JOURNAL OF NORTHEASTERN UNIVERSITY

Volume 25 Issue 04, 2022 ISSN: 1005-3026 https://dbdxxb.cn/ Original Research Paper

EXTRACTION OF VARIOUS HIGH- AND LOW-LEVEL FEATURES FROM MUSIC AUDIO SAMPLES USING DIGITAL SIGNAL PROCESSING, CEPSTRAL ANALYSIS AND ONSET DETECTION

Atul Kumar Rai[1], Sachi Gupta[2],Gaurav Agarwal[3]

[1]Department of Computer Science &Engineering (DS), ABES Institute of Technology, Ghaziabad

[2]* Department of Information Technology, IMS Engineering College, Ghaziabad

[3]Department of Computer Science & Engineering, KIET Group of Institutions, Ghaziabad

atulrocks@gmail.com, shaurya13@gmail.com, , gaurav13shaurya@gmail.com

Abstract

Music is a form of art that has a significant role in the life of all humans. Music helps in the development of the mind, makes the listener feel better and is even a source of energy and motivation for many. Moreover, music having certain characteristics is found more suitable to listen to than others in certain cases. Inspired by the downfall in the quality of recommendations given by various music playback platforms due to the use of a user-metadata system, the aim here has been to understand the effect of music on the human psychology and realize how music emotionally affects a person. Research has been done in Neuromusicology and light has been shed on the various kinds of effects of music and the various ways that music can affect the brain. Various kinds of features were extracted from a number of audio signals (songs) and analysed on the basis of the knowledge gained.For music is not just lyrics that we interpret or beats that we tap to, music has a great impact on our brain, perception, memory, emotions and nervous system.

Keywords: Music Information Retrieval; Human Perception; Neuromusicology

1. Introduction

Musicis a form of art comprised of sounds that are found to be pleasant to the human ear. These sounds are found to have certain characteristics (Timbral, Tonal or Rhythmic) that prove to be appealing to our minds thereby giving us a good listening experience, be it in terms of correlation or enhancement of our mood. Music can have a variety of beneficial effects on humans including improved memory and concentration, enhancement of functioning of different parts of the brain and psychological and emotional support.Moreover, the same music played at different tempos could have different responses made to that musical piece. The piece played at a faster tempo could feel happy or enticing as compared to at a slower tempo when it could feel sad or soothing. Sad or Happy responses to music could be totally non-universal1but however enticing or soothing responses are generally similar in most cases.

On an average, an American spends about 4.5 hours a daylistening to music2 using music playback platforms like Google Play Music, Spotify, YouTube, etc. to stream music on Smartphones or personal computers. Recommendations made by such platforms often include already streamed content and content which has been created by the same artist or belongs to the same genre which limits the content that reaches the user and restricts the user's

Submitted: 21/10/2022 **Accepted**: 24/11/2022

experience.Another problem with these recommender systems is the cold – start problem which means that new artists/ new singles are not recommended to the user even though they might match the user's choice or listening preferences. This is due to the recommender system being dependent on user-metadata i.e. the user's previously viewed content resulting in recommendations of music from the same artist or belonging to the same genre which is a very narrow playlist and neither of these two guarantees suitability with respect to the listener's mood.

The goal of this study is to understand the effect that music has on the human mind and how the same is produced, extract various features from music samples and analyse them on the basis of Neuromusicology theory correlating them to the perceptions of humans.Neuromusicology could help music platforms provide a better experience to users while at the same time helping the users benefit from better music recommendations to suit their listening preferences and help them feel more satisfied.

The study involves the use of Python (v 2.7) programming language to analyse, clean and model the data. It involves the use ofAubio, Mido and Music21 toolkits in Python to interpret the audio signal and analyse its features. It also involves the use of Scipy, Statsmodels, Numpy and Matplotlib toolkits/modules for mathematical and graphical modelling of the extracted feature data set from the audio samples which is stored in .csv (Comma Separated Values) format.

How music affects the human mind

Different genres of music have different effects on autonomic nervous system activity, and such effects cannot simply be explained by musical preferences of the individual. There is more to it than meets the eye. - Andrew Manson, Eastern Kentucky University.

Ocean waves seem to be pleasant and soothing because they have a sound cycle of approximately 12 cycles per minute which is the breathing rate of an average sleeping human, hence the connection in our brain. Also, it is associated with a holiday. We can't process different kinds of sounds resulting in noise and our productivity is inhibited and lowered down to up to 33%. – Julian Treasure, TED

All this happens due to release of Dopamine (the Love drug) in the body. While listening to music we are analysing the patterns and based on what musical knowledge and pattern recognition we have developed we predict what will come up next. Now the surprise lies in whether we actually get to hear what we are expecting or we don't. These little nuances give us the pleasure while hearing music. – Valorie Simpoor, McGill University

The reasons for the effect of music on the human brain could be one or more from among the following: -

- Associated Learning – General perceptions based how one's mind has been shaped since childhood 3.
- Musical Expectations – patterns are easier to remember and anything with a pattern sparks the mind more. Music has rhythm which makes it predictable. It gives us a definite structure to remember and associate the piece with and similarly becomes easier to recall at a

later time. This is also the reason music can divert our mind very easily. But this fact also helps us make associations and make us remember certain things (text, methods of doing things, happenings) with the music that is associated with the same and forever reminds us of the same (content or event in life) 4.

- Expressive Emotional Movement – fast/active movements, higher tempo rhythms/accelerating tempos and high-pitched voice tone (especially increasing pitch) is found to increase tension 5.
- Activating Sound – Sudden changes in the acoustic environment enhance perceptual processing of subsequent visual stimuli that appear in close spatial proximity 6.

The first two are based on learning and development while the second two are more universal music feature-based explanations.

Neuromusicology7is the study of the relationship existing between the human nervous system and the interaction of the human brain with music (Roehmann, 1991). The sounds we usually hear such as musical tones proceed within our body along a marked path,entering as plain sound waves into the inner ear or cochlea whose function is to sort complex sounds into their elementary frequencies. The cochlea then transmits these elementary frequencies in the form of trains of neural discharges via separately tuned fibres of the auditory nerve to the auditory cortex, which is present in the temporal lobe of the brain,where specialized cells respond to certain frequencies. Overlapping tuning curves of neighbouring cells prevent gaps in the system. However, the way that the human brain responds to music is not so simple. Sequences of tones are grouped together by the brain and relationships identified instead of each tone being interpreted individually8.Entering of music into our brains results in the triggering of pleasure centres that release dopamine, a neurotransmitter responsible for making us feel happy. In certain cases where the brain is familiar with the music, it can even anticipate the most pleasurable peaks and prime itself with an early dopamine rush.

