slamGR

Laurens Paul Hoogenboom

April 2021

1 Human assessment of data

Even for humans it can be a difficult task to classify music genres at times. Frankly this depends on the musical talent of that person, but some songs are hard to label even for professionals. One example of a song that cannot simply be put in one genre is Red Hot Chilli Pepper's Dani California. As a musician my personal opinion is that this song could just as well be justified as Funk as Rock. Regarding solely the bass line I would not be able to tell this was a rock song, had I not heard it prior. The guitarist also uses a wahwah-pedal, which is a signature Funk sound.

For the remainder is this report only the mini set of data will be used, as the larger set resulted in a server error every time it was downloaded. In this set the first 5 songs were extracted and converted to .wav format. When I listened to these samples I came to the conclusion that the labels were already very generalised.

Track 1 is just a beyonce-like pop/R&B song. That is fair enough, as it is closest to pop out of the 4 labels. The second song was supposed to be jazz, but sounded more like a compilation of random chimes. Although jazz is infamous for stretching what degree of rythm and tonality can still be considered jazz, I would say that this is not what people generally mean when the say jazz. The 3rd track is clearly rock, unfortunately the signer is singing out of tune and the guitarist seems to strum whenever they feel like it. All in all it's a terrible performance, but clearly rock. The 4th track is labeled as rock, even though the most emphasised sound is a saxophone. In 60's rock 'n roll songs. like Johny B Goode, the saxophone was often used for solo's and overall melody. Even though modern Rock and Rock 'n roll are very different kinds of music, some elements like high BPM, the key of the music and distorted guitars are similar. the 5th track is labelled as classical. Whether it be modern classical or traditional classical music, it is hard to miss. Out of the four genres, this is the most distinct in that it only uses instrumentation from a certain era, no electronic amplification or effects are used and as a result has a way of being either calm or bombastic that is not seen in the other four labels. If there are vocals in classical music they are most commonly opera vocals, but rarely are there vocals like we are used to in modern music.

This assessment might seem overly critical and slightly snobby, but there is a good reason for that. Doing machine learning data quality is very important. The humanly detectable characteristics of music might propagate to different quantifications of the data. The sound-type of brass instruments is very specific and might be significant in the spectral data. However, since brass instruments are more common in jazz and classical music one ought to be careful when picking training data.

2 Power Spectral Density

When computing the power spectral density (PSD) of our audio signals it is important to remember that they are real signals. This means that -10Hz and 10Hz are identical signals. As a consequence the energy of the the power spectrum is evenly distributed over the negative and the positive frequencies. However in audio we don't particularly care about negative frequencies and thus this half-plane can be disregarded. In order to compensate the magnitude of every non-zero frequency should be doubled, so no energy is lost in the process.

3 (Short Term) Fourrier Transform

The short term Fourrier transform (STFT), as opposed to the discrete time transform (DFT) subdivides the signal into smaller portions, making it easier to get insight in the time evolution of the frequency spectrum of the signal. This is especially relevant in music because songs can be subdivided into part like a verse, chorus, bride, breakdown etc. Transitions between parts can happen quite abruptly and it can be useful to minimise spectral data overlap between these parts. The placement of the parts can also play a role in determine the genres, as these a lot vary. However, a certain degree of overlap will be necessary to prevent boundary artefacting.

3.1 Window size

When selecting window size for the STFT it is important to realise the window size affects both the time and frequency resolution a wider window will cause a longer duration of the track to be converted to one time instance in the STFT. A narrower window will require more steps to complete the STFT of the track. The narrower a window gets, the better it represents one moment of time in the track, this means that spectral data for that one moment of time will be better represented in the final STFT. Thus, the time resolution increases with a narrower window.

The frequency resolution decreases with a shorter window because a longer period of time allows for better distinction between similar frequencies. In audio the different frequencies are overlayed in only a couple channels (often just one.). In just half a period it would be difficult to see the sinusiodals of 1000Hz and 1001Hz diverge, but over a longer period of time, this becomes more apparent.

