Musical Genre Classification of Audio Signals

Mrugank V Alat
Pradeep Kumar Govindaraju
Sachin Mohla

What is Music Genre Classification ?

- Musical Genres are nothing but categorical labels created by us.
- Perceptually we classify music based on -

Instrumentation content | **Rhythmic structure** | **Harmonic**

What makes it a machine learning problem?

- Automating music genre classification is a necessary task in structuring and organizing music online.
- Other applications include : smartphone music/media players, PANDORA! Drinkify! & other online radio stations

Music Genre classification a challenging problem!!

- Hard to systematically and consistently describe due to their inherent subjective nature.
- Feature extraction is not straightforward as in the case of speech signals.
- Many music genres share similarity which makes feature extraction difficult.
- Genres evolve as time progresses!!

Problem Statement

Using machine learning algorithms on raw audio signals for music genre classification. We investigate relevant features extraction for our problem and discuss the classification of 10 different Music Genres using various machine learning classification techniques. We discuss all results in details and hint on relevant features particularly required for genre classification.

Feature Extraction

Timbral Texture Features -

A. Spectral Centroid

B. Spectral Rolloff

C. Spectral Flux

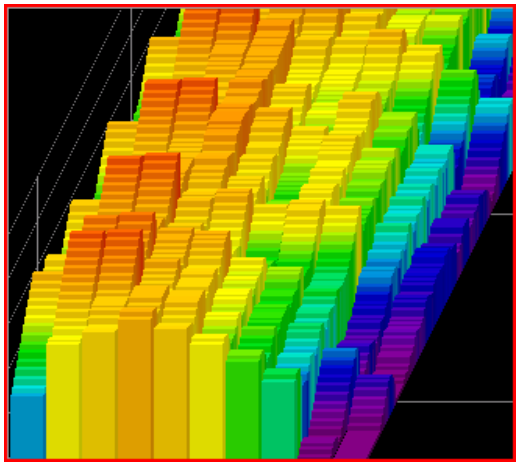
D. Time domain zero crossings

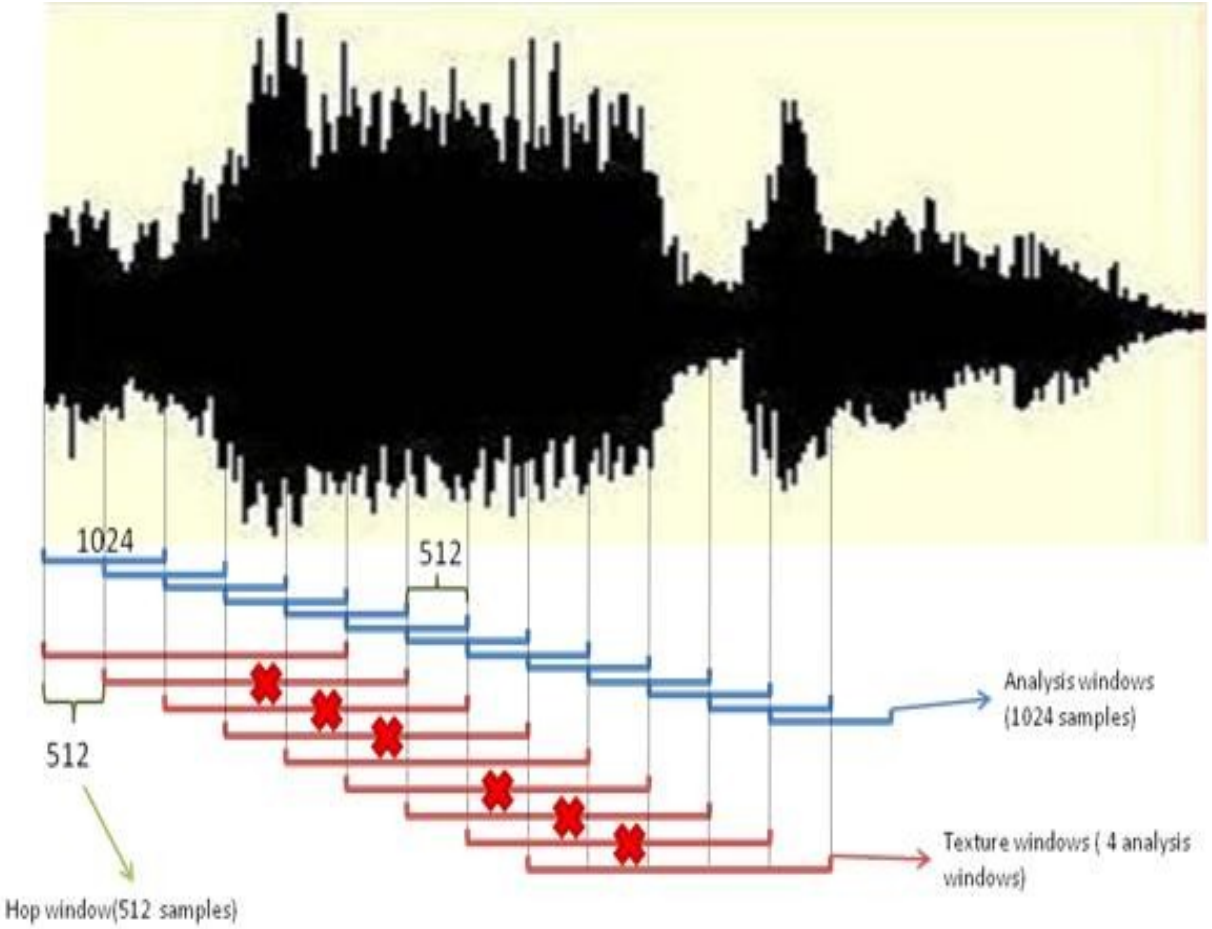
E. Low Energy

F. MFCC

Rhythmic Content Features

STFT – Short Term Fourier Transform





Features – Spectral Centroid

- Centre of gravity of the magnitude spectrum of the STFT.
- Higher centroid values => brighter texture / more high frequencies

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]}$$

Features – Spectral Rolloff

- Frequency below which 85% of magnitude distribution is concentrated.
- Measure of spectral shape

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n]$$

Features – Spectral Flux

- Measure of the amount of local spectral change.

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2$$

Features – Time Domain Zero Crossing

- Provides a measure of noisiness of the signal

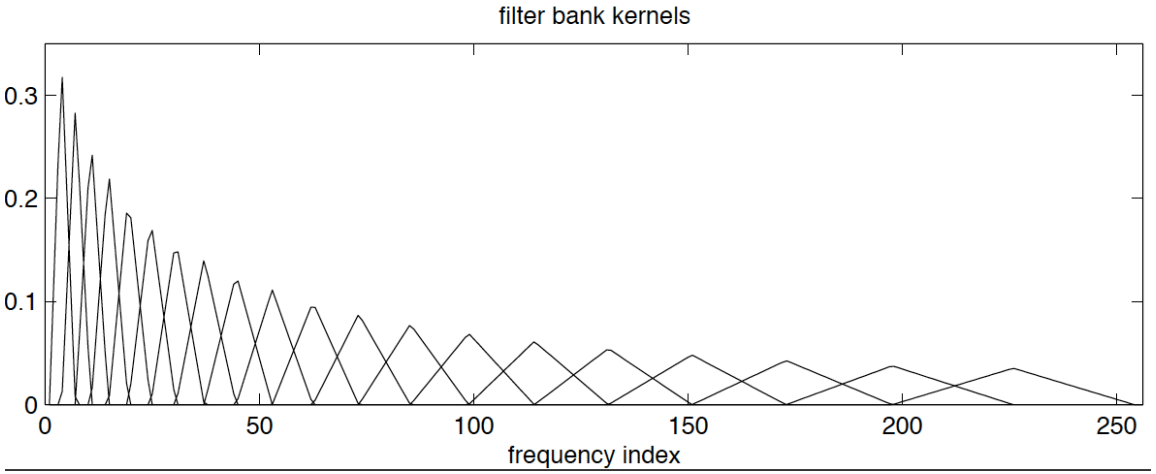
$$Z_t = \frac{1}{2} \sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])|$$