Julian Treasure (The Sound Agency) describes 4 ways that sounds affect us 9. The former also states that our relation with sound has become more or less unconscious. We deal with sound involuntarily suppressing sounds that are unpleasant and random and highlighting others that sound pleasant and melodious. The following are the 4 ways: -

- Physiological - Ocean waves seem to be pleasant and soothing because they have a sound cycle of approximately 12 cycles per minute which is the breathing rate of an average sleeping human, hence the connection in our brain. Also, they are associated with a holiday at the beach.
- Psychological - Sounds of birds chirping gives us the feeling of happiness and hence are often included in movie scores to add meaning and feeling to a movie scene.
- Cognitive – The way that sound affects our productivity. Multiple sounds resulting in noise are difficult for the human brain to process simultaneously and our productivity is inhibited and lowered down to up to 33% 10.
- Behavioural – A natural human tendency is to distance oneself from unpleasant sounds while moving closer to pleasant sounds. Different kinds of music affect our mood differently

and can make us do the same things differently. For example,the way we drive is influenced by the songs we are listening to while driving and the music we generally listen to.

Further, there have been many studies that validate various different hypotheses regarding the effect of music on the human brain. One such study (Mei-Ching, Pei-Luen, Yu-Ting &Keh-Chung, 2013)11that analysed stroke Patients with Unilateral Neglect observed that the patients performed better on tasks requiring them to cut straight lines in half or to recognize and record objects in simulated real-life scenes while they were listening to classical music as compared to their performance while they were made to listen to white noise or nothing (silence).The highest performance scores were thus observed amongst the participants under the classical music condition while the lowest scores were observed under the silence condition. Also, most participants after listening to classical music rated their arousal as highest.

Listening to music helps exercise better and longer which is because it can drown out our brain's cries of fatigue 12. As our body starts to get tired and wishes to stop exercising, signals are sent to the brain asking for a break. These signals are competed against by music signals thereby helping us to override those signals of fatigue.However, this is usually beneficial for low and moderate-intensity exercise cases only (Table. 1). Listening to music while exercising allows us to push through the pain to exercise longer and harder andalso helps us to use our energy more efficiently. A study (Bacon, Myers, & Karageorghis, 2012) 13 has also showed that 7% less oxygen was consumed by cyclists who listened to music while doing the same work (cycling) as compared to thecyclists who cycled in silence. However, it was also observed that there was a ceiling effect ofthe music at around 145 beats per minute. Music with tempo higher than this didn't add much motivation.

We can even be affected by short pieces of happy or sad music. A study (Logeswaran& Bhattacharya, 2009)14 showed that participants interpreted a neutral facial expression as happy or sad after hearing a short piece of musicto match the tone of the music they heard. This effect was most notable with facial expressions close to neutral but was also observed with others. The tempo of music is also found to have an effect on the heart rate, blood pressure and respiratory rate. Experiments (Bora, Krishna &Phukan, 2017) 15 involving 70 individuals were conducted making the people listen to two different kinds of music, fast tempo (120-130 beats per minute) and slow tempo (50-60 beats per minute) and observing the variations caused in respiratory rate, heart rate and blood pressure. The variations caused in blood pressure are notable but those in breathing and heart rate are not very significant.

Music Information Retrieval (MIR): -

Music information retrieval (MIR) is the retrieval of information from music and involving many real-world applications. MIR is not a completely separate science but includes knowledge of psychology, musicology, signal processing, psychoacoustics, academic music study, optical music recognition, informatics, machine learning, and computational intelligence and often a combination of some of these. Techniques for MIR include Spectral Analysis (using Fourier Transform to form Spectrum), Linear Predictive Coding (LPC), Cepstral Analysis (extraction of MFCCs) and Zero-Crossing Rate (ZCR). Extraction of lyrics is done using

Natural Language Processing (NLP). Features extracted from music could be of the following kinds 16& 17: -

- Low Level Features of audio – are those features which have little or no significance to users but are used to calculate the high-level features that will actually be understandable to the user. These low-level features usually talk about the timbre and loudness of the music which in turn are related to the more user understandable features. For example, Spectral Envelope Shape, Temporal Evolution, time variance, etc. have little or no meaning to the average music listener. However, these shall be used to determine higher level (rhythmic) features which are user understandable.
- High Level Features of audio – are more user understandable features in contrast to the low-Level Features and are usually derived from the low-level features after analysis. They are usually related to the characteristics of music that listeners speak about. For example, the tempo of the music piece is a high-level audio feature.

Another classification of music audio features could be presented as follows -

- Timbral Features - When a musical piece has frequencies which are multiples of the frequency of the central pitch of the music, the musical piece is said to have harmony (2f, 3f, 4f, etc. along with f). This is known as timbre of the music. However, in case of a noise signal, there are many frequencies in the frequency spectrum of the sound which are not multiples of the central frequency or rather there is no central frequency in such a case. There is no timbre in noise.While contributions of features such as harmony, loudness, melodic expectation, pitch height, and tempo have been examined thoroughly in past studies, it is still unclear how (and which) Timbral features contribute most directly to musical tension. The Timbral features that have been examined in past work (from widely varying methodological approaches) include roughness, brightness, spectral flatness, and density 18.Mel Frequency Cepstral Coefficients (MFCCs) are Low-Level audio featuresrepresenting the timbre of the audio. The values of the MFCCs on the Mel-scale depict an approximation of the frequency resolution of the auditory system by mapping the actual frequency to perceived pitch.
- Rhythmic features –rhythm description aims at turning acoustic events occurring in time into more abstract notions of timing, tempo and metrical structure. In metrical structure, tempo is the rate of beats at a given metrical level. The metrical structure also describes the beat pattern or the regular temporal structure that the music piece follows. Rhythmic structures are identified based on the measurement of periodicity of events which are represented by onsets in the music piece which can be identified by pitch/energy changes (low-level features). These onsets help estimate beat-positions in the music as well as the loudness of the beat which helps to identify the start and end of the beat cycle.
- Tonal features –areone of the main aspects of western music. The key of a song is a collection of pitches having a certain central pitch or tonic and two other pitches which are the dominant pitch(5th degree of the scale – lying above the tonic) and the sub-dominant (4th degree of the scale – lying below the tonic), also known sometimes as the chord to the pitch.There are two basic tonal key modes, themajorand minor mode, differing in their characteristics and in terms of the positioning of tones and semitones within their perspective

scales. Scales are composed of sequence of notes and each pair of consecutive notes forms an interval that is defined by a ratio (whole-tone or semitone).

