SUPERVISED LEARNING FOR GENRE CLASSIFICATION OF AUDIO TRACKS

Author: Angel Bergantiños Yeste
Director: Dr. Sergio Escalera Guerrero
Deparment of Applied Mathematics and Analysis

Context and motivation

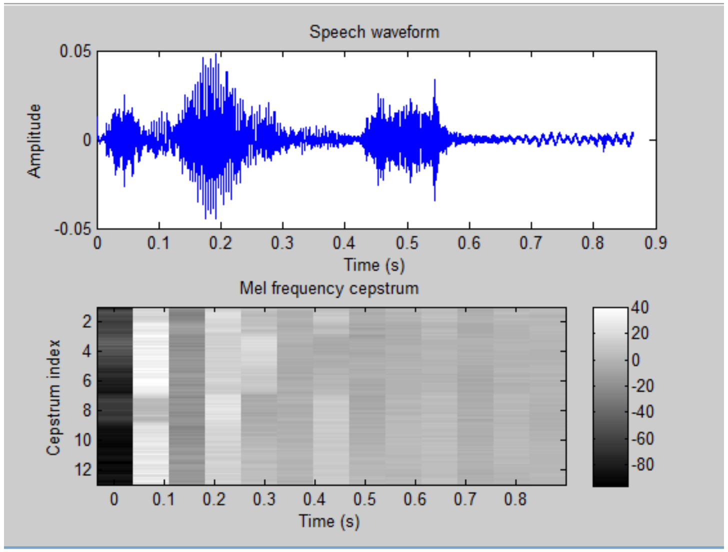
History

Telephones were the first to transmit audio through electricity

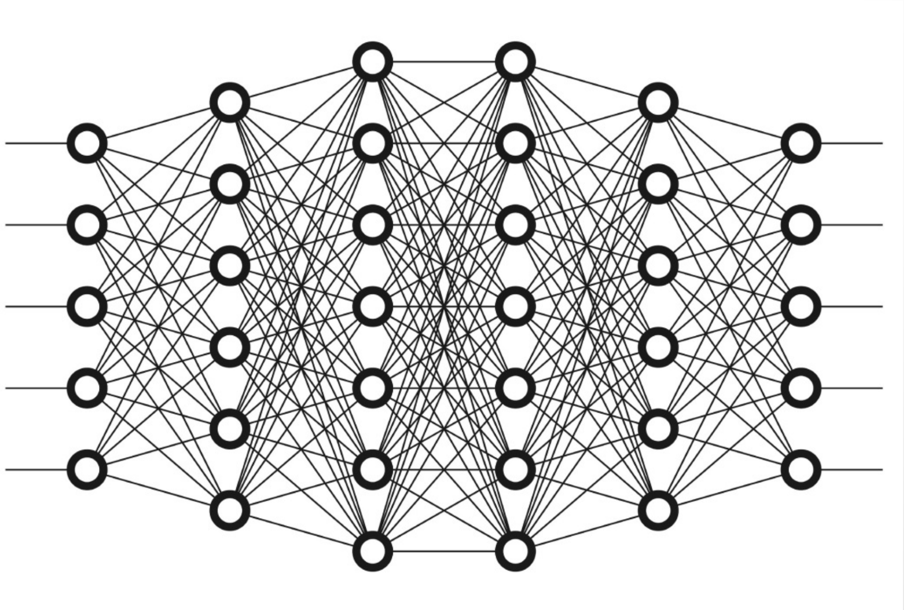
Thomas Stockham with the first digital recorder (1976)

First iPod released by Apple in 2001

State of the art



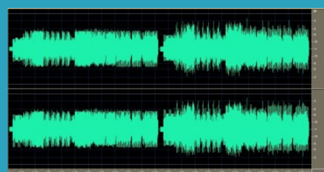
MFCC to extract information of a song's frequencies



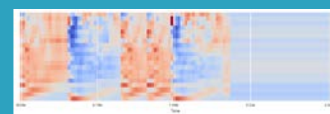
Deep learning techniques are being used in bigger studies