Features – Low Energy

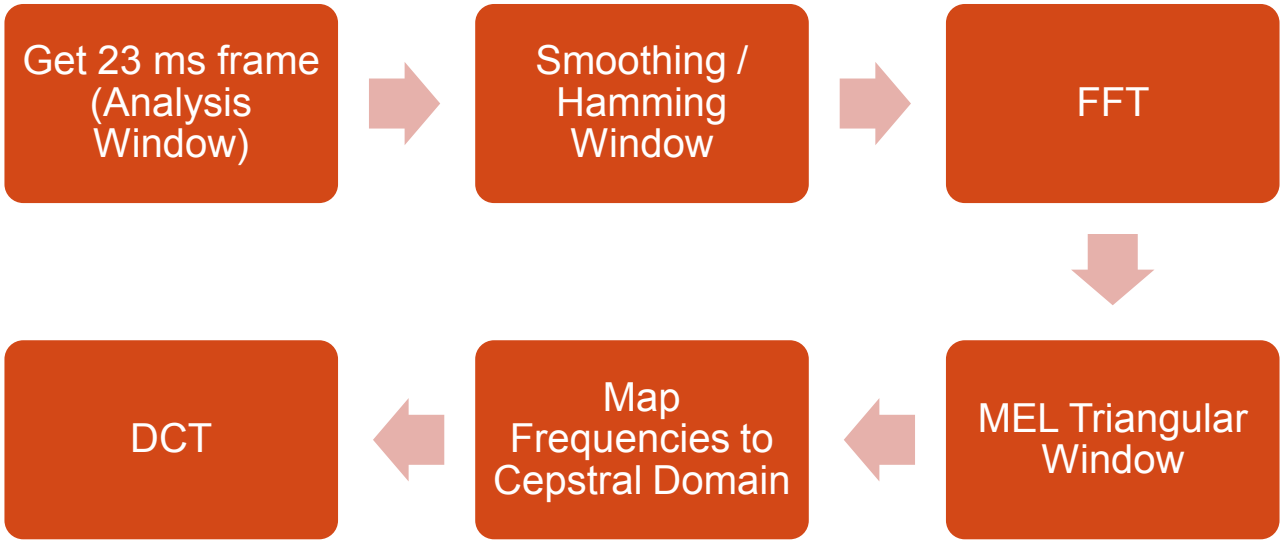
- Only feature based on texture window rather than analysis window.
- % of analysis windows that have less RMS energy across the texture window.

MFCCs

- Mel Frequency Cepstral Coefficients - perceptually motivated feature
- Gives us a model that gives a good approximation of the human auditory system.
- Representation of the short term power spectrum of a sound based on the mel scale.



MFCCs Feature Extraction



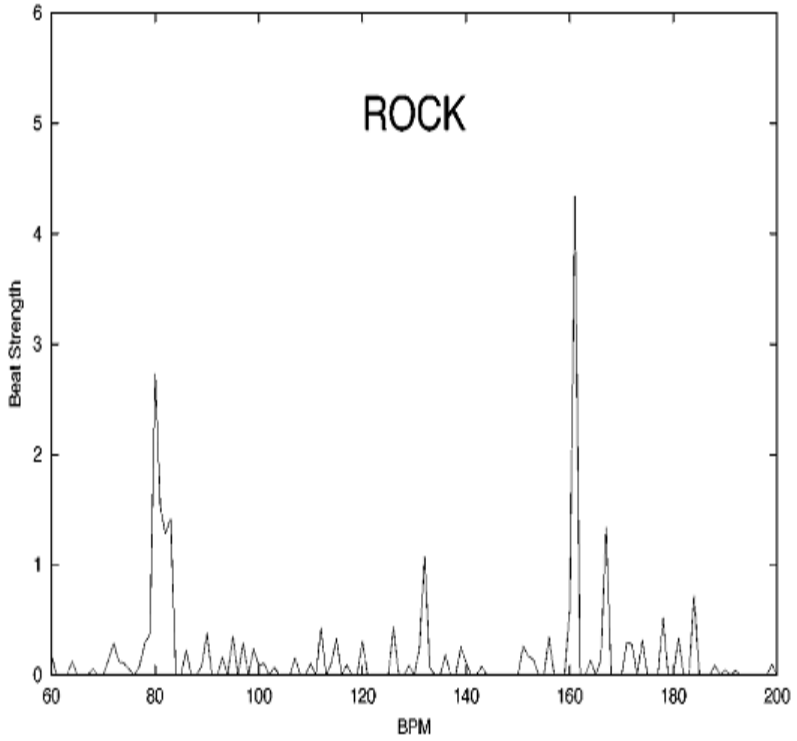
Now how do we model every song using the MFCC coefficients?

- Explaining the Analysis and Texture Windows
- To model each song we take the running statistics of every analysis frame.
- We finally model the song by fitting a multivariate gaussian distribution that gives us the mean of the MFCC's coefficient retained and their covariance

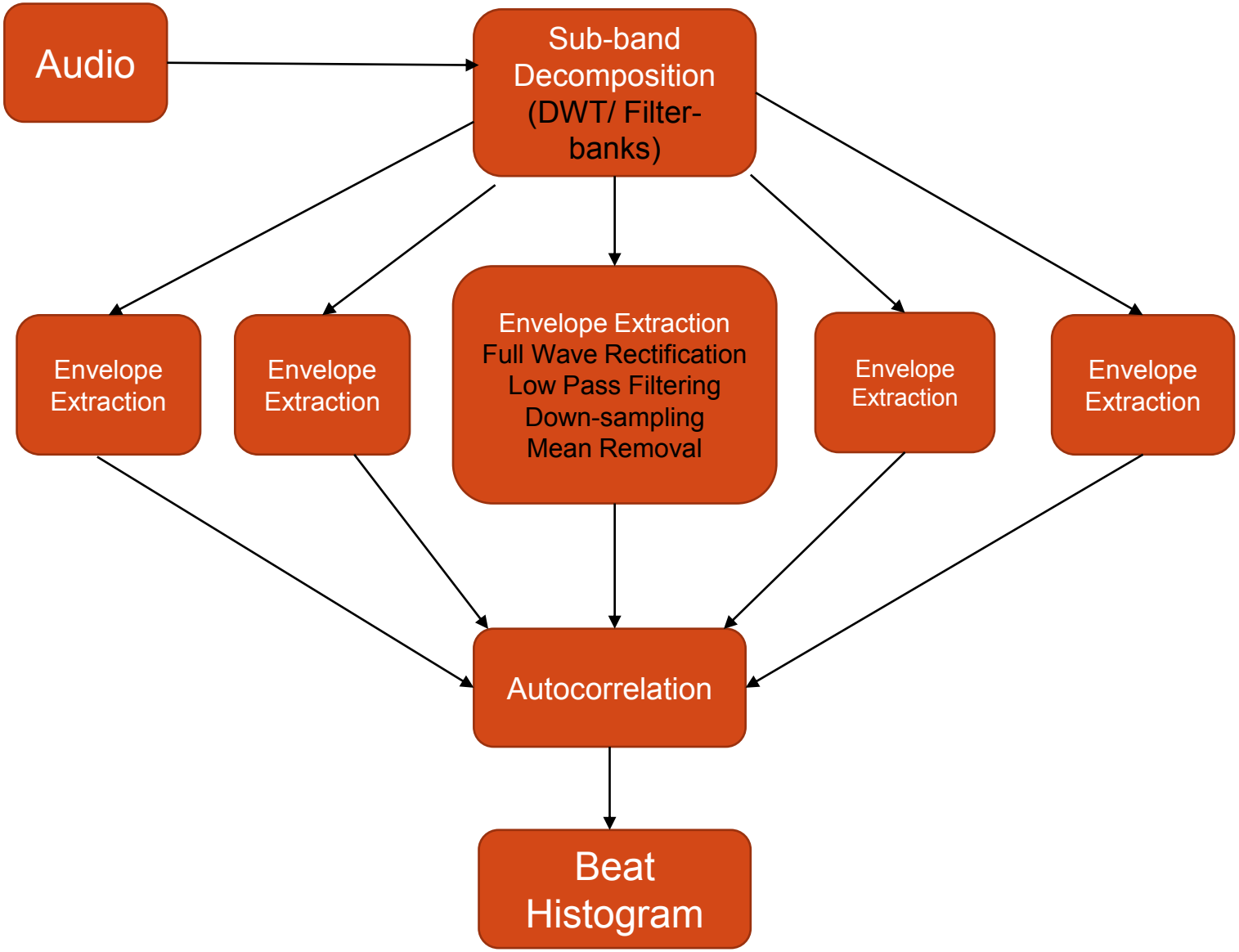
MFCC feature vector = 10 elements,
5 - mean vector , 5 - standard deviation vector