Spectral Analysis - Music Information Retrieval techniques include Spectral Analysiswhich works on the fact that the audio signal at any point of time may be composed of various frequencies (Sine waves). The combination of various such frequencies may lead to a particular waveform. Spectral analysis is concerned with these constituent frequencies and not the waveform itself. Spectral Analysis decomposes the signal snippet into its constituent sine wave frequencies by making use of fast Fourier Transform (FFT). These sine waves when combined would result in generation of the signal snippet being analysed.The FFT results can then be used in generation of the Frequency Spectrum of the signal given which is plotted as amplitudevs. frequency graph which depicts the various frequencies that are present in the signal along with their amplitudes. However, this does not mean that it is necessary to combine all the sine waves to get the signal; it is just a way of visualizing and analysing the signal.The frequency spectrum only depicts the signal at a particular point of time i.e. for a given fixed time window. A spectrogram is used to generate a time varying view of the signal with the plot being frequency vs. time and the graph being colour coded to represent magnitude of a particular frequency component.

Cepstral Analysis19 – isanother technique that is made use of herefor information retrieval from the provided music audio samples. A Cepstrum is obtained by taking the logarithm of the estimated spectrum of a signal and then performing Inverse Fourier Transform (IFT) and can be seen as information about the rate of change in the different spectrum bands.Reversing the first four letters of the word "spectrum" led to the derivation of the word “cepstrum” with operations on cepstra being quefrency (amalgamating the first five letters of “frequency”) analysis, liftering (reversing the first three letters of “filtering”), or Cepstral analysis. Cepstra can be of different types includingreal cepstra, complex cepstra, phase cepstra, and power cepstra. The power cepstrum in used in the analysis of human speech and often as a feature vector for representing the human voice and musical signals. For these applications, the power spectrum is usually first transformed using the Mel scale resulting in the generation of the mel-frequency cepstrum or MFC with its coefficientsknown to us as the Mel-Frequency Cepstral Coefficients or MFCCs. The mel-frequency cepstrum finds applications in pitch detection, voice identification, etc.Cepstral Analysis refers analysis done on the cepstrum obtained to derive certain features of the signal.

Mel Scale 20 The Mel Scale is a perceptual scale of pitches judged by listeners to be at equal distances from each other i.e. equal increments of pitch are perceived by the listeners from one frequency band to the next. This is due to the human ear acting as a filter with there being a greater number of filters (closely spaced) in the lower frequency range while the hearing filters are spaced widely in the higher frequency range. This means that the human ear perceives the same pitch increment over a greater frequency difference at higher frequencies as compared to the lower frequencies. This causes a non-linearity in the Frequency scale while the Mel Scale overcomes this non-linearity. The 1000Hz pint on the frequency scale acts as the reference point corresponding to 1000 Mels. This as a Psychoacoustic Scale is shown in figure 1.

$$m = 2595*\log_{10} (1 + f/700)$$

$$f = 700*(10^{m/2595} – 1)$$

Conversion from Mels to Hertz and vice – versa.

m = pitch/ melody in Mels, f = frequency in Hertz

The Mel-Frequency Cepstrum (MFC) in sound processing is regarded as a representation of the short-term power spectrum of a sound. The cepstrumis obtained on a non-linear Mel scale of frequency from a linear cosine transform of a log power spectrum.The linear predictive cepstrum differs from the mel-frequency cepstrum21 in the way that the MFC has equally spaced frequency bands on the Mel scale, which approximates the human auditory system's response more closely. This human auditory adaptive representation allows for better representation of sound, for example, in audio compression. Mel-frequency Cepstral Coefficients (MFCCs) derived from the Cepstral representation of the audio clip collectively represent a Mel-Frequency Cepstrum.

The Mel scale takes the human hearing perception into account and Mel Frequency analysis therefore includes closely spaced filters at the lower frequency range and not so much at higher frequencies. These closely spaced filters help to accommodate for the better differentiating capability of the human ear at lower frequencies while not so much at lower frequencies. These filters are triangular in shape and known as Mel Frequency Filter Banks as shown in figure 2. The output from each filter bank constitutes one MFCC and represents the energy present in the range of that filter bank.

2. Materials and Method

Data

The methods discussed below were applied to standard MIR datasets as well as personal music collections after standardizing both to the same format and sampling rates to avoid any ambiguities due to different audio qualities of the files tested. Standard data sets used include the freely available Hainsworth dataset [A] which consists of 245 song samples of the English language from a variety of genres, some of which are considered easy to track (pop/rock) while others more difficult (classical/jazz). All the excerpts in the Hainsworth dataset were not of the same duration (range: 40s – 90s), however, all were made to conform to the 705kbps bitrate standard. Besides, samples from the personal music collection (i.e. author’s music collection) were analyzed which were again made to conform to the 705kbps standard. This set included songs in the English and Hindi languages encompassing various genres, ranging from songs of the 1990s to the first half of the 2010s. The chosen music samples were in the .wav (Waveform Audio File Format) which is sampled at 44100Hz with a single channel. This sampling rate was chosen because the Nyquist Theorem 22 states the requirement of a sampling frequency at least twice the highest frequency component of the signal which is 20,000Hz i.e. the maximum frequency audible to humans.

Audio File Inclusion

Only a 45 second (at max.) sample of the music of was analyzed constituting a total of 45*44100 = 19,84,500 samples. For the Hainsworth dataset, the first 45 seconds of the excerpt were chosen except in cases where the duration of the sample was less than 45 seconds itself and only as much samples could be formed. For music from the personal collection, a 45 second excerpt of the duration 00:01:15 to 00:02:00 is chosen.The next task performed was filtering of the music sample. An A-weighted 23 filter was applied onto the music signal to account for the human auditory system which is more sensitive to differences at lower frequencies and lesser to higher frequencies.