Method whole overview

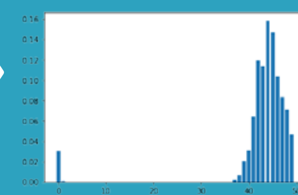
RAW SONG



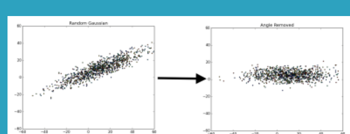
MFCC



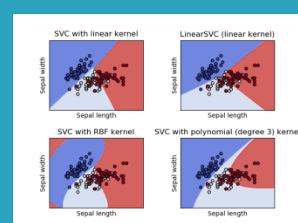
HISTOGRAMS



PCA



LEARNING

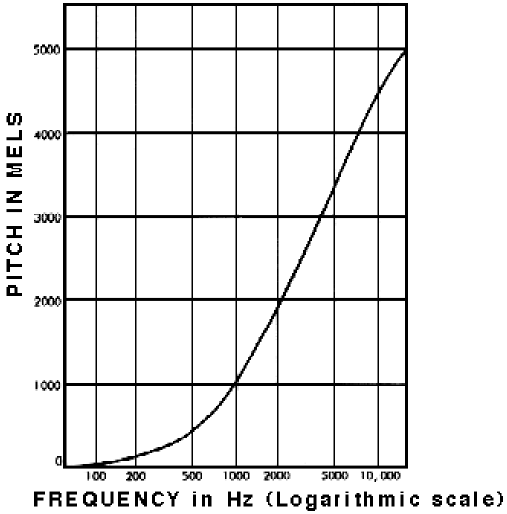


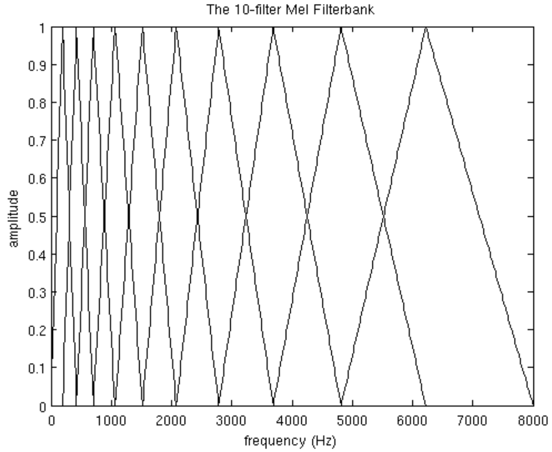
GENRE

Method: MFCC

Mel Scale

$$m = 2595 \cdot \log_{10}\left(1 + \frac{F}{7000}\right)$$



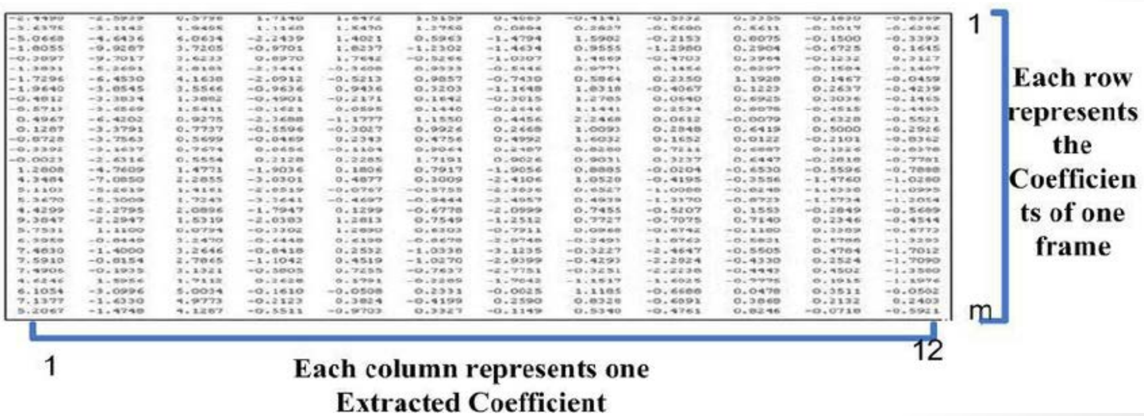


Method: MFCC

Discrete Cosine Transform:

$$C_n = \sum_{k-1}^{k} (log D_k) \cos\left[m(k - \frac{1}{2}) \frac{\pi}{k} \right]$$

Resulting matrix:



Method: PCA and Feature representation

Naive representation

Every frame is concatenated after another.