Rhythmic Content Features

- Obtained from Beat Histogram of the song.
- Used to represent main-beat and sub-beat characteristics through feature vectors.



Computing Beat Histogram



Feature Extraction

- 4-dimensional feature vector extracted from beat histogram.
  - Relative Amplitude of 1st and 2nd peak.
  - Ratio of their amplitudes.
  - Overall Sum of beam histogram.

Classification - Dataset Used

- Gtzan MARSYAS (Music Analysis Retrieval and Synthesis For Audio Signals)
- 1000 audio tracks
- 30 seconds long
- 10 genres of 100 tracks each.
- The tracks are all 22050Hz Mono 16-bit audio files in .au format.

K-nearest neighbor

Genre	Metal	Classical	HipHop	Pop	Blues	Country	Disco	Jazz	Reggae	Rock
Metal	20	-	-	-	3	-	3	-	-	4
Classical	-	28	-	-	-	-	-	1	-	1
HipHop	-	-	10	6	1	-	4	-	7	2
Pop	-	-	1	26	-	1	2	-	-	-
Blues	4	-	-	-	9	5	1	-	3	8
Country	1	2	1	-	5	14	1	1	1	4
Disco	1	-	4	1	2	1	17	-	1	3
Jazz	-	1	1	7	2	5	2	8	1	3
Reggae	-	-	4	8	1	4	-	1	11	1
Rock	6	1	6	-	5	1	2	1	-	8
Accuracy	**66.67%**	**93.33%**	**33.33%**	**86.67 %**	**30%**	**46.67%**	**56.67%**	**26.67 %**	**36.67%**	**26.67%**

Total Correct Classification : 50.33%

K-nearest neighbor

4 Genres only

Genre	Metal	Classical	HipHop	Pop
Metal	28	-	1	1
Classical	1	29	-	-
HipHop	3	-	21	6
Pop	-	-	2	28
Accuracy	93.33%	96.67%	70%	93.33%

Hard K Means

Distance metric : Kullback-Lieber (KL) Divergence

$$D = 0.5 * (KL(P||Q) + KL(Q||P)$$

Genre	Accuracy
Metal	90%
Classical	94.2%
HipHop	68.33%
Pop	88.57%

*Meaningful results only in case of 4 genre classification

SVM Classifier - Polykernel

$$K(x,y) = \left(\sum_{i=1}^{n} x_i y_i + c\right)^2 = \sum_{i=1}^{n} \left(x_i^2\right)\left(y_i^2\right) + \sum_{i=2}^{n}\sum_{j=1}^{i-1} \left(\sqrt{2}x_i x_j\right)\left(\sqrt{2}y_i y_j\right) + \sum_{i=1}^{n} \left(\sqrt{2c}x_i\right)\left(\sqrt{2c}y_i\right) + c^2$$

- Accuaracy = 63.7%

```
=== Confusion Matrix ===

 a  b  c  d  e  f  g  h  i  j   <-- classified as
83  0  1  0  5  0  4  1  0  6 |  a = metal
 0 86  3  0  0  3  0  6  0  1 |  b = classical
 2  1 40 10  1  1  6  0 22  7 |  c = hiphop
 0  0  1 75  0  3  8  7  4  2 |  d = pop
 9  1  3  0 64  6  1  5  2  9 |  e = blues
 1  2  0  7  7 60  6  9  2  6 |  f = country
 3  0  4 13  2  1 56  0  6 15 |  g = disco
 3  8  0  0  5  3  0 68  1 12 |  h = jazz
 1  0  6 13  3  7  2  2 61  5 |  i = reggae
13  2  4  4  7 11 11  8  3 37 |  j = rock
```

SVM Classifier – Polykernel [4 genres]

- Accuracy = 93.3162%

```
=== Confusion Matrix ===

  a  b  c  d   <-- classified as
 96  0  3  1 |  a = metal
  2 94  2  1 |  b = classical
  5  0 78  7 |  c = hiphop
  0  1  4 95 |  d = pop
```

Boosting SVM

- Adaboost : Iterations 1000
- Accuracy = 66.7341 %

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j   <-- classified as
 89  0  2  0  1  0  3  0  0  5 |  a = metal
  2 79  1  0  0  2  1 12  1  1 |  b = classical
  3  0 53  4  3  2 10  0 12  3 |  c = hiphop
  0  1  4 76  0  4  9  1  3  2 |  d = pop
  6  0  3  0 65  7  4  5  5  5 |  e = blues
  1  2  0  9  5 67  5  1  2  8 |  f = country
  2  1  8 11  2  3 61  1  2  9 |  g = disco
  1  9  0  1  3  7  1 71  3  4 |  h = jazz
  0  0 10  8  1  4  4  2 64  7 |  i = reggae
  9  3  3  7 10 14 14  3  2 35 |  j = rock
```

Neural nets

- Only 1 hidden layer used.
- The average classification accuracy for 10 genres is 61.47 %.

a	b	c	d	e	f	g	h	i	j	<-- classified as
85	1	3	0	4	0	3	2	0	2	a = Metal
1	75	0	0	1	2	1	13	2	4	b = Classical
4	0	45	6	5	0	6	0	21	3	c = Hip-hop
0	2	3	69	0	4	9	4	7	2	d = Pop
4	1	4	0	58	12	3	6	4	8	e = Blues
1	7	1	5	7	58	6	4	3	8	f = Country
2	2	11	15	0	4	55	0	2	9	g = Disco
1	11	0	4	6	8	0	63	1	6	h = Jazz
1	0	12	7	6	6	7	1	57	3	i = Reggae
11	3	5	2	10	12	18	5	4	30	j = Rock

Neural nets (4 genres)

- Using only 4 genres, the accuracy increases to 92.8%

```
  a   b   c   d   <-- classified as
 95   0   4   1 |   a = Metal
  3  95   0   1 |   b = Classical
  5   0  78   7 |   c = Hip-hop
  1   1   5  93 |   d = Pop
```

Neural nets implementation

	Accuracy
Beat Histogram (BH) Features (4)	24.368 %
MFCC (10)	58.4429 %
STFT (9)	47.9272 %
BH and MFCC (14)	58.1395 %
BH and STFT (13)	51.1628 %
MFCC and STFT (17)	**60.0607 %**
Complete (23)	**61.474 %**

- Number of hidden layers- 1
- Number of nodes-10
- Learning rate- 0.03

Thank you

Questions ??

One good thing about music, when it hits you, you feel no pain. ☺