This is followed by the task of determination of a window-size and accordingly, a hop-size (both in terms of number of samples). The window-size is the number of samples considered at once from the input sample to extract the required features. This size is used by the Fast Fourier Transform (FFT) function and for onset and tempo detection. The hop-size is the number of samples taken into consideration foreach iteration of the feature extraction algorithm. It wasmade sure that the window-size was not an integral multiple of the hop-size so that exactly 'n' number of hops did not fit in a single window 24. This was to make sure that there was overlap at the boundary of the window with a single-hop being a part of two consecutive windows which helped avoid window-end inconsistencies due to cross-window validation.

Multiple acoustic features that can be extracted from the audio signal are usually calculated by first creating overlapping frames of the audio signal with overlap up to 50%. This frame splitting is done using windowing functions like the Hann function and then transformed using the FFT function of the computer software (like MATLAB) which returns us the Discrete Fourier Transform the audio signal window snippet. This is known as the Short-Time Fourier Transform (STFT) of the audio signal.

Feature Extraction

Once the audio file has been suitably included, the task is to extract the various kinds of features from the selected 45 second sample (00:01:15 – 00:02:00). These features include the tempo (beats per minute) (Rhythmic, high-level feature), key and mode of the song (Key – Pitch/Note, Mode – minor/major) (Tonal, high-level feature) and the Mel Frequency Cepstral Coefficients (MFCCs) (Timbral, low-level features). Feature extraction makes use of Python programming language and associated toolkits/modules.

1. Tempo (Beats per Minute) –

$$\text{Beats per Minute (BPM)} = \frac{\text{Total no. of beats observed with certain confidence}}{\text{Total time (in minutes) of sample}}$$

The extraction of the tempo of a music sample from the audio input requires the estimation of the onset times which indicate the fall of the next beat (lines of lyric usually start with the onset of the beat). The onsets/beats found in each frame are recorded and a collection of all onsets over all frames of the audio sample is made which gives us the total number of beat occurrences over a frame. Dividing this by the frame size gives us the number of beats per sample of the audio which when multiplied by the sampling rate of the audio signal results in the number of beats per second of the audio. This is the intended tempo (number of beat occurrences per minute is referred to as BPM – beats per minute). The tempos obtained for each frame are then averaged over the entire sample to obtain the average music tempo.

Another method of extracting the tempo for a music sample would be to calculate the total number of onsets/beat occurrences till a point of time in the sample and divide the same by the time of the sample that has been scanned so far. This gives us an instantaneous tempo at each frame/onset which again, when averaged over the entire sample results the average music tempo.

2. Key (Pitch and Scale)-

Extraction of pitch is done on the basis of the analysis of the energy present in each frequency range. Each musical note corresponds to a certain frequency. An energy spike at the corresponding frequency indicates the presence of that note in the musical piece and this presence is recorded. A collection of all such note recordings over two octaves (4th- 6th Octave) has been made and fed to a MIDI stream in the form of a MIDI Object 30 (Pitch, Onset, Start, Velocity and Last). The MIDI Object stream is put under comparison with the known scales to find the scale which closest resembles the notes obtained from the music piece.

The key of a song is comprised of the pitch, which is the root note of the song and the mode which is the scale/chord associated with the root node defined as major chord or minor chord. A musical scale 28 is constituted by 7 musical notes with the first note being the root note of the song followed by increments of a Whole Tone, twice, a Semi Tone, once, Whole Tones again, thrice and finally another Semi Tone (for theoretical knowledge and background on musical scales and notes, one may refer to documents on the World Wide Web). The root note of a scale i.e. the key of the song is the key that the song begins in. A musical chord 29 is comprised of 3 notes i.e. a triad. A musical triad consists of a root or first note, the third note, and the fifth note. It is possible to form a triad on any note of a major scale such as the C major scale, for example, which consists of the notes C D E F G A B. All triads are minor or major except the triad formed on the seventh or leading-tone as the root note, which is a diminished chord. A triad formed using the note C itself would consist of the first note of the C major scale (C itself), the third note (E) and then the fifth note (G) of the scale. The interval from C to E is a major third i.e. of four semitones, hence this triad being called C major (major chord). In contrast, the triad formed upon the same C major scale but with D as the root note would have only 3 semitones between the first (D) and third note (F) and would thus be called D minor (minor chord).

3. Mel Frequency Cepstral Coefficients (MFCCs) –

The sounds produced by humans are functions of their vocal tract which includes the vocal cords, tongue and teeth as well. The shape of the vocal tract of the human that leads to the production of any particular sound gets manifested into the shape of the envelope of the frequency spectrum and this is what the MFCCs are supposed to represent. They are an improvement over the LPCs (Linear Prediction Coefficients) and LPCCs (Linear Prediction

Cepstral Coefficients). The following are the steps followed to calculate the MFCCs from a given audio signal -

i. Frame the signal into short frames – The audio signal is continuously changing hence there is a need to break it down into smaller parts so that it can be analysed one part at a time. The frames are kept very small in size so that the signal snippet can be approximated as constant. The size of the frae is same as the hop size. An excessively large frame yields only an estimate/average of the signal (and its features) over time and a frame too small makes it unable to generate enough samples to extract features from. Next, we obtain the power spectrum of the signal to identify the frequencies present in that particular time frame of the signal. This is inspired from the cochlea which a part of the human auditory system which vibrates at different parts corresponding to different frequencies. These frequencies (sine waves) combined together could regenerate the original signal however it does not mean that this is how the signal was indeed created originally.

ii. For each frame calculate the estimate over the periodogram of the power spectrum – The human cochlea cannot distinguish clearly between two close frequencies. This ability to distinguish diminishes further as we move towards higher frequencies (hence, the Mel filter banks are closely spaced at lower frequencies and move further apart at higher frequencies). The Mel filter banks get wider at higher frequencies as we are less concerned about variations and only about how much energy is present at each spot. For each frame, we calculate the energy stored in the samples and the energies in the various frequency ranges.

iii. Apply the mel filter bank to the power spectra, sum the energy in each filter – to calculate the energies, we multiply the power spectrum with each filter bank and then add up the coefficients (which would be 0 for all positions which lie outside a particular bank and coefficients corresponding to particular filter bank are generated by the product of the frequency amplitude and the filter height). This results in a summation of energies in a given filter bank with an added correlation with the human auditory perception differences.

Choosing and generating the filter banks25 – the upper frequency for the filter bank range is limited to half the sampling rate (Nyquist theorem). The filter banks should be equally spaced on the Mel scale with equal mel-widths so we convert the frequency range to Mels and then linearly split the range into as many filter banks as we want. Then we convert back the filter bank ranges from Mel to frequency (inverse log, they are now no longer linear). We then have to round off these frequencies to the nearest FFT bins which give us a certain number of points on the FFT scale (power spectrum x-axis). A filter bank begins at the first point, peaks at the second and ends at the third (symmetrical). The second bank would begin at the second point (peak of the first), peak at the third point (end of the 1st bank) and end at the third point (start of the third filter bank). This results in a 50% overlap of consecutive filter banks.

iv. Take the logarithm of all filter bank energies – we don’t hear loudness on a linear scale. Generally, to double the loudness of the sound we hear we need to increase the energy of the sound signal to 8 times. Hence, the variation is more difficult to perceive at higher frequencies (energy is directly proportional to frequency). This means that the energy vs. loudness graph is rather exponential than linear/polynomial.

v. Perform Discrete Cosine Transform (DCT) on the log filter bank energies – DCT de-correlates the coefficients of consecutive filter banks and makes them more independent of each other which is not the case originally due to the complex nature and layering of most songs. DCT also limits the number of features to enhance the performance.

vi. Keep DCT coefficients 2-13, discard the rest 26 – These are the finally obtained 12 MFCCs. Filter banks are spaced out (non-linearly) over the range of 0 – 20,000Hz. However, in practicality, the human ear is most sensitive to the frequency range 2000 – 5000Hz 27 while most songs have their frequency components in the range of about 50 – 8000Hz (Octaves 2 through 8). The 1st filter bank centres on the frequency of 0Hz which is inaudible and hence of no significance. Hence, the 1st filter bank is discarded. Similarly, 14 onwards filter banks are also representatives of frequencies beyond the practical hearing range and the usual song component frequencies, thereby eliminating the need for their inclusion and consideration. Hence, these are also discarded.

3. Results

A set of 15 features were successfully extracted from the music samples using various Digital Signal Processing and Power Spectrum analysis methods while taking into consideration various factors including sampling rate, size of frames and umber of bins to be chosen for the Fast Fourier Transform function. These signal processing methods were also coupled with digital musicology methods (MIDI streams) to analyze discrete signal samples for their features. The set of 15 features included MFCC coefficients 1 through 12, the Tempo (Beats per Minute), Pitch (Key of the song) and Mode (Major and Minor for the scale). Features 1 through 12 (the MFCCs) were successfully extracted using the 6 steps mentioned under Feature Extraction -> Mel Frequency Cepstral Coefficients (1st sub-heading). The Tempo was successfully extracted using the method mentioned under Feature Extraction -> Tempo (2nd sub-heading). The Pitch and Mode were collectively extracted successfully using the method mentioned under Feature Extraction -> Pitch (3rd sub-heading).

A sample set of features as extracted for 4 different songs from the personal music collection (‘Tera ChehraKitnaSuhanaLagta Hai’ by Jagjit Singh, ‘Agar Tum Saath Ho’ by Alka Yagnik and Arijit Singh, ‘Humble’ by Kendrick Lamar and ‘Wish You Were Here’ by Pink Floyd) has been shown below in Table 2. The first song mentioned above is widely regarded as a Ghazal, the second as a Bollywood song, the third as a Hip-Hop/Rap number and the fourth as an all-time great Classic Rock track. The first sample taken was from the song ‘Tera ChehraKitnaSuhanaLagta Hai’ by the renowned singer Jagjit Singh and is one of his best known Ghazals. The sample was found to have a tempo of approximately 101 Beats per Minute with the actual31 known tempo of the song also known to be 100 Beats per Minute. The Pitch determined using the algorithm discussed under Feature Extraction -> Pitch was C♯ in the Minor Scale with the actual pitch known to be E♮-Major which is known to be the same as C♯-Minor since both the former and latter are relative scales 32 33 of each other (relative major and relative minor respectively).

The second sample taken was from the song ‘Agar Tum Saath Ho’ sung by Alka Yagnik and Arijit Singh and was observed to have a tempo of approximately 124 Beats per Minute with the actual34tempo known to be 125 Beats per Minute. The pitch obtained using the algorithm was E♭ in the Major Scale (i.e. D♯ Major) with the actual pitchknown to be D♯ Major as well.
The third sample taken was from the song ‘Humble’ by Kendrick Lamar which was observed to have a tempo of approximately 152 Beats per Minute and the pitch A♮ in the Minor Scale with the actual35 tempo and pitch known to be 150 Beats per Minute and C♯ in the Minor Scale (relative Minor to E-Major) respectively. In this specific case, the extracted and actual known song keys do not match.
The fourth sample mentioned in the table was taken from the song ‘Wish You Were Here’ by the band Pink Floyd and was observed to have a tempo of about 121 Beats per Minute and pitch G in the major scale. The actual tempo and ¬key 36 of the song are known to be 123 Beats per Minute and G Major respectively and hence both are good matches.
The error in the tempo extraction algorithm (considering the above three cases) was found to be: -

$$\frac{1}{4} \times \sqrt[2]{\left(\frac{101-100.74}{101}\right)^2+\left(\frac{125-124.48}{125}\right)^2+\left(\frac{150-151.98}{150}\right)^2+\left(\frac{123-120.99}{123}\right)^2}$$

$$= 0.0054\ (0.54\%)$$

Similar analysis was done on the samples of the Hainsworth music data set and the results recorded. The 12 MFCCs, tempo (bpm) and key (pitch and mode) of each excerpt, as obtained were stored in a Comma Separated Values (.csv) file. The tempos obtained for the excerpts were compared against their corresponding known tempos obtained from the online source and the error calculated as –

$$\text{Error} = \frac{\text{absolute value of (obtained bpm} - \text{ known excerpt bpm from online source)}}{\text{known excerpt bpm from online source}}$$

The accuracy of the method discussed above on the Hainsworth dataset was found to be 99.16%.
However, in some cases it was also observed that the samples read were not enough to obtain sufficient data from to calculate the values of the audio features which resulted in NULL (‘nan’) values being returned by the algorithm. This indicated the possibility of using a larger frame size but such cases were onlyobserved for a few frames among thousands of total frames. Hence, the current method of feature extraction seemed suitable for a majority of the cases. Another possible reason for the NULL output values could have been the inability of the algorithm/method invoked to detect enough instances of the feature (Beat Onsets, Key Onsets or Energy Spikes in the inverse spectrogram envelope).

4. Discussion

The various samples from the standard Hainsworth dataset and from the personal collection analysed led to collection of 15 features each in most cases except where the onsets detected were not enough to estimate the tempo of the piece. The song samples mentioned above were observed to have different tempos, pitches, and MFCC features, which were all recorded in a CSV file. The first 12 features from the feature set of each song i.e. the MFCCs are not directly interpretable as they are low level features. The tempo and key of the song are rhythmic and tonal features which are human-understandable.

The song sample with the highest tempo among the shown three was from the song 'Humble' by Kendrick Lamar which is known to be of the Hip-Hop and Rap genre. The song was acclaimed as a very energetic and powerful one with its heavy beats and background ever since its release in 2017.Such songs, as discussed above (refer to Table 1) are suitable for workouts and exercise. The song sample with the slowest tempo observed among the shown 3 samples was from the song 'Tera ChehraKitnaSuhanaLagta Hai' by Jagjit Singh which is one of his best-known works as a Ghazal singer. The song is widely regarded as a relaxing song with a romantic tone.

'Wish You Were Here' was found to have the highest song key (G Major) among all three song samples. No comments for certain could be made on the values of the 12 Mel Frequency Cepstral Coefficients since they are low-level features and must be converted into high-level features to make them human interpretable. Three of the four song samples ('Humble', 'Wish You Were Here' and 'Agar Tum Saath Ho') were observed to have a Major scale while the fourth ('Tera ChehraKitnaSuhanaLagta Hai') was found to have a Minor scale.

The algorithm discussed above successfully extracted various features from song samples of the .wav format sampled at 44000Hz with a single channel. Tempo was extracted with 99.46% accuracy while Pitch was extracted with 75.00% (three out of four correct) accuracy as far as the 4 samples mentioned in the table are concerned. The extracted features supported the expectations developed on the basis of various theories and previous works. 12 Mel Frequency Cepstral Coefficients were successfully extracted by choosing from amongstthe filter banks (2 through 13) placed on the spectrum of the Inverse Fast Fourier Transform of the audio signal frame, taking Inverse Logarithm and performing Discrete Cosine Transformation. These describe the energy content of the signal at different frequency intervals.

The tempo was successfully extracted from the audio samples by detecting the beat onsets in each frame of the sample. A majority of music tempos were found to belong to the range 110 – 140 Beats per Minute. The algorithm for Pitch and Mode detection worked with a certain level of accuracy. It extracted the exactly correct song key (both pitch and mode) for three of the samples and an incorrect key for one sample. The pitch and mode of the music samples were extracted successfully by observing the musical notes that comprise the audio sample and matching the stream of notes obtained with the known musical scale compositions.

Also, frequency distribution histograms for the Hainsworth and personal musical sample datasets were plotted as shown in Figure 3 and 4. The histograms showed that for both datasets, majority of the sample tempos lied in the range 120 – 140 Beats Per Minute with the maximum number of samples falling in the 120-130 BPM bin. The histograms were plotted for the range of 80 to 200 BPM and form a hill like shape with almost no samples lying in the range 80 – 100 BPM and 160 – 200 BPM while around 90% of the samples fell in the range of 100 – 160 BPM. This indicates that music produced usually belongs to this tempo range and especially in the range 120 – 140 BPM which would be most common to observe for any new musical piece analysed.

However, different samples of audio differed visibly in terms of their tempos, pitch and modes while the MFCC values also differed but were not directly interpretable. Most values were determined with good accuracy while others could be improved by checking the algorithm/method and performing suitable modifications.

References

1. Egermann, H., Fernando, N., Chuen, L., & McAdams, S. (2015). Music induces universal emotion-related psychophysiological responses: comparing Canadian listeners to Congolese Pygmies. Frontiers in psychology, 5, 1341. doi:10.3389/fpsyg.2014.01341

2. McIntyre H. (2017, November, 9). Americans Are Spending More Time Listening To Music Than Ever Before [Reading]. Retrieved from https://www.forbes.com

3. Lehmann, J., & Seufert, T. (2017). The Influence of Background Music on Learning in the Light of Different Theoretical Perspectives and the Role of Working Memory Capacity. Frontiers in psychology, 8, 1902. doi:10.3389/fpsyg.2017.01902

4. Huron, D. (n.d.). Musical Expectation [Reading]. Retrieved from https://csml.som.ohio-state.edu/Music829D/Notes/Expectation.html

5. Lefevre, M. (2004), Playing with sound: The therapeutic use of music in direct work with children. Child & Family Social Work, 9: 333-345. doi:10.1111/j.1365-2206.2004.00338.x

6. McDonald, J. J., Störmer, V. S., Martinez, A., Feng, W., & Hillyard, S. A. (2013). Salient sounds activate human visual cortex automatically. The Journal of neuroscience : the official journal of the Society for Neuroscience, 33(21), 9194-201.

7. Kent, D. (2010). The Effect of Music on the Human Body and Mind.

8. How Does Music Affect Your Brain? (2007, June, 07) [Reading]. Retrieved from https://www.ashford.edu/online-degrees/student-lifestyle/how-does-music-affect-your-brain

9. Treasure, J. (2009, July). The 4 ways sound affects us [Lecture presented at TEDGlobal 2009]. Retrieved from https://www.ted.com/talks/julian_treasure_the_4_ways_sound_affects_us?language=en

10. Zwaag, M.D., Dijksterhuis, C., Waard, D.D., Mulder, B., Westerink, J.H., &Brookhuis, K.A. (2012). The influence of music on mood and performance while driving. Ergonomics, 55 1, 12-22.

11. Mei-Ching C., Pei-Luen T., Yu-Ting H. &Keh-chung L. (2013). Pleasant music improves visual attention in patients with unilateral neglect after stroke. American Journal of Occupational Therapy, 27:1, 75-82, DOI: 10.3109/02699052.2012.722255

12. Cooper, B. B. (2013, December, 6). The Surprising Science Behind What Music Does To Our Brains [Reading]. Retrieved from https://www.fastcompany.com/3022942/the-surprising-science-behind-what-music-does-to-our-brains

13. Bacon, C. J., Myers, T. R., & Karageorghis, C. I. (2012, August). Effect of music-movement synchrony on exercise oxygen consumption. J Sports Med Phys Fitness, 52(4), 359-65.

14. Logeswaran, N., & Bhattacharya, J. (2009). Crossmodal transfer of emotion by music. Neuroscience letters, 455 2, 129-33.

15. Bora, B & Krishna, M &Phukan, K.D. (2017). The effects of tempo of music on heart rate, blood pressure and respiratory rate – A study in gauhati medical college. Indian Journal of Physiology and Pharmacology. 61. 445-448.

16. McKinney, M., &Breebaart, J. (2003). Features for audio and music classification.

17. Subramanian H. (2004). Audio Signal Classification (Unpublished master’s thesis). Electrical Engineering Department, Indian Institute of Technology, Bombay, India.

18. Farbood, M. M., & Price, K. (2014). Timbral features contributing to perceived auditory and musical tension. In Proceedings of the 13th International Conference on Music Perception and Cognition. Seoul, Korea.

19. Imai, S. (1983, April). Cepstral analysis synthesis on the mel frequency scale. In Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83. (Vol. 8, pp. 93-96). IEEE.

20. Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. The Journal of the Acoustical Society of America, 8(3), 185-190.

21. Wong, E., & Sridharan, S. (2001). Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification. In Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on (pp. 95-98). IEEE.

22. Sachi Gupta, Gaurav Agarwal, Hybrid fuzzy-based Deep Remora Reinforcement Learning Based Task Scheduling in Heterogeneous Multicore-processor, Microprocessors and Microsystems, Volume 92, 2022, 104544, ISSN 0141-9331, https://doi.org/10.1016/j.micpro.2022.104544.

23. Mel Frequency Cepstral Coefficient (MFCC) tutorial. (n.d.) [Reading]. Retrieved from http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

24. Agarwal G, Om H. An efficient supervised framework for music mood recognition using autoencoderbased optimised support vector regression model. IET Signal Process. 2021;15:98–121.https://doi.org/10.1049/sil2.12015

25. The human hearing range - what can you hear? (n.d.) [Reading]. Retrieved from https://global.widex.com/en/blog/human-hearing-range-what-can-you-hear

26. Ronald P. (1992). Introduction to Music. USA: McGraw-Hill. 81

27. Forte A. (January 1, 1979). Tonal Harmony in Concept and Practice. NY: Holt, Rinehart and Winston; 3rd edition

28. Tunebat (www.tunebat.com) [For finding pitch and tempo of songs]

29.Agarwal, Gaurav; Om, Hari; Gupta, Sachi; "A learning framework of modified deep recurrent neural network for classification and recognition of voice mood", Int J Adapt Control Signal Process, 36 (8), 1835– 1859, 2022

30. S. Hainsworth, Data from: "Techniques for the automated analysis of musical audio," [Dataset] Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 2004. Retrieved from http://www.marsyas.info/hainsworth.zip

Table 1 - Common Exercise Tempos and Tempos (bpm)
(source :www.scienntificamerican.com)

Pace\Exercise	**Running**	**Walking**
Slow	140 – 150 bpm	100 – 110 bpm
Moderate	150 – 160 bpm	110 – 125 bpm
Faster	160 – 175 bpm	125 – 135 bpm

Table 2 - Sample set of 15 features extracted from 3 songs chosen at random, over duration of 45 seconds each

Song	**MFCC1**	**MFCC2**	**MFCC3**	**MFCC4**	**MFCC5**	**MFCC6**	**MFCC7**	**MFCC8**	**MFCC9**	**MFCC10**	**MFCC11**	**MFCC12**	**Tempo (BPM)**	**Pitch**	**Mode**
Tera ChehraKitna Suhana	2.2767549	-0.3404742	0.3579539	-0.028353	-0.0468483	-0.1620961	-0.0216786	-0.0955584	-0.1752817	-0.1193951	-0.1457302	-0.0195425	100.74	C#	Minor
Agar Tum Saath Ho	2.2565295	-0.4734121	0.1914611	-0.5297528	-0.0417666	-0.2831263	-0.1215863	-0.3105153	0.00444913	0.16392513	0.23394406	0.07610869	124.47	C#	Major

Humble – Kendrick Lamar	1.22002233	-0.3482428	0.0083439	-0.1868111	-0.0155231	0.15721232	0.06377706	0.09610684	0.06130609	-0.0431956	0.01136505	0.0102065	151.98	F#	Major
Wish You Were Here	1.8542867	-1.6961405	0.5493195	-0.3537776	-0.1493355	-0.0185418	-0.0780571	-0.3301412	0.0820423	-0.1753286	0.01637394	0.02679603	120.99	G	Major

Figure 1 -Non-linearity of Mel-Scale with respect to Hertz Scale.

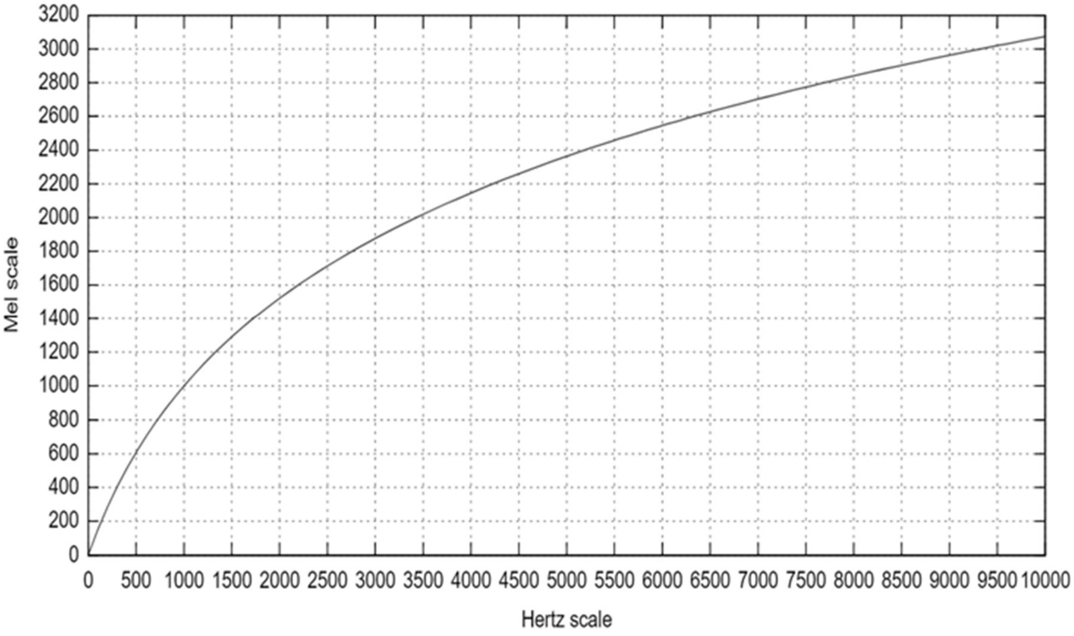


Figure 2 - Mel-Scale Triangular Filter Banks.

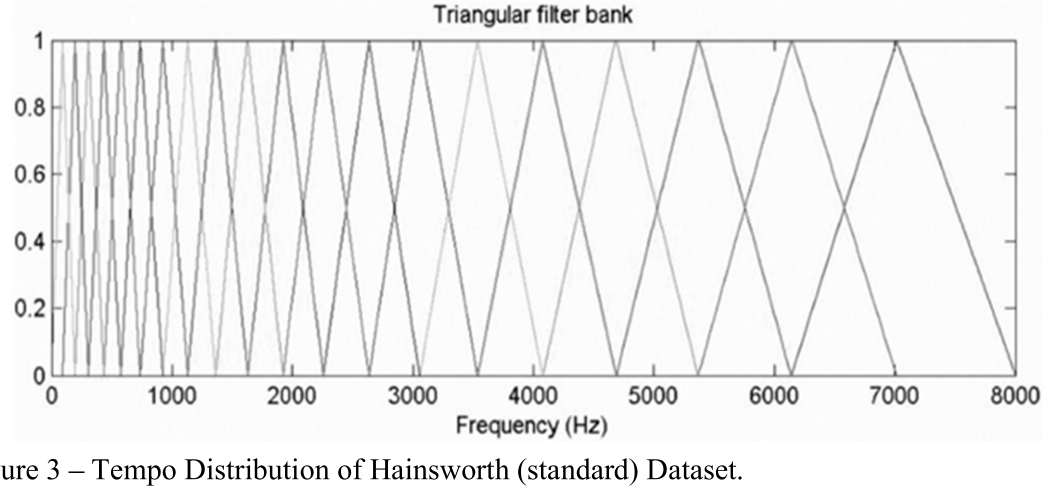


Figure 3 – Tempo Distribution of Hainsworth (standard) Dataset.

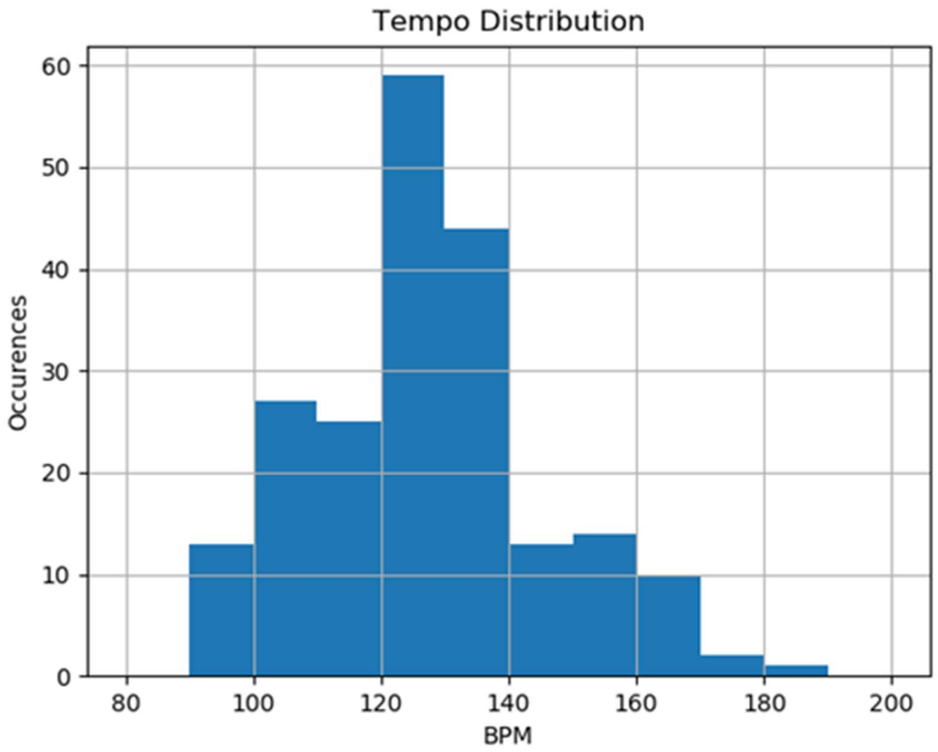
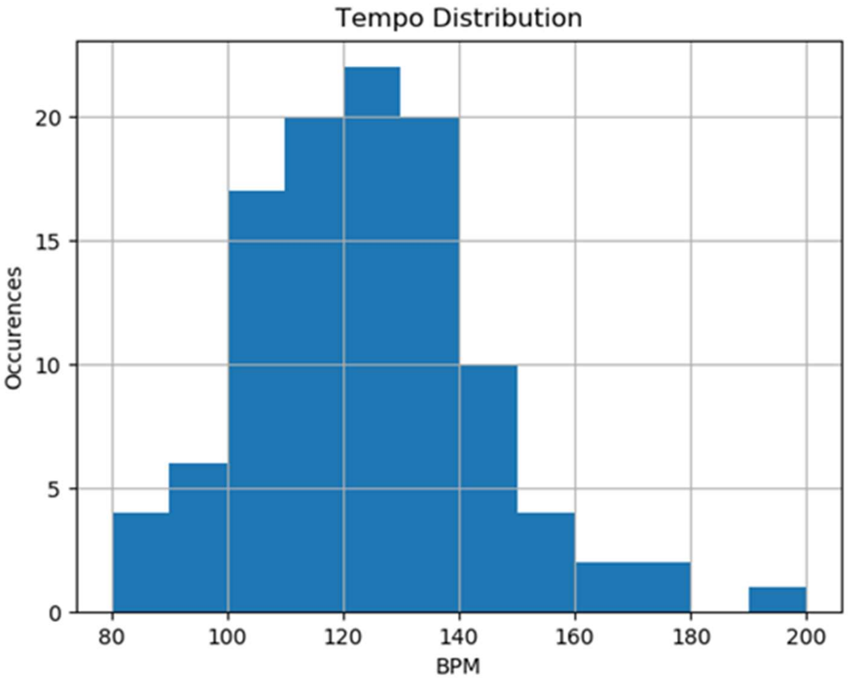


Figure 4 – Tempo Distribution of Personal Dataset.


Tempo Distribution
Occurences
BPM