Enhancing Music Genre Classification with Neural Networks by using Extracted Musical Features

Denys Flederus
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
d.r.flederus@student.utwente.nl

ABSTRACT

The use of mel-frequency cepstral coefficients (MFCCs) has proven to be a powerful tool in music and voice recognition, and sound recognition in general. This paper is focused on investigating what data we can use along with MFCCs to increase the accuracy of music genre classification. The results of this process are analyzed to gain insight into the characteristics of different music genres. In this paper, MFCCs are considered for music genre classification using a multilayer perceptron (MLP) neural network. Measuring the effect of augmenting MFCCs with additional audio features at the input of the MLP. Following this there is an analysis of the effects different features, e.g. zero-crossing rate, spectral bandwidth etc., have on the accuracy of classifying genres and what the results show about the similarities and relatedness of music genres. Finally, an analysis of the results of classifying a selection of songs from the metal genre.

Keywords

multilayer perceptrons, neural networks, music genre recognition, mel-frequency cepstral coefficients, music genres

1. INTRODUCTION

Music taste is highly subjective and styles of music range widely. Therefore, music has historically been grouped into genres and even a hierarchy of subgenres, with new genres and subgenres being recognized very often due to innovation, experimentation, new technologies and combining existing genres. This grouping of music into genres helps people more easily discover music that is similar to the music they already listen to. Online availability of music and increasingly growing local storage options make us able to discover and listen to more music.

Classifying music genres is hard because the distinction between genres is nonlinear highly dimensional. Extracting features from the music files should help distuingish between music genres. This paper explores the effect of several musical features on the accuracy of music genre classification. To do this MFCCs are used as a basis to train a simple neural network. Then, the features are individually appended to the MFCC input and their accuracies are compared. Well-performing features are subsequently combined and added to the MFCC input to attempt to further increase accuracy. Additionaly, a few metal songs from outside the dataset are evaluated with the final best performing neural network and are analyzed and discussed.

32nd Twente Student Conference on IT Jan. 31st, 2020, Enschede, The Netherlands.

Research Questions.

RQ1 Which features of music can best be used in addition to MFCCs to define the genre of a musical piece?

RQ2 Which genres are the most similar based on their musical content?

In this paper, the following questions will be answered. RQ1 will be answered in the process of applying neural networks with differing inputs to a dataset. The answer of RQ2 will follow from analysing the results produced by the methods used in RQ1.

Contribution.

This research attempts to provide a deeper understanding of content-based musical analysis. A better understanding of this allows music services to better cater to a users' musical taste by being able to provide the user with more relevant or even unknown but interesting music, promoting the exploration of a wider array of music. While the focus in this research is on existing and established genres, it also opens up oppurtunity to extend the same ideas to more functional genre classes that correlate directly to, for example, people's moods. This research also shows the effect of every explored feature on every genre, allowing a better understanding of which musical features can be uniquely used to help classify specific genres.

2. BACKGROUND

2.1 Probabilistic classifiers

The field of machine learning contains many kinds of algorithms. This paper will exclusively consider the class of probabilistic classifiers. Instead of generating a single output, probabilistic classifiers generate a probability value for every possible output. In practise, many classifiers can be adapted to give a probabilistic output for each class considered [5]. The concept of neural networks comes from the way how neurons function in the brains. The network receives some input which gives a signal to the neurons in the next layer. This signal is processed and changes based on a mathematical function. Each neuron then propegates their own signal to the next set of neurons where the same thing happens until the end of the network is reached, resulting in a final output. This neural network learns by

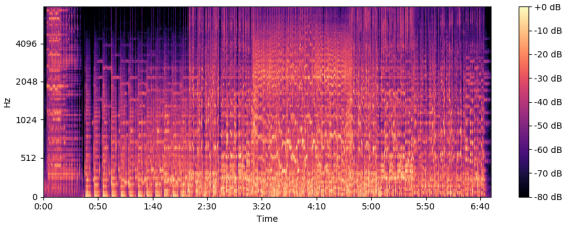


Figure 1: Mel-frequency spectrogram generated from *.wav*-file converted from *.mp3*-file: Pink Floyd - Time [1973]

training the neurons, tweaking their internal parameters, using known training data as a feedback measure. After being trained for a number of epochs, every epoch represents the training data once. This network should be able classify unknown data. There are many different ways to design and set up a neural network.

2.2 Music genres

This research will consider ten wide genres that most known western music could be classified as. The genres are as such: *Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae* and *Rock.*

Cultural and historical context differ widely from genre to genre. In practise some genres contain a much wider array of different kinds of music than others. Some genres are closely related, like blues, rock and metal. Rock originated from blues and metal was derived from rock. Whereas other genres can be very distinct such as classical which covers much older music than the other genres.

2.3 Musical characteristics

So what are the features that we can measure and base our classifications on? The accuracy of music genre classification is highly reliant on proper feature extraction. Tzanetakis and Cook considered: *spectral centroid, spectral rolloff, spectral flux, time domain zero crossings, mel-frequency ceptral coefficients, analysis window, texture window and low-energy feature* in their 2002 paper [11]. While this research did not involve machine learning, later research in 2018 still considers many of these features [2].

These features are extracted by applying the Fourier transform on an audio file, like *.mp3* or *.wav*. From this Fourier transform you can also visualize the music by extracting the mel-frequency cepstral coefficients and plotting sound frequencies on a y-axis and time on the x-axis with sound intensity displayed on a third axis represented by a colored heatmap as seen in figure 1.

Mel-frequency cepstral coefficients.

MFCCs are the $\log_2$-scaled powers of the Fourier transform on a mel scale, as opposed to a hertz scale. This scaling represents audio in a way that a human listener would interpret audio, because humans do not perceive volume and pitch linearly.

3. RELATED WORK

3.1 Music Genre Classification

The first attempts at content-based music genre classification were published in 2002 and 2003 most notably the following articles:

- *Musical genre classification of audio signals* [11].

- *Factors in automatic musical genre classification of audio signals* [7].

- *A Comparative Study on Content-Based Music Genre Classification* [6].

In these papers, binary classifiers such as support vector machines were used as a, then novel, way of using machine learning to classify music genres. However, the first paper still relied on statistical analysis of musical features, reporting a 61% accuracy rate. The results in the second article, at a 71% accuracy rate using Linear Discriminant Analysis [7], were not very impressive, compared to the reported 70% accuracy rate humans are able to classify genres [9]. Later that year, the third article was published and reported just under 80% accuracy on the same dataset[1] using support vector machines on Daubechies wavelet coefficients [6].

Support vector machines have very effective in music genre classification. More recent research focuses on further aspects of this topic like the article *Musical genre classification using support vector machines and audio features* [8]. However, my paper is more related to studies like *Long short-term memory recurrent neural network based segment features for music genre classification* [1], making use of neural networks. The above study in particular aims to improve music genre classification performance using recurrent neural networks, or more specifically long short-term memory (LSTM).

4. METHODOLOGY

This research uses the GTZAN[2] dataset containing 100 30-second long music files for each music genre discussed in subsection *Music Genres* of section *Background*, totaling 1000 music pieces. The music files are single channel *.wav* files at 22050 Hz. Furthermore, the following technologies are used to analyze the music files and construct learning algorithms:

- *Python 3.6*[3].

- *Librosa*[4], a Python library for audio analysis.

- *Tensorflow*[5], a Python library for machine learning.

- *Keras*[6], an API for Tensorflow.

In this research, machine learning methods are used to tackle the known music genre classification problem. I examine the effects that using different musical features as input have on single-label classification accuracy and I analyze the confusion matrix generated from a trained model.

F1 scores.

The most important statistic extracted from a confusion matrix is the F_1 score. This is a function of the precision, the amount of true positives divided by the amount of samples in that class, and recall, the amount of true positives divided by the total positives of that class across all classes. F_1 is related to accuracy but reflects on how certain you can be that the predicted class is the correct class. A high difference in accuracy and F_1 score comes from a difference in type I errors, false positives, and type II errors, false negatives.

[1]The same dataset is used in this paper.
[2]http://marsyas.info/downloads/datasets.html
[3]https://www.python.org/
[4]https://librosa.github.io/librosa/
[5]https://www.tensorflow.org/
[6]https://librosa.github.io/librosa/

4.1 Features

The following features are used along with MFCCs to train neural networks and are directly extracted using the *Librosa* library. Figure 8 in the appendix shows each feature graphed for one song of each genre.

Zero-crossing rate.

Zero-crossing rate or zcr is a measure of how many times the signal goes from negative to positive or vice versa within some timeframe. Zcr is calculated with the following formula:

$$zcr = \frac{1}{n-1}\sum_{i=1}^{n-1} 1_{\mathbf{R}<0}(s_i s_{i-1})$$

Where s is a signal of length n and $1_{\mathbf{R}<0}$ an indicator function that returns 0 or 1. The value of the zcr tends to be higher for percussively intense music and lower for music that lacks percussive elements.

Root-mean-square.

Root-mean-square or rms is defined directly by the mathematical function of taking the square root of the mean of the square of each value in the set of considered values.

$$rms = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + ... + x_n^2)}$$

Rms describes the the surface area between the x-axis and the signal line which can be interpreted as the energy of the signal. A loud signal, goes high on the y-axis, will have higher rms and therefore contain more energy.

Spectral centroid.

Spectral centroid [4] measures the weighted mean of frequencies in a signal, the weights come from the spectrogram.

$$centroid(t) = \frac{\sum_{i=0}^{n-1} S[k_i, t] \times freq[k_i]}{\sum_{j=0}^{n-1} S[j, t]}$$

Where t is a timeframe, n is the number of frequency values per timeframe, S is the spectrogram generated by the Fourier transform and the $freq$ describes the array of frequencies in the k^{th} row of S. Spectral centroid is associated with the brightness of a sound, the presence of higher frequencies.

Spectral bandwidth.

Spectral bandwidth measures the difference between the highest and lowest frequencies within a timeframe [4].

$$bandwidth(t) = \sum_{i=0}^{n-1} S[k_i, t] \times (freq[k_i, t] - centroid[t]^p)^{\frac{1}{p}}$$

Where the variables and functions are the same as described in the *Spectral centroid* section and p is defaulted to two, resulting in the second order spectral bandwidth. This is normalized to a weighted standard deviation of the differences of the highest and lowest frequency values. Spectral bandwidth describes the dynamic range.

Spectral contrast.

Spectral contrast divides the spectrogram into seven bands. Every band contains a value representing the energy contrast, which is a function of the mean, peak (highest) and valley (lowest) energies. A more in-depth definition is given by Jiang et al. [3]. Higher spectral contrast signifies a clearer sound where low spectral contrast indicates a noisier or more muddled sound.

Spectral flatness.

Spectral flatness [10] is closely related to decibel, it describes the tonality of music. It is defined as the ratio of the geometric mean to the arithmetic mean of a power spectrum.

$$flatness = \frac{\sqrt[t]{\prod_{i=0}^{t-1} x(i)}}{\frac{\sum_{i=0}^{t-1} x(i)}{t}}$$

Where $x(i)$ returns the amount of frequency values in the timeframe t that fall within a certain frequency. A low spectral flatness signifies a clearly distinguishable tone and a high spectral flatness indicates noise, or a combination of a lot of different tones with a high variance in frequency.

Spectral roll-off.

The spectral roll-off for some frame is the frequency number below which some percentage of the spectral energy is concentrated (85% by default).

Chroma.

The chroma feature consists of twelve bands pertaining to the musical notation of notes; **C, C#, D, D#, E, F, F#, G, G#, A, A#, B**. All frequencies within some timeframe are associated with one of these bands and added to their band on a time scale. For example, 440 Hz is the frequency of the A note of the fourth octave, 220 Hz and 880 Hz are also an A note of a lower and higher octave respectively. This feature does not care about octaves. Guitar music specifically, usually plays in the key of E, due to the fact that the lowest open string is tuned to E. You would expect that the E and it's minor (G) or major (G#) third and the dominant fifth (B) to have a high presence in the chromagram of guitar music.

5. RESULTS

5.1 Preprocessing

Using only the audio data as a floating point time series, 22050 numbers per second, the trained neural network, as described in the next section, but with an input layer of size 22050, reached a accuracy of 83.91% on the training set and a 26.23% accuracy on the the testing set. This is a case of enormous overfitting, meaning that the network only works on data it has already seen. In practise, MFCCs, which are directly extracted from this time series performs much better at a test accuracy of 74.37% as shown in table 2.

