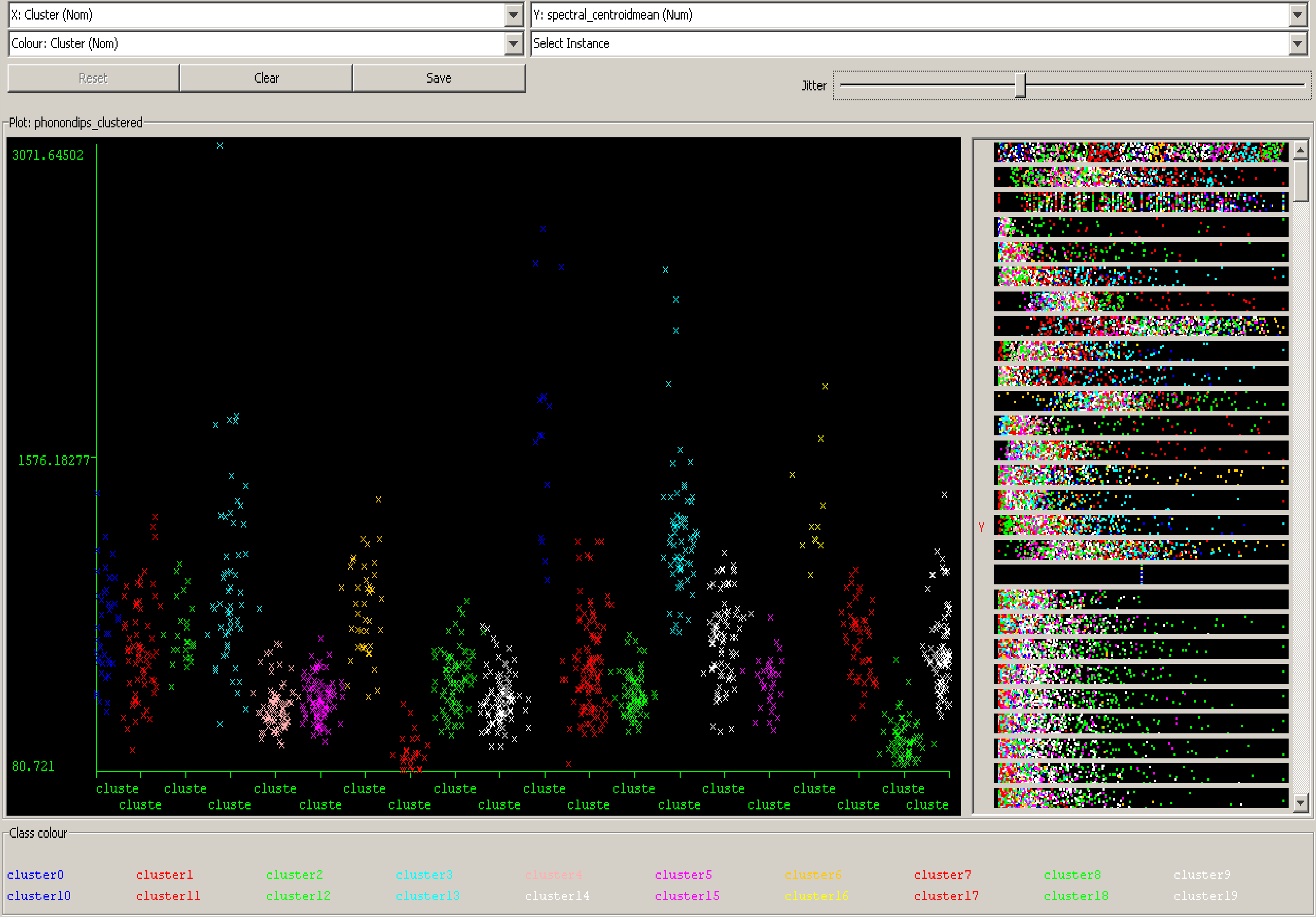


SMSTools2
File View Analysis Window Help
Input sound 1 - C:\Datos\Emilia\TestSounds\hotel_california-short.wav
0 10 20 30 40 50 60 70
HPCPGrama 1 - C:\Datos\Emilia\TestSounds\hotel_california-short.wav
A G F # # B #
0 10 20 30 40 50 60 70
HPCPMean 1 - C:\Datos\Emilia\TestSounds\hotel_california-short.wav
A # B C # D # E F # G #
HPCPKeySpace 1 - C:\Datos\Emilia\TestSounds\hotel_california-short.wav
f# d bb
Db/C# A F Db
c# a f
E C Ab
g# e c
B G Eb
eb/d# b g eb
F# D Bb
f# d bb
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9



- http://www.chordpickout.com/index.html


Kick samples vs snare samples
spectral centroid
zero crossing rate
0
1000
2000
3000
4000
5000
6000
0
500
1000
1500
2000
2500
3000
3500


X: Cluster (Nom)
Y: spectral_centroidmean (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Save
Jitter
Plot: phonondips_clustered
3071.64502
1576.18277
80.721
cluste
Y
Class colour
cluster0
cluster1
cluster2
cluster3
cluster4
cluster5
cluster6
cluster7
cluster8
cluster9
cluster10
cluster11
cluster12
cluster13
cluster14
cluster15
cluster16
cluster17
cluster18
cluster19

Day 3 Lab

- Get your Lab 2 working
  - Make sure that training data = 100% accurate
  - Try the test snares and test kicks
    - Write down your accuracy and parameters
    - Change the number of features
    - Add or replace current features with different values
      - (e.g., mirbrightness, mirrolloff)

- Demo - tonality
- Demo – tempo