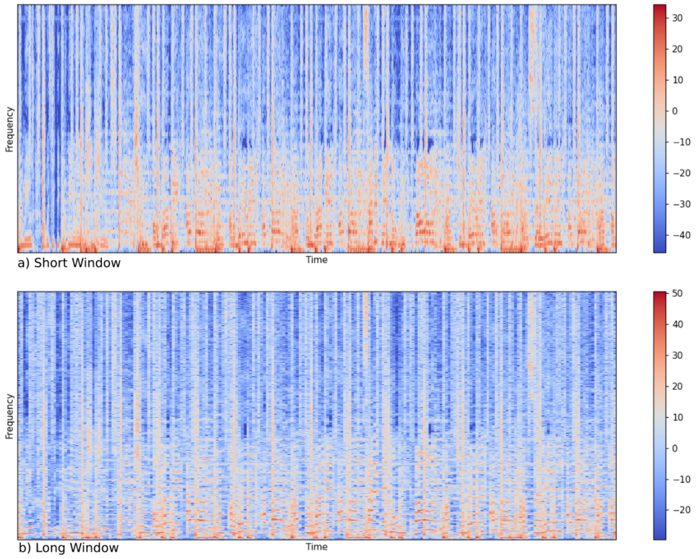

Figure 1: Above: Short window of 216 sample wide. — Below: Long window of 3898 samples wide — Both windows used 50% window width as step size.

Also, Different instruments, or different strings of the one instrument may play at frequencies very close to each other. However, these different instruments or strings will be differently dampened. This further increases the different behaviours of similar frequencies over a relatively long period of time.

Selecting a window size is a matter of making a trade-off between either resolutions and finding a workable compromise.

Rao says 40ms window 10ms step

FMA samples at 44.1kHz with windowsize=2048 and stepsize=512 samples.

Both take 4th window as step size. This equals roughly a window of 46ms.

It seems like some tens of milliseconds is a viable window order of size.

window 1949 step 512 sample rate 2205

All tracks are 420984 samples long. In order to not lose data we want the shifted window to fit an integer amount of times in the track. This way the step size can be chosen as an integer fraction of the window size and the NOLA condition will be met. In order to do this the window size will be around 80ms. The window was chosen to fit an integer amount of times in the track, such that the step size any integer fraction of the window and not break the NOLA condition.

Aditionally, one could check is the constraint overlap add condition is met. With python the *scipy.signal.check_COLA()* function can be used.

3.2 Step Size

Both RAO and FMA use a step size of approximately a fourth of the window size. This causes a 75% overlap of successive windows. From this is seems a relatively large overlap is required. The smaller the overlap the least computations required to completely cover the whole track, meaning faster performance. NOLA

Additionally, the input signal needs complete windowing coverage. i.e. The window and step size are such that every moment of time of the input signal is covered by a window at least once. This limits the choice of window and step size.

4 Spectral Leakage

non-LTI Operations on a signal may induce new frequency components. The creation of these new components is called spectral leakage. Aliasing, caused by sampling, is an example of spectral leakage. However, commonly this is used to refer to the effect of windowing.

5 Spectrogram

Depending on the type of audio considered the amplitude of the signal (a.k.a. volume) can vary significantly. The volume of a whisper, compared to a harrier jet taking off is insignificant and this can be seen on a linear scale. Or actually, this cannot be seen on a linear scale because the volume of the whisper would be indistinguishable from no sound at all if the volume of a jet plane is in the same plot.

In such a case a logarithmic scale would make more sense. A logarithmic scale can also by considered regional resolution scaling in one or multiple axis. This makes for a plot where a volume in the order of 1s can still be distinguished while a volume in the order of 1000s can also still fit in the window.

Humans perceive sound on a specific logarithmic scale called decibel.

$$dB = 20log_{10}(Vol)$$

The factor of 20 causes the Db to increase with 10dB if the volume is scaled with a factor 10. This is a convenient way of classifying ”loudness” of a sound and can thus be used to classify music. Because of the way different types of music is listened to the mixing of these songs will be different. For example, dance music has a ”kick” which is not just hear, but also felt. The kick in house song is relatively loud, while the kick of a drum in rock music is not that loud relatively speaking.

6 Time-Domain features

Perhaps the best known time-domain feature is tempo. Virtually everyone is able to clap along with a song every first beat in a measure. Especially since virtually every song is in 4/4 time signature, regardeless of the tempo. While on the topic, time signature is maybe the second-best known feature. Often people don't understand what is different in 3/4th song instead of a 4/4th song, when asked to hum a waltz most people instinctively hum a 3/4th melody, like "bum pah pah; bum pah pah".

Especially time signature would be useful, as jazz is the only of the four labels that regularly deviates from 4/4th. Rock rarely uses a 3/4th signature, like in Money by Pink Floyd. However pop is know to never do this, as part of what makes pop is that it is predictable.

7 Frequency Domain Features

at this time it seems likely that k-means will be a valid choice, because the way clustering is used works well with the way the features are implemented. However, this might be proven completely wrong.