First-off, extract the mfcc's from the 1-second clip, this gives a matrix of shape (20, 44). The librosa function defaults to 20 MFCCs per timeframe and there are 44 timeframes per second (22050 Hz gives ∼500 datapoints per timeframe). After extracting, the data is normalized to numbers between -1 and 1. Then the (20, 44)-shaped matrix gets flattened, resulting in a one-dimensional array of size 880. This is the input for the neural network. Other features are combined with MFCCs, these featurues are processed the same way and appended to the end of the MFCC array. The ten genres are one-hot encoded as specified in table 1.

blues	**1**	0	0	0	0	0	0	0	0	0
classical	0	**1**	0	0	0	0	0	0	0	0
country	0	0	**1**	0	0	0	0	0	0	0
disco	0	0	0	**1**	0	0	0	0	0	0
hip-hop	0	0	0	0	**1**	0	0	0	0	0
jazz	0	0	0	0	0	**1**	0	0	0	0
metal	0	0	0	0	0	0	**1**	0	0	0
pop	0	0	0	0	0	0	0	**1**	0	0
reggae	0	0	0	0	0	0	0	0	**1**	0
rock	0	0	0	0	0	0	0	0	0	**1**

Table 1: One-hot encoding of genres.

When the data of all 30.000 1-second clips is processed they are randomly shuffled into a 80/20 split of training and testing data.

5.2 Neural Network Model

The actual neural network is not the focus in this research. The model used is a MLP with an input layer of size 880 + 44×(additional features), 880 comes from the (20, 44)-shaped matrix generated by extracted MFCCs from a 1-second clip and 44 comes from the size of additional features used as input, 12 for *chroma* and 7 for *spectral contrast* and 1 for all other features.

Increasing the sample size of the dataset.

The GTZAN dataset only offers 1000 clips to train on, to artificially increase the size of the dataset I ended up cutting every 30-second clip into evenly sized 1-second clips, resulting in a 30 times larger dataset. You can see this drastically improved to accuracy on training on MFCCs while also reducing overfitting in figure 2. The implication is that only the short-term characteristics of the features is being considered, long-term context lost.

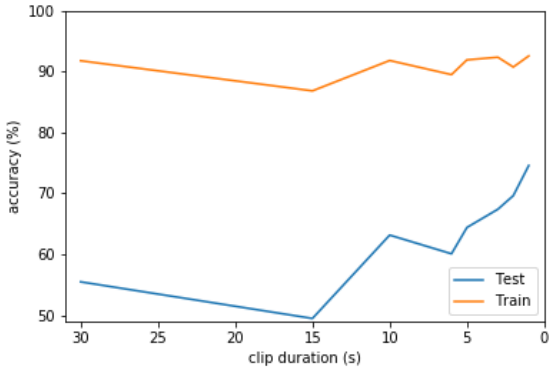


Figure 2: Test accuracy increases as the clip duration becomes lower and the dataset increases in size. Datapoints are at 30, 15, 10, 6, 5, 3, 2 and 1 seconds.

Network architecture.

The model has three hidden layers of size 400, 200 and 100 neurons respectively. Other options for the configuration of hidden layers where: 600-300-150, 1600-800-400, 800-400-200 and 1200-600-300. Which performed similarly, but were much slower in execution. Extra hidden layers beyond the third did not noticably improve accuracy anymore.

I tried different activation functions and optimizers as well as different network depths with varying amounts of neurons per layer. All layers are densely connected. For simplicity's sake, in multilayer models, every next layer is half the size of the previous one. This makes sense because the network is working towards a small output of ten neurons. I ended up using the ReLu activation function instead of Softplus for the hidden layers. They resulted in very similar accuracy, but Softplus is much slower due to the more complex mathematics involved. The output layer uses the Softmax activation function.

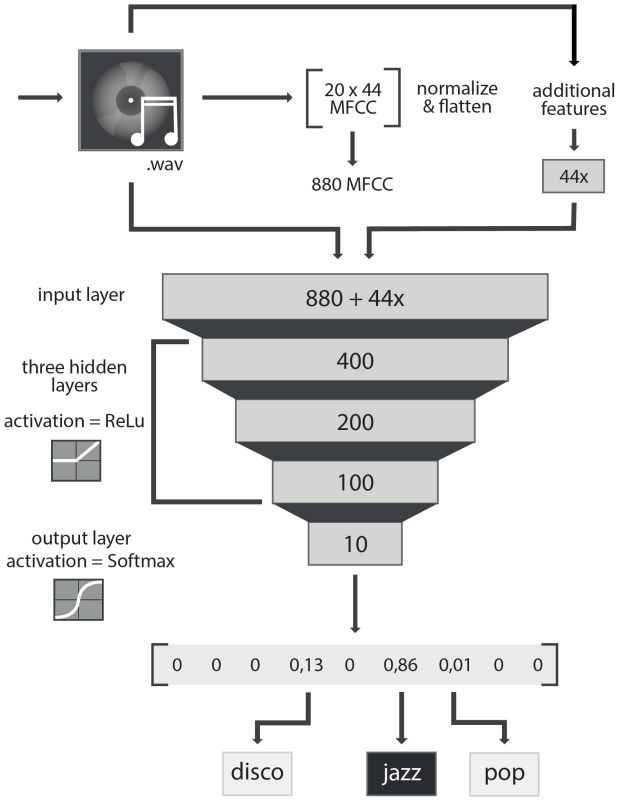


Figure 3: Visual representation of the neural network architecture and how a 1-second clip is prepared for- and processed by- the neural network.

The neural network outputs a one dimensional matrix of size 10 with numbers between 0 and 1, being the probability or level of certainty per respective class. The class for which this value is the highest will be considered the output. This is for 1-second clips. For classifiying longer clips preprocess them into 1-second clips, extract the features and feed them into the network second by second. Then combine the results by taking the average probability for every output class.

Furthermore, cross validation is applied during testing with a 80/20 split, batch size is 32 and loss is calculated by categorical crossentropy which works with one-hot encoding. The neural networks trains for 20 epochs. Figure 4 shows that further training does not improve accuracy.

Average accuracy and standard deviations are calculated from the results of the neural network trained on the features by training the neural network five different times on a different random training and testing split each time. Every feature that does not improve on the accuracy significantly, is not further considered.

5.3 Testing results

Table 3 shows the results of the average performances over five runs of neural networks trained with MFCCs and the repective feature in the table. The test averages high-

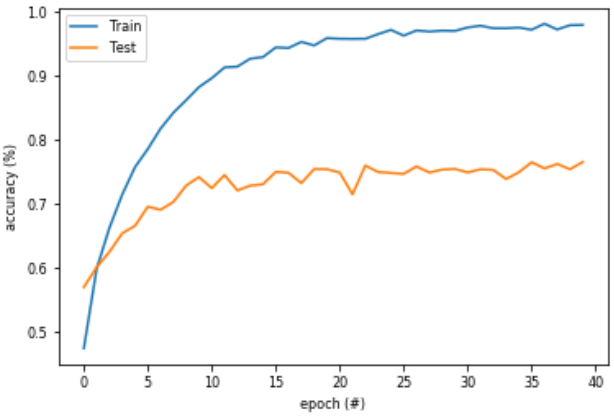


Figure 4: Accuracy of training on MFCCs over number of epochs using the GTZAN dataset cut into 1-second clips.

lighted in bold perform at least one standard deviation bettee in accuracy than using only MFCCs and input and are considered for further improving accuracy by combining them at input. Therefore, table 4 shows the results of combining MFCCs with *zero-crossing rate* and textitroot-mean-square at the input of the neural network. This further improves the accuracy by 0.94% compared to zero-crossing rate with MFCCs, 0.65% compared to root-mean-square with MFCCs and an overall increase of 1.89% compared to only using MFCCs. This shows that there is a positive correlation bewteen zero-crossing rate and root-mean-square and the underlying characteristics of music genres.

features	train avg	train stddev	test avg	test stdev
mfcc	91.38	0.86	74.37	0.92

Table 2: Accuracy and standard deviation of using only MFCCs.

features	train avg	train stddev	test avg	test stdev
zcr	92.23	0.96	**75.32**	0.56
rms	91.96	0.73	**75.61**	0.60
centroid	91.09	1.07	74.33	0.41
bandwidth	91.14	0.72	74.50	0.99
contrast	91.40	0.52	74.78	1.14
flatness	91.75	1.07	73.99	1.04
roll-off	90.86	0.60	74.57	0.68
chroma	91.33	0.89	73.18	0.64

Table 3: Average accuracies and standard deviations of MFCCs augemented with features.

features	train avg	train stddev	test avg	test stdev
zcr+rms	92.79	0.67	76.26	0.81

Table 4: Average accuracies and standard deviations of MFCCs augmented with zcr and rms.

5.4 Results - confusion matrices

Figure 5 shows the confusion matrices on the testing data. For every genre it shows what percentage was correctly classified, the middle downward diagonal, and how much was wrongly classified. Along with the standard deviations because the data was taken as the accuracies of five neural networks per feature.

Table 6 shows the F_1 scores for every genre, calculated from the confusion matrices from figure 5. You can see

Table 5: Statistics from just **MFCC** confusion matrix figure 5a.

genre	precision	type I	type II	recall	F_1 score
blues	77.80	22.20	23.63	0.77	0.77
classical	88.68	11.42	8.78	0.91	0.90
country	63.90	36.10	29.46	0.68	0.66
disco	68.60	31.40	29.26	0.70	0.69
hip-hop	68.88	31.12	29.50	0.71	0.70
jazz	80.90	19.10	34.70	0.70	0.75
metal	84.38	15.62	15.26	0.85	0.85
pop	83.44	16.56	23.80	0.78	0.81
reggae	64.88	35.12	23.34	0.74	0.70
rock	61.50	38.50	40.62	0.60	0.61

Table 6: Statistics from **zero-crossing rate** & *root-mean-square* confusion matrix figure 5j.

genre	precision	type I	type II	recall	F_1 score
blues	80.90	19.12	20.72	0.80	0.80
classical	88.14	11.88	8.88	0.91	0.89
country	69.14	30.90	29.74	0.70	0.70
disco	71.48	28.52	31.78	0.69	0.70
hip-hop	68.34	31.56	23.58	0.74	0.71
jazz	79.58	20.46	22.02	0.78	0.79
metal	84.36	15.64	10.64	0.89	0.87
pop	83.78	16.24	19.90	0.81	0.82
reggae	70.84	29.14	28.60	0.71	0.71
rock	66.94	33.08	40.68	0.62	0.64

and compare the impact of each feature per genre. The most notable observations on the impact on F_1 scores are:

Blues: *Zero-crossing rate* has a positive impact. *Spectral flatness* and *chroma* features have a negative impact.

Classical: Already is very easy to distinguish, no features have a notable impact. The highest overall F_1 score of 0.92 is achieved here by *spectral centroid*.

Country: Just like for blues, *zero-crossing rate* has a very positive impact. *Spectral flatness* and *chroma* features have a negative impact.

Disco: *Root-mean-square* has the most positive impact here. *Spectral centroid*, *spectral bandwidth* and, to a lesser extent, *chroma* features have a negative impact.

Hip-hop: *Zero-crossing rate* and *root-mean-square* have a positive impact. Other features except *spectral roll-off* have a negative impact.

Jazz: All features have some positive impact, mostly *root-mean-square* and *spectral contrast*.

Metal: *Spectral roll-off* has a postive impact. Surprisingly *zero-crossing rate* has the only negative impact.

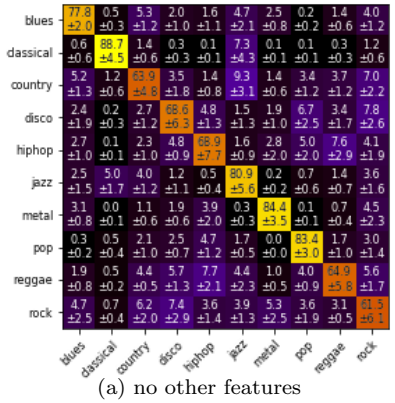
Pop: All features, except *zero-crossing-rate* and *spectral flatness*, have a positive impact, especially *root-mean-sqaure*.

Reggae: *Zero-crossing rate* and *root-mean-square* have a positive impact. *Choma* features have an exceptionally negative impact compared to the rest.

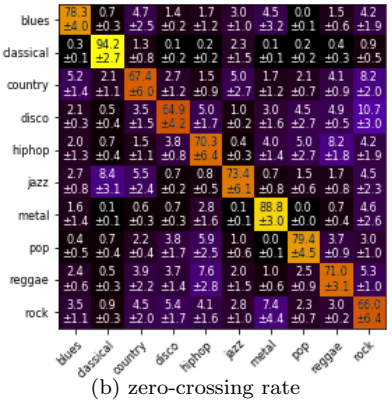
Rock: *Zero-crossing rate*, *root-mean-square* and *spectral bandwidth* have a positive impact. Again *chroma* features have the most negative impact. This time, the worst F_1 score across all the results at 0.60.

The most notable correlations between genres in figure 5j are displayed in table 7.

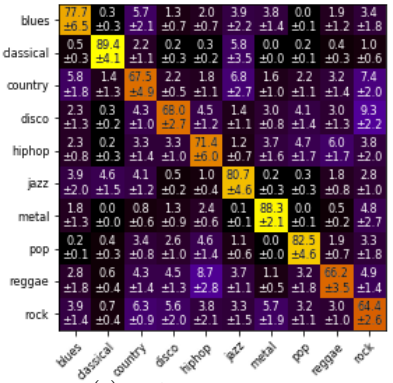
Figure 5: Confusion matrices of the average accuracies (%) and their standard deviations of neural networks with MFCCs as input along with other features. Average accuracy and standard deviation taken from the results of five neural networks trained for each variation of data input.



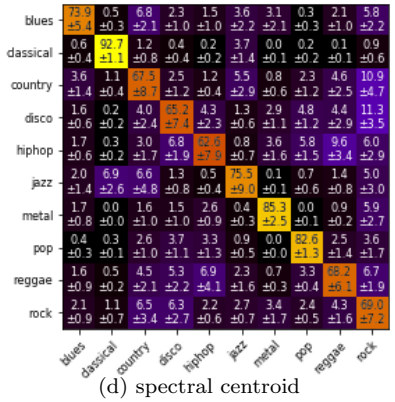
(a) no other features



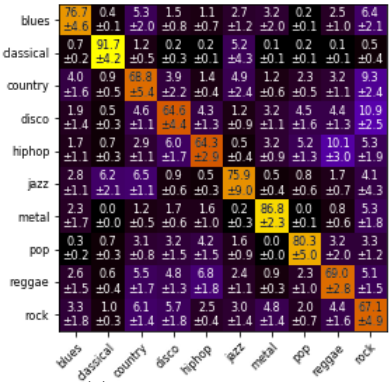
(b) zero-crossing rate



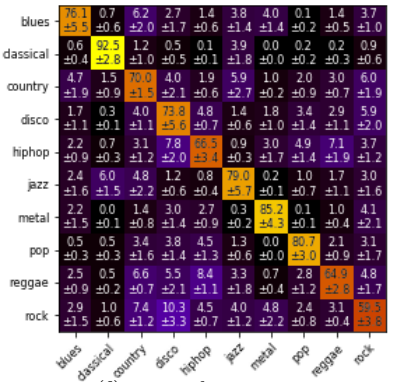
(c) root-mean-square



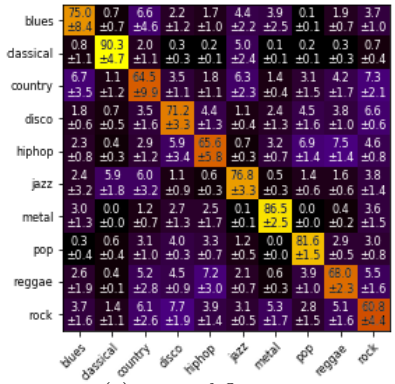
(d) spectral centroid



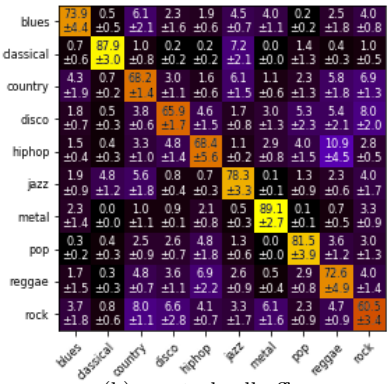
(e) spectral bandwidth



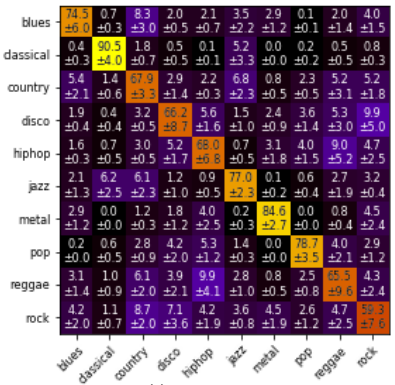
(f) spectral contrast



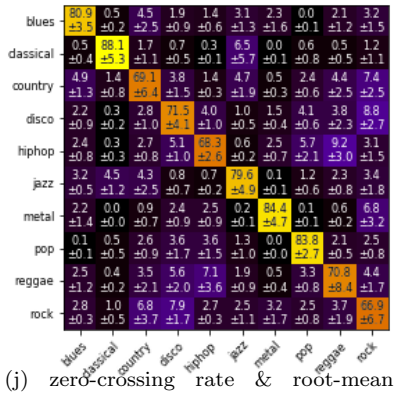
(g) spectral flatness



(h) spectral roll-off



(i) chroma



(j) zero-crossing rate & root-mean square

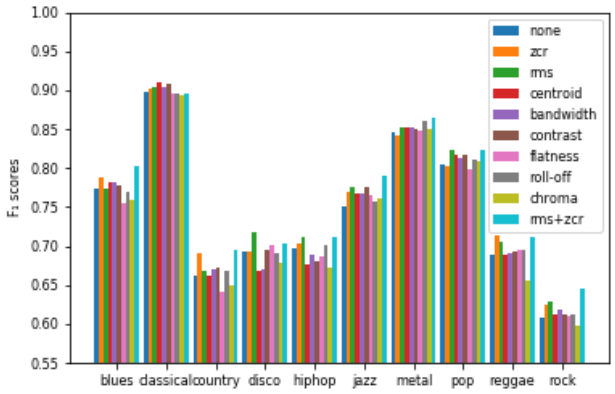


Figure 6: F_1 scores calculated for every genres from the confusion matrices from figure 5. F_1 scores and other statistics can be found in table 5, table 6 and tables 8 through 15 in the appendix.

genre	misclassified as	(%)	swapped (%)
hip-hop	reggae	9.2	7.1
disco	rock	8.8	7.9
rock	disco	7.9	8.8
country	rock	7.4	6.8
reggae	hip-hop	7.1	9.2
metal	rock	6.8	3.2
rock	country	6.6	7.4
classical	jazz	6.4	4.5
hip-hop	pop	5.7	3.6
reggae	disco	5.6	3.8
hip-hop	disco	5.1	4.0

Table 7: Highest probabilities of misclassification between genres from figure 5j.

5.5 Deeper analysis of the metal genre

To analyze the overlapping of genres deeper, four songs from different bands in the metal genre have been analyzed. The bands are as follows: *Avenged Sevenfold*, *Alter Bridge*, *Gojira* and *Opeth*. These bands all fall under the general metal genre along with subgenres like hardrock, progressive metal, death metal and post-grunge. Some songs I have deliberately chosen because they are unlike conventional metal music or because they contain long or frequent passages that sound unlike conventional metal music.

The songs are prepared to be classified by a neural network with the architecture as proposed in this paper using MFCCs, *zero-crossing rate* and *root-mean-square* as input. The songs are classified on a per second basis and the weighted average of outputs represent to final output. This is shown in figure 7.

Avenged Sevenfold - So Far Away [2010].

A 5 minute and 27 second long song from the album *Nightmare*. A mostly acoustic country-ish rock ballad with a 30 second long guitar solo starting at 2:22. At 3:23 there is an interlude followed by another guitar solo that keeps going as the outro verse, with more intense vocals, ends the song.

Scores highest, but not decisively, on country and somewhat lower on hip-hop, jazz and metal. The 15.7% metal prediction is a result of the guitar solo.

Alter Bridge - Metalingus [2004].

A 4 minute and 20 second long song from the album *One Day Remains*. High tempo heavy hardrock song with a groove feeling, and with thundering guitar riffs and drums throughout the song.

Decisively classified as country by the neural network, with a 18.9% prediction for rock.

Gojira - Born in Winter [2012].

A 3 minute and 51 second long song from the album *L'enfant Sauvage*. The song with a clean guitar riff along with accompanying bass and drums to form a low intensity atmospheric soundscape with low range vocals. At 1:50 the instrumentation and vocals get a lot more intense followed by a heavy, 16 second long, breakdown section at 2:20. Then the intense instrumentation and vocals continue until 3:06 where it continues with the same sound it had in the intro until the end of the song.

Classified as mostly jazz with also a relatively high prediction for country.

Opeth - Blackwater Park [2001].

A 12 minute and 8 second long song from the album *Blackwater Park*. A very long and straightforward death metal song with harsh vocals where riffs are repeated very often. At 2:48 there is a 2 minute and 26 seconds long interlude with clean guitar. At 7:09 there is a short guitar solo after which the drums become a lot more intensive until 10:56. Where a heavy instrumental part is played which ends abruptly at 11:19. The song finishes with an acoustic guitar outro.

Classified as both country and hip-hop, and as rock to a lesser extent.

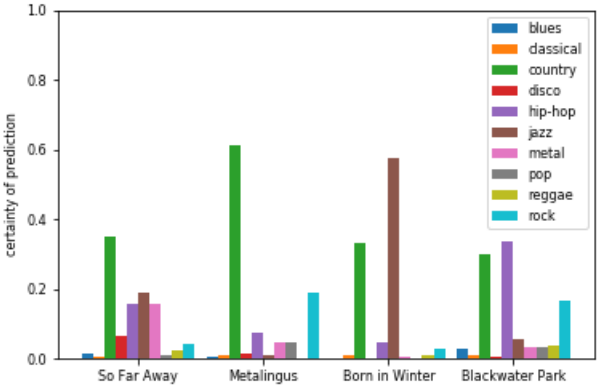


Figure 7: Music genre classification output of four metal songs.

None of the considered songs are correctly classified, with *So Far Away* having the highest prediction for the metal genre at 15.7%. This probably follows from the fact that the GTZAN dataset uses mostly heavy metal from the 80's and 90's, like Metallica and Iron Maiden, which can differ a lot from modern metal and more extreme metal genres like death metal. Considering figure 5j you can see that, if metal is wrongly classified, it will most often be misclassified as rock music. Which shows that the classes of metal and rock have some overlap. Rock music is most often wrongly classified as country or disco. There seems to be a lot of overlap between the classes of country and rock and the classes of rock and metal. This could explain why country is so dominantly present in classifying

these metal songs. Aside from that, these songs were chosen specifically because some of these songs stray from a conventional metal sound. This would explain the high prediction rate of jazz for *Born in Winter*.

6. CONCLUSIONS

This paper explored the accuracy of using MFCCs as input for a MLP neural network to classify music genres using the GTZAN dataset by classifying music on a per-second basis, resulting in a 74.37% accuracy. To further improve this, multiple musical features that can be extracted directly from the music file were used to augment the MFCC input data. Here *zero-crossing rate* and *root-mean-square* improved the accuracy the most, with accuracies of 75.32% and 75.61% respectively. Using both of these features further improved the accuracy to 76.26%, a 1.89% increase in accuracy over just using MFCCs.

For every combination of inputs for the neural network a confusion matrix was generated, which apart from the accuracy of correct classification per genre also shows what percentage of music clips was wrongly classified and as which genre it was misclassified. From this the F_1 scores for every genre are calculated which also takes into consideration other genres being misclassified as a certain genre to give an impression of to what extent you can trust the output of the neural network.

Finally, four songs from four different metal bands were analyzed using a neural network with the proposed features, (zcr and rms). Results show that the *country* genre has the highest overall prediction rate. This can be explained by the overlap of the genres of *blues* and *rock*, and *rock* and *metal*. This probably also has to do with the evolution of metal music as a genre over the past few decades.

6.1 RQ 1

Overall, the combination of *zero-crossing rate* and *root-mean-square* have the most positive effect on the accuracy of the neural network.

6.2 RQ 2

Table 7 shows the highest percentages of genres being misclassified as other genres and the percentage of misclassification in the other direction. From this table and figure 5j it can be concluded which genre pairs have the most bidirectional overlap. This research shows that the following pairs of genres are very similar based on their musical content:

- *Classical* and *jazz*
- *Country* and *rock*
- *Disco* and *rock*
- *Reggae* and *hip-hop*
- *Metal* and *rock*

7. FURTHER RESEARCH

This research can be continued by exploring other combinations and input sizes of features, and applying them to more powerful neural networks.

Interestingly, by error, I trained a network on 1 second of MFCCs and 30 seconds of *spectral contrast*, it resulted in a training accuracy of around 95% and a testing accuracy of around 90%.

This was very unexpected and warrants further research. There could be an argument to treat some features long-term, calculate them over the whole input, while treating other features on a short-term, which was the case in this paper. Furthermore, analyzing music based on their content features can be used for music recommendation systems and personalized music profiles.

8. ACKNOWLEDGEMENTS

I would like to thank my supervisor dr. Elena Mocanu who guided me throughout the research and paper writing process. I would also like to thank illustrator Myrthe Majoor for designing the visual representation of the neural network architecture.

A Jupyter notebook with code and examples can be found at: *https://github.com/DenysRF/mgr*

9. REFERENCES

[1] J. Dai, S. Liang, W. Xue, C. Ni, and W. Liu. Long short-term memory recurrent neural network based segment features for music genre classification. In *Proceedings of 2016 10th International Symposium on Chinese Spoken Language Processing, ISCSLP 2016*, 2017.

[2] A. Elbir, H. Bilal Cam, M. Emre Iyican, B. Ozturk, and N. Aydin. Music genre classification and recommendation by using machine learning techniques. In *Proceedings - 2018 Innovations in Intelligent Systems and Applications Conference, ASYU 2018*, 2018.

[3] D. Jiang, L. Lu, H. Zhang, J. Tao, and L. Cai. Music type classification by spectral contrast feature. In *Proceedings. IEEE International Conference on Multimedia and Expo*, Aug 2002.

[4] A. Klapuri and M. Dacy. Signal processing methods for music transcription, chapter 5. *Springer Science & Business Media*, 2007.

[5] G. R. Kommu, M. Trupthi, and S. Pabboju. A novel approach for multi-label classification using probabilistic classifiers. In *2014 International Conference on Advances in Engineering and Technology Research, ICAETR 2014*, 2014.

[6] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. *SIGIR Forum (ACM Special Interest Group on Information Retrieval)*, (SPEC. ISS.):282–289, 2003.

[7] T. Li and G. Tzanetakis. Factors in automatic musical genre classification of audio signals. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, volume 2003-January, pages 143–146, 2003.

[8] A. B. Mutiara, R. Refianti, and N. R. A. Mukarromah. Musical genre classification using support vector machines and audio features. *Telkomnika (Telecommunication Computing Electronics and Control)*, 14(3):1024–1034, 2016.

[9] D. Perrot. Scanning the dial: An exploration of factors in the identification of musical style. *Proc. of ICMPC 1999*, 1999.

[10] D. Shlomo. Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters*, 2004.

[11] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

APPENDIX

Table 8: Statistics from **zero-crossing rate** confusion matrix figure 5b.

genre	precision	type I	type II	recall	F_1 score
blues	78.30	21.74	20.22	0.79	0.79
classical	94.20	5.78	14.60	0.87	0.90
country	67.38	32.62	27.80	0.71	0.69
disco	64.86	35.18	22.26	0.74	0.69
hip-hop	70.32	29.68	29.54	0.70	0.70
jazz	73.42	26.62	17.60	0.81	0.77
metal	88.78	11.18	22.40	0.80	0.84
pop	79.36	20.64	18.20	0.81	0.80
reggae	71.00	28.96	28.18	0.72	0.71
rock	66.02	33.94	45.54	0.59	0.62

Table 9: Statistics from **root-mean-square** confusion matrix figure 5c.

genre	precision	type I	type II	recall	F_1 score
blues	77.74	22.32	23.38	0.77	0.77
classical	89.40	10.66	8.60	0.91	0.90
country	67.54	32.46	34.48	0.66	0.67
disco	67.96	32.14	21.68	0.76	0.72
hip-hop	71.38	28.60	29.12	0.71	0.71
jazz	80.72	19.24	27.30	0.75	0.78
metal	88.26	11.76	19.08	0.82	0.85
pop	82.46	17.58	17.98	0.82	0.82
reggae	66.22	33.80	21.70	0.75	0.70
rock	64.38	35.54	40.78	0.61	0.63

Table 10: Statistics from **spectral centroid** confusion matrix figure 5d.

genre	precision	type I	type II	recall	F_1 score
blues	73.16	26.84	15.68	0.82	0.77
classical	92.74	7.26	9.84	0.90	0.92
country	71.20	28.80	40.3	0.64	0.67
disco	66.42	33.62	31.24	0.68	0.67
hip-hop	63.52	36.54	21.72	0.75	0.69
jazz	76.74	23.30	20.94	0.79	0.78
metal	84.78	15.22	12.76	0.87	0.86
pop	82.44	17.48	18.68	0.82	0.82
reggae	68.18	31.80	28.00	0.71	0.70
rock	68.98	31.04	52.74	0.57	0.62

Table 11: Statistics from **spectral bandwidth** confusion matrix figure 5e.

genre	precision	type I	type II	recall	F_1 score
blues	78.22	21.80	20.66	0.79	0.79
classical	91.42	8.62	10.36	0.90	0.91
country	70.08	29.90	36.96	0.65	0.68
disco	67.88	32.14	31.94	0.68	0.68
hip-hop	64.34	35.58	20.82	0.76	0.69
jazz	80.02	19.98	24.08	0.77	0.78
metal	86.04	13.94	15.12	0.85	0.86
pop	79.78	20.18	15.96	0.83	0.82
reggae	69.52	30.50	29.94	0.70	0.70
rock	64.96	35.00	41.8	0.61	0.63

Table 12: Statistics from **spectral contrast** confusion matrix figure 5f.

genre	precision	type I	type II	recall	F_1 score
blues	75.40	24.68	17.80	0.81	0.78
classical	92.36	7.66	10.20	0.90	0.91
country	71.14	28.84	38.88	0.65	0.68
disco	76.38	23.62	43.26	0.64	0.70
hip-hop	66.22	33.76	26.18	0.72	0.69
jazz	77.76	22.32	21.74	0.78	0.78
metal	85.00	15.00	14.88	0.85	0.85
pop	80.16	19.84	14.78	0.84	0.82
reggae	66.42	33.56	24.28	0.73	0.70
rock	61.18	38.80	36.08	0.63	0.62

Table 13: Statistics from **spectral flatness** confusion matrix figure 5g.

genre	precision	type I	type II	recall	F_1 score
blues	72.98	27.08	17.96	0.80	0.76
classical	92.20	7.80	11.94	0.89	0.90
country	66.50	33.48	36.44	0.65	0.66
disco	73.14	26.88	34.24	0.68	0.71
hip-hop	66.76	33.22	25.08	0.73	0.70
jazz	78.38	21.66	23.52	0.77	0.78
metal	86.10	13.94	16.04	0.84	0.85
pop	81.62	18.42	21.08	0.79	0.81
reggae	69.50	30.48	29.22	0.70	0.70
rock	61.10	38.86	36.30	0.63	0.62

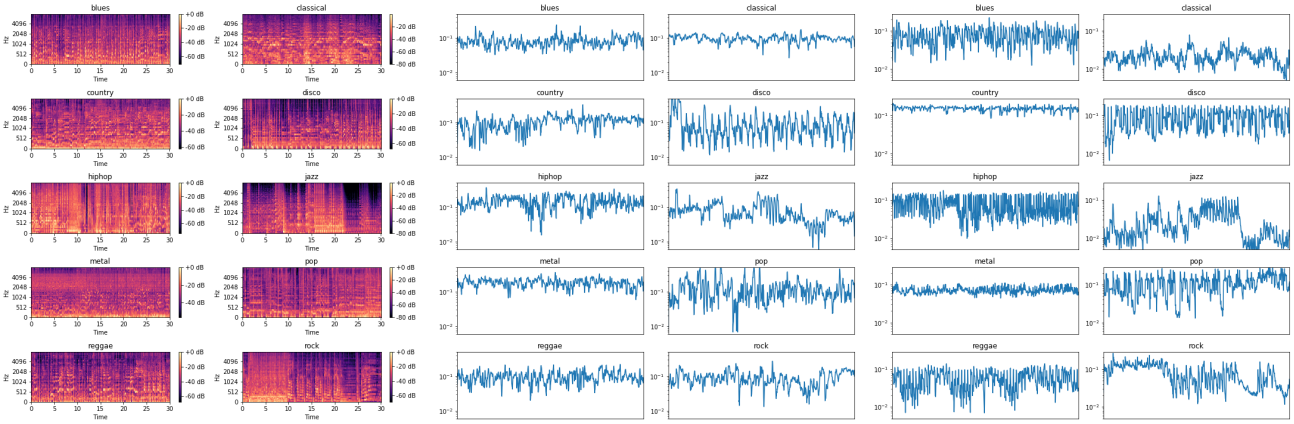
Table 14: Statistics from **spectral roll-off** confusion matrix figure 5h.

genre	precision	type I	type II	recall	F_1 score
blues	74.96	25.00	19.58	0.79	0.77
classical	88.34	11.76	8.72	0.91	0.90
country	69.26	30.74	34.94	0.66	0.68
disco	68.90	31.10	30.50	0.69	0.69
hip-hop	68.08	31.92	24.38	0.74	0.71
jazz	78.92	21.14	27.66	0.74	0.76
metal	88.08	11.90	16.72	0.84	0.86
pop	81.66	18.26	18.82	0.81	0.81
reggae	71.88	28.12	33.62	0.68	0.70
rock	60.40	39.58	34.58	0.64	0.62

Table 15: Statistics from **chroma** confusion matrix figure 5i.

genre	precision	type I	type II	recall	F_1 score
blues	75.84	24.26	21.56	0.78	0.77
classical	91.94	8.04	12.34	0.88	0.90
country	68.90	31.14	38.32	0.64	0.66
disco	69.76	30.16	33.04	0.68	0.69
hip-hop	68.14	31.84	31.10	0.69	0.68
jazz	77.54	22.54	23.02	0.77	0.77
metal	84.24	15.78	12.58	0.87	0.86
pop	79.02	20.96	15.86	0.83	0.81
reggae	68.82	31.14	36.18	0.66	0.67
rock	58.28	41.80	33.66	0.63	0.60

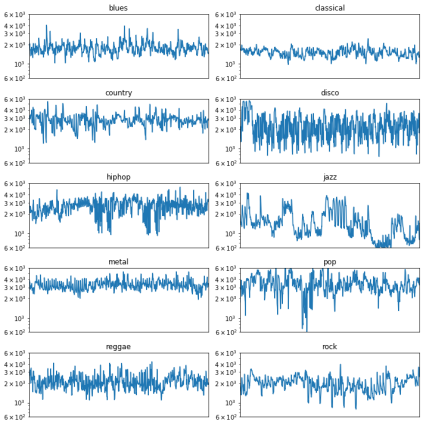
Figure 8: All features graphed for one song of each genre.



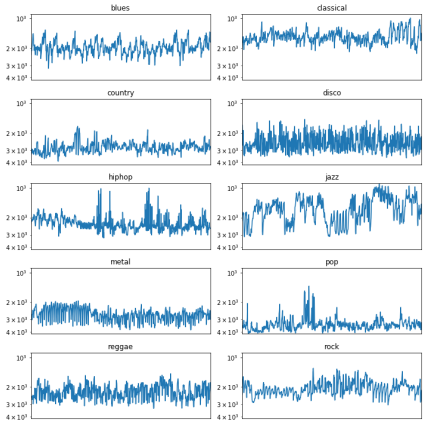
(a) mel-frequency spectogram

(b) zero-crossing rate

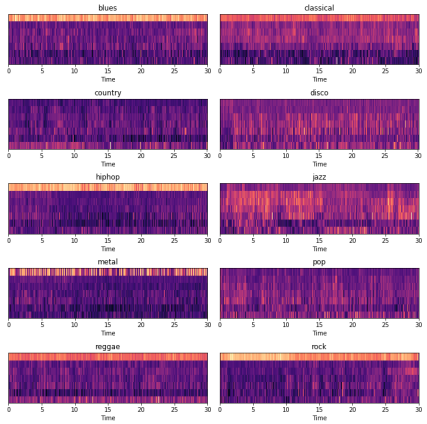
(c) root-mean-square



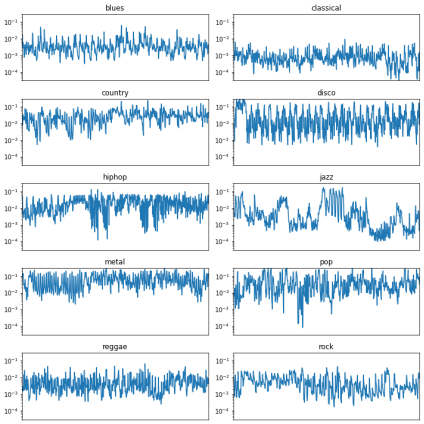
(d) spectral centroid



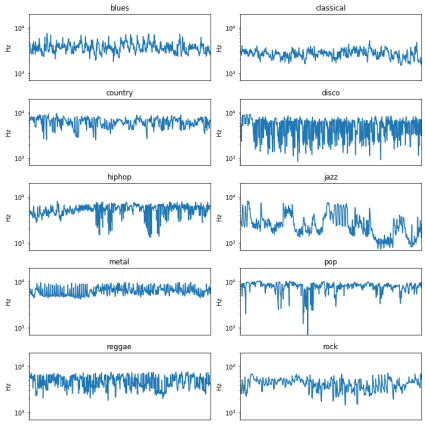
(e) spectral bandwidth



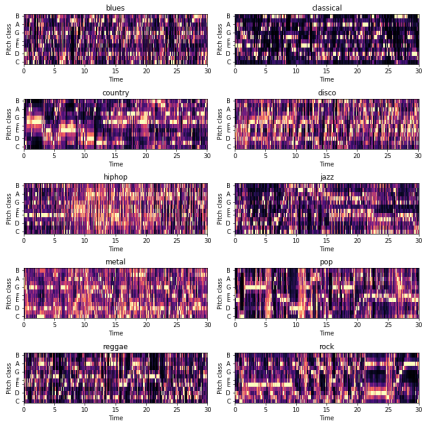
(f) spectral contrast



(g) spectral flatness



(h) spectral roll-off



(i) chromagram