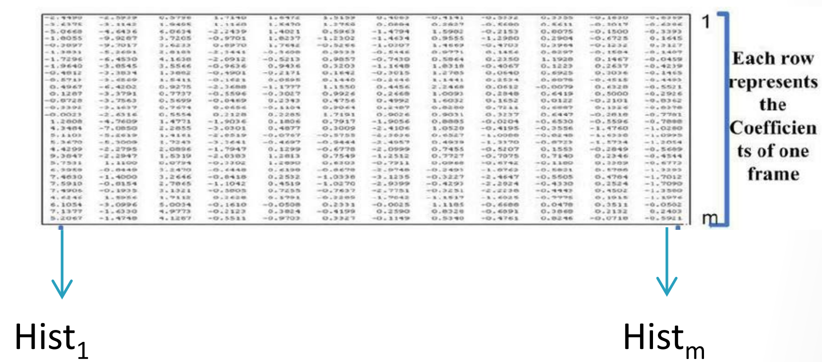
$[Frame_1][Frame_2]\ldots[Frame_N]$

Where each Frame has as many values as MFCC.

$Frame1 = [MFCC_1, \ldots, MFCC_M]$

Histogram representation

Each histogram is formed with the MFCC of every frame.



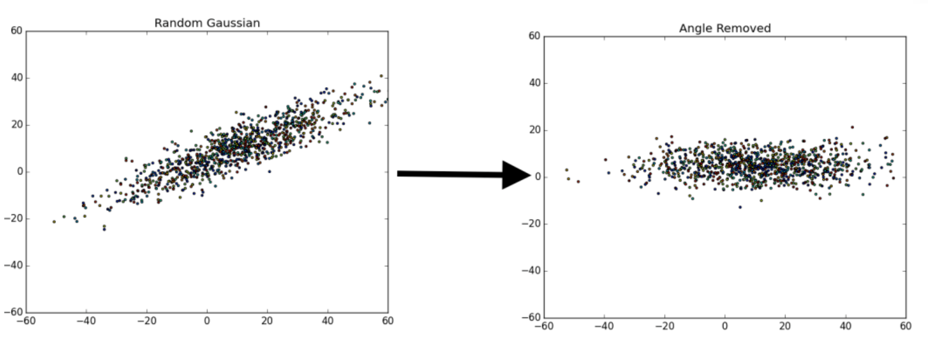
And then, concatenate all histograms.

$[Hist_1][Hist_2]\ldots[Hist_M]$

Method: PCA and Feature representation

Naive: $[\text{Frame}_1][\text{Frame}_2]\ldots[\text{Frame}_N]$

Histogram: $[\text{Hist}_1][\text{Hist}_2]\ldots[\text{Hist}_M]$

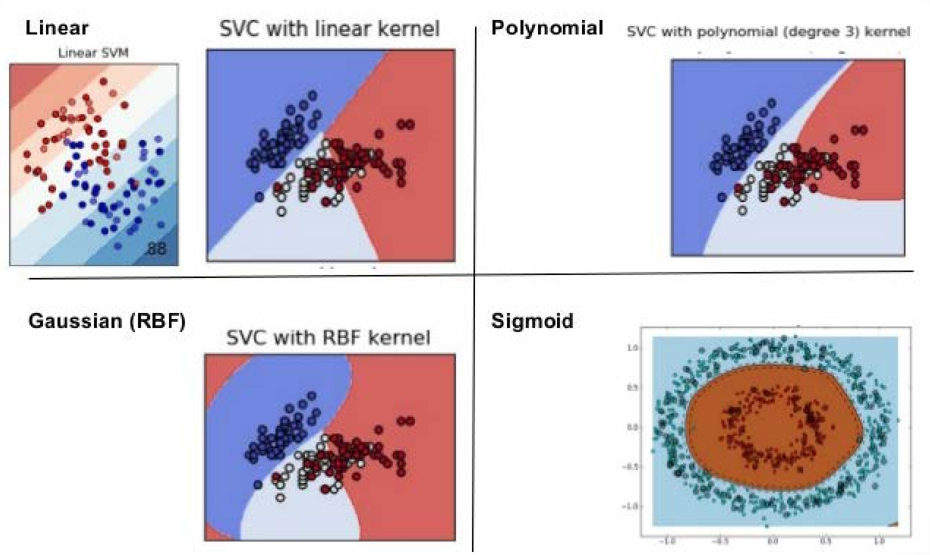


Method: learning schemes

Support Vector Machine

Different kernels:

- Rbf
- Linear
- Poly
- Sigmoid



Results: Data

Marsyas dataset: http://marsyasweb.appspot.com/download/data_sets/

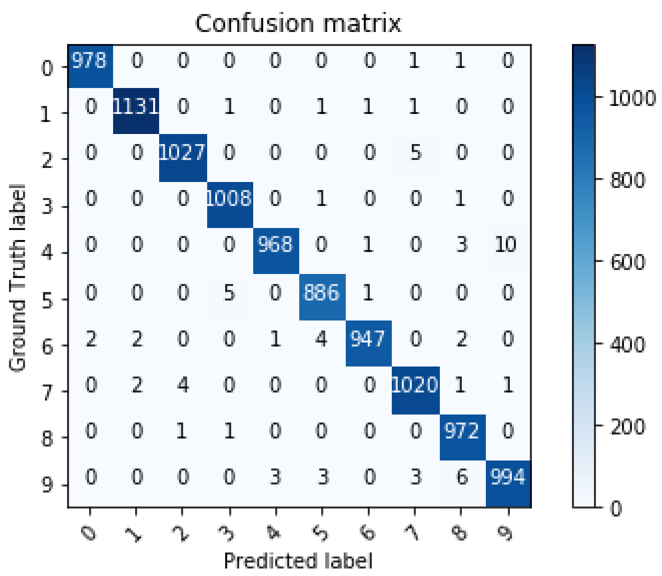
Genres

Blues	Jazz
Classical	Metal
Country	Pop
Disco	Reggae
Hip hop	Rock

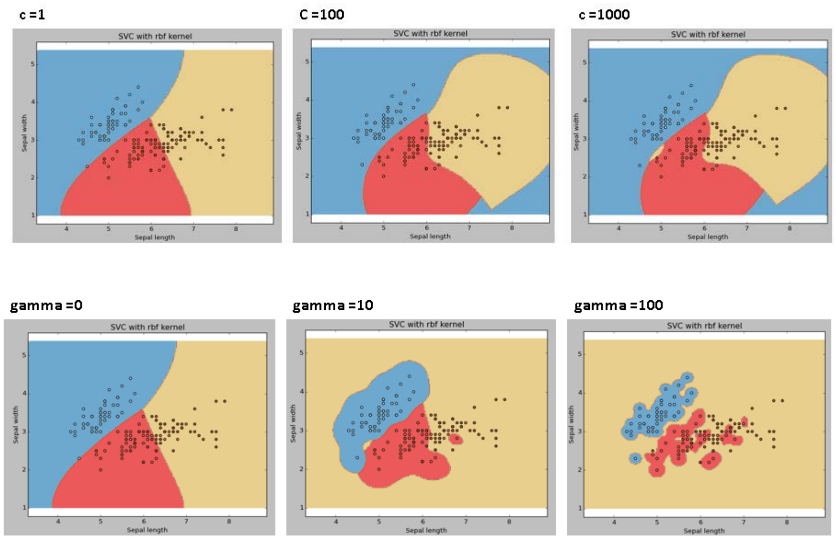
Properties

- **Sample rate:** 22050Hz
- **Channels:** 1 (Mono)
- **Frame rate:** 22050 fps

Results: Evaluation protocol and method parameters.



Confusion matrix

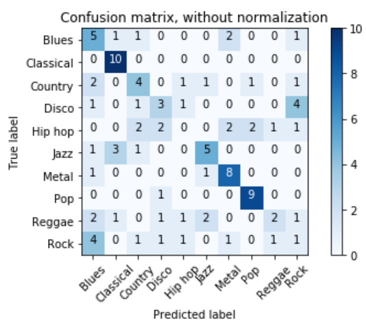


Gamma and C parameters

Results: Naive

Naïve representation without PCA

n_mfcc	rbf	Linear	Poly	Sigmoid
5	15	38	38	10
10	15	44	41	10
15	22	47	39	11
20	24	46	41	11



n_mfcc	rbf $\gamma = 10^{-7}$	Linear C=1	Poly $\gamma = 10^{-7}$	Sigmoid $\gamma = 10^{-9}$
15	50	47	42	34

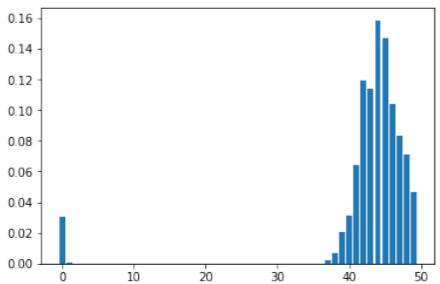
Naïve representation with PCA

n_mfcc	rbf	Linear
5	12	34
10	12	42
15	12	44
20	12	41

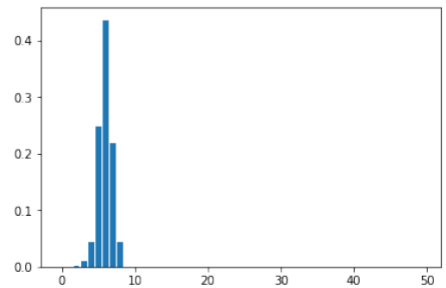
n_mfcc	rbf $\gamma = 10^{-7}$	Linear $C = 10^{-2}$
15	46	45

Results: Histograms

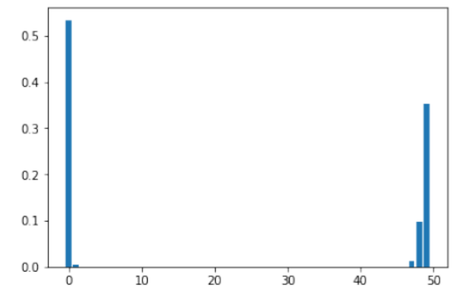
Histogram representation



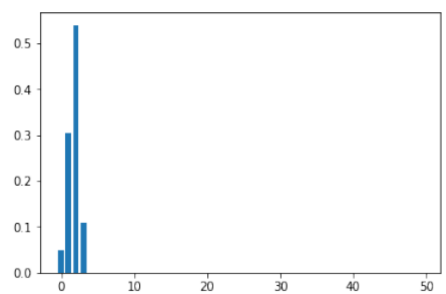
(a) Feature 1



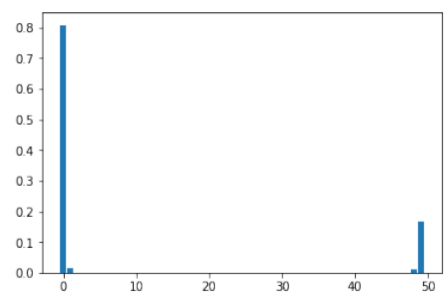
(b) Feature 2



(c) Feature 3



(d) Feature 4



(e) Feature 5

Results: Histograms

Histogram representation without PCA (linear)

n_mfcc	10	20	30	40	50	60	70	80	90
5	25	41	41	45	41	41	40	41	41
10	25	42	44	50	50	50	48	49	51
15	25	42	44	53	51	50	54	57	59
20	25	42	44	53	52	51	54	55	57

Histogram representation with PCA (linear)

n_mfcc	10	20	30	40	50	60	70	80	90
15	25	42	44	53	51	50	54	57	59

Histogram representation without PCA (rbf)

n_mfcc	10	20	30	40	50	60	70	80	90
5	24	37	33	35	34	35	31	32	32
10	24	37	37	33	35	35	33	31	31
15	24	37	37	35	37	33	31	31	29
20	24	37	37	35	37	32	30	31	30

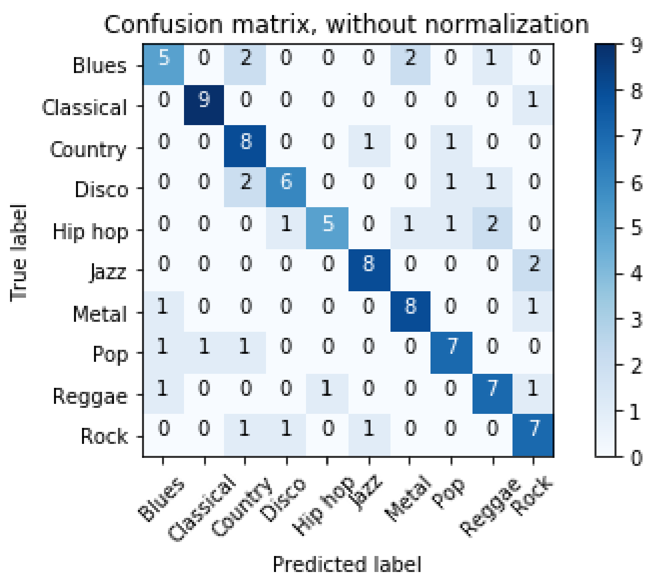
Histogram representation with PCA (rbf)

n_mfcc	10	20	30	40	50	60	70	80	90
15	24	37	37	34	38	36	34	31	30

Results: Histograms

Best results

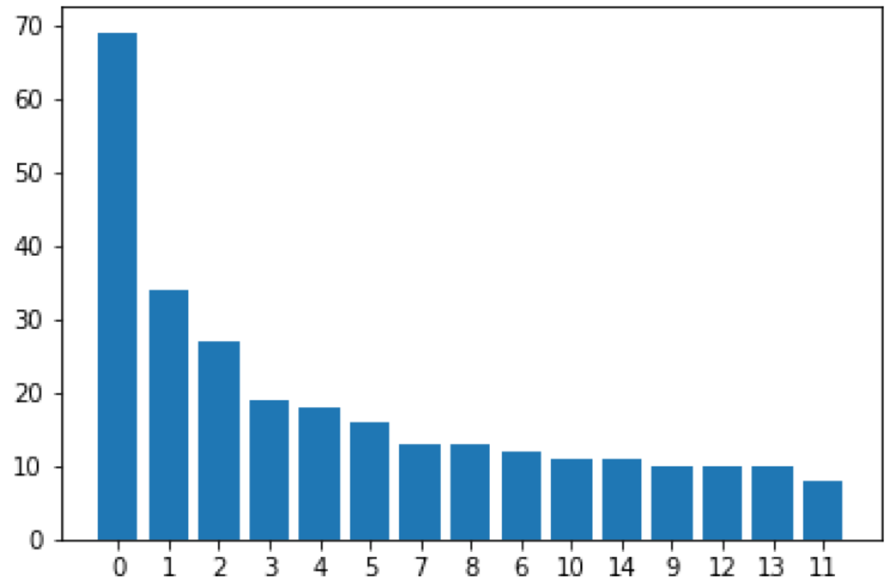
n_mfcc	rbf $\gamma = 1$	Linear $C = 10$
15	70	65



set 1	set 2	set 3	set 4	set 5
61	69	74	64	75
set 6	**set 7**	**set 8**	**set 9**	**set 10**
71	72	65	71	70

Results: Feature relevance

```
Feature ranking:
1. feature 3 (0.011840)
2. feature 0 (0.011585)
3. feature 0 (0.010474)
4. feature 3 (0.010153)
5. feature 1 (0.010096)
6. feature 8 (0.010084)
7. feature 0 (0.009707)
8. feature 10 (0.009629)
9. feature 2 (0.009549)
10. feature 1 (0.009488)
11. feature 2 (0.009082)
12. feature 0 (0.008644)
13. feature 3 (0.008399)
14. feature 3 (0.008375)
15. feature 11 (0.008210
16. feature 9 (0.008207)
17. feature 0 (0.008100)
18. feature 0 (0.008054)
19. feature 0 (0.007974)
20. feature 1 (0.007854)
21. feature 7 (0.007851)
22. feature 2 (0.007696)
23. feature 2 (0.007666)
24. feature 3 (0.007388)
25. feature 1 (0.007354)
```



Results: Extra

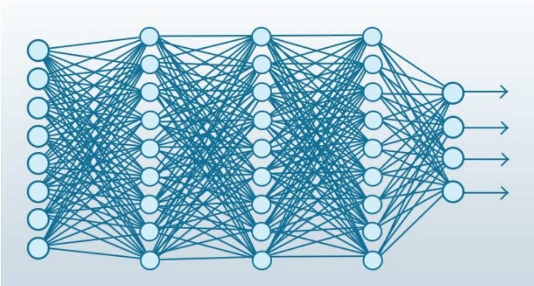
AdaBoost: Worse performance (18%)

Naïve-Bayes: Lower accuracy, but similar to SVC default results (56%)

Conclussion and future work

Bigger is better: with more songs, we can learn better.

With a bigger dataset, deep learning can work better and allows the creation of a more accurate model.



Other features of music can be also used as input to work with the frequency for further audio analysis.

Thanks for your attention!