A Review of Music Analysis Techniques

Kaustubh R Kulkarni[1], Sowmiya Raksha R Naik[2]

[1]*Adhoc Faculty Member, Department of Computer Engineering and Information Technology,*
Veermata Jijabai Technological Institute, Mumbai, India.
[2]*Assistant Professor, Department of Computer Engineering and Information Technology,*
Veermata Jijabai Technological Institute, Mumbai, India.

---***---

Abstract - *Music analysis can be done in various ways. Music piece at hand can be decomposed into its component instrumental tracks and/or vocal tracks and also its constituent noise. The instrument being played in each of the instrumental tracks or its class can be recognized. Perceptual attributes of music like pitch and tempo can be estimated. Musical notes constituting the music piece can also be extracted and recognized. This paper reviews the various approaches that have been used for music analysis.*

***Key Words*: music; music analysis; musical instrument extraction; musical instrument recognition; musical note extraction; musical note recognition; musical pitch and tempo analysis.**

1. INTRODUCTION

This paper reviews the following approaches to analyze a piece of music:

1. Instrument extraction and recognition
2. Pitch estimation
3. Tempo estimation
4. Extracting and recognizing notes

Section 2 reviews some of the techniques for musical instrument extraction and recognition. Pitch estimation techniques have been discussed in section 3. Tempo estimation techniques have been discussed in section 4 and notes extraction and recognition techniques have been discussed in section 5. Section 6 concludes by summarizing the techniques used in all the previous sections.

2. MUSICAL INSTRUMENT EXTRACTION AND RECOGNITION

The music signal has been represented using a matrix by Serrano et al [1]. Complex non-negative matrix factorization involves finding two non-negative real matrices $W \in R_{\geq 0}^{M \times K}$ and $H \in R_{>0}^{K \times N}$ such that the magnitude of the complex matrix $S \in C^{M \times N}$ that represents the music signal can be expressed as in equation 2.1 [1]:

$$V(m,n) = |S(m,n)| \approx \sum_{k=1}^{K} W(m,k)H(k,n)e^{i\emptyset_k(m,n)} \ldots (2.1)$$

where

$$m \in (0, \ldots, M)$$

$$n \in (0, \ldots, N)$$

and the matrix

$$\emptyset_k \in C_{\geq 0}^{M \times N}$$

contains phase information for every element in matrix S [1]. The columns of matrix W have been used to obtain information about the spectral energy distribution of a sound source, while the rows of matrix H have been used to obtain the timing and intensity information about the corresponding source. Serrano et al [1] have found that the inclusion of phase information helps in more accurate factorization in cases where there are partial overlaps between individual tracks using normal non-negative matrix factorization.

Serrano et al [1] have modified the factorization to consider all possible shifts s in the pitch due to vibrato by modifying the complex matrix factorization equation 2.1 as [1]:

$$V(m,n) = |S(m,n)| \approx \sum_{k=1}^{K} \sum_{s=1}^{S} W(m-s,k)H_s(k,n)\, e^{i\emptyset_k(m,n)} \ldots\ldots\ldots\ldots (2.2)$$

One of the variants of non-negative matrix factorization is non-negative matrix partial co-factorization (NMPCF). It has been used by Hu and Liu [2] to separate the music signal into singing voice portion and accompaniment portion. Let X denotes the spectrogram matrix of music signal x where X_{ft} represents the magnitude of the f^th^ frequency bin at the t^th^ time frame. X has been factorized by applying NMF as X=UV as follows [2]:

$$X = U_S V_S + U_A V_A, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.3)$$

where the subscript S indicates the voice component and the subscript A indicates the accompaniment component. The equation 2.3 is called co-factorization of matrix X. The non-vocal portions of signal x consists of accompaniment component only and have been used by Hu and Liu [2] in co-factorization as a part of prior knowledge. Hu and Liu [2] have also used a spectrogram of clean singing voice, as prior knowledge.

The characteristics of music like dynamics, timbre, tonality, rhythm have been used to extract features from music. These features in the form of feature vectors can be given as inputs to the neural network for instrument recognition. Neural networks have been used by Masood et al [3] for instrument recognition. Masood et al [3] have extracted the following features:

1. 'The number of sign changes in the signal per second is called the zero crossing rate'. [3]
2. 'The frequency below which a certain percentage of the energy in the sound is confined is called the spectral roll off'. [3]
3. 'The measure of the symmetry of the spreading of the spectrum around the mean of the distribution is called the skewness'. [3]
4. 'The proportion (denoted between 0 and 1) of the energy within the sound above a certain frequency is called the brightness'. [3]
5. 'The ratio between the geometric mean and arithmetic mean of the distribution is called flatness'. [3]
6. 'The measure of how short and wide or tall and thin a spectrum is kurtosis'. [3]
7. 'Root mean square energy is computed by taking the root of the average of the square of the amplitude'. [3]
8. The Mel-frequency Cepstral Coefficients (MFCCs).

A sequence of frames, where each frame is described in terms of the above stated features, has been given as an input to a neural network classifier. [3] Such a sequence of frames is called a context window. Sharma [4] has found such sequence of frames useful because adjacent frames have strong correlation and may form patterns in the time domain.

Uhlich et al [5] have applied short time Fourier transform (STFT) on context window for feature vector generation. This feature vector has been fed to a deep neural network (DNN) with rectified linear unit (ReLU) layers. The phase information from input features has then been combined with the magnitude information from DNN outputs to estimate the STFT of the target instrument. Finally the signal has been converted back to the time domain using an inverse STFT. [5]

Multiple multi-layered perceptrons (MLP) can be stacked together and trained as a whole. [4] Non-linear activation functions such as sigmoid [4]:

$$f(x) = \frac{1}{1+e^{-x}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.4)$$

or hyperbolic tan[4]:

$$f(x) = \tanh(x) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.5)$$

are used at each layer in a multilayer perceptron network. Sharma [4] has designed a neural network that estimates background music. Sharma [4] has tried various types of artificial neural networks such as MLP, stacked MLP, convolutional neural network combined with hidden Markov model (CNN-HMM), restricted Boltzmann machine (RBM), stacked auto encoder and deep belief network. The SNR values in decibels have been calculated for the estimated background music predicted by the neural network models. [4]

The harmonics of a musical instrument may overlap, causing destructive or constructive interference. Every musical instrument has a characteristic spectral signature, which Donnelly and Sheppard [6] have put to use in instrument classification. An amplitude threshold has been established for each instrument signal to distinguish the harmonics from the noise part. The peaks that exceed the threshold have been extracted and the frequency location of each peak has been noted [6]. The fundamental frequency $f0$ has been identified as the significant peak with the lowest frequency. The ratios of frequencies of peaks to the fundamental frequency values have been computed. The vector of ratios has then been clustered using k-means clustering algorithm to learn the locations of those harmonics that are important to each instrument. For each cluster, the mean and standard deviation have been noted and used as the instrument's spectral signature. [6]

Another signature for musical instruments has been proposed by Singh and Kumar. [7] It is in terms of the durations of the attack state, the steady state and the decay state. Singh and Kumar [7] have also suggested the following features, related to the music timbre, for musical instrument recognition:

1. Zero Crossing Rate
2. Spectral Centroid
3. Spectral Roll-Off
4. Spectral Flux
5. Mel Frequency Cepstral Coefficients (MFCC)

Singh and Kumar [7] have also extracted rhythmic features from the beat histogram, which are as follows:

1. Rhythmic regularity
2. Beat strength which is evaluated using the following statistical measures [7]:

    a. Mean
    b. Variance
    c. Standard deviation
    d. Skewness (third order central moment)
    e. Spectral Kurtosis (fourth order central moment)
    f. Spectral crest factor
    g. Spectral spread
    h. Dynamic range

Singh and Kumar [7] have used Gaussian Mixture Model (GMM) trained using expectation minimization (EM) algorithm for instrument class recognition after the features have been extracted.

Bhalke et al [8] have first detected the silent parts by dividing the signal into frames and then have compared the energy of each frame with an energy threshold. Then those silent frames have been removed and MFCC features have been extracted from the resulting signal. Dynamic time warping (DTW) score has been computed by comparing each frame with a reference template. DTW has been used to measure the similarity and to find the optimal alignment between two time series (here, music signals) which may vary in speed. Finally classification has been performed using k nearest neighbor (k-NN) classifier with k=1. [8]

The continuous wavelet transform (CWT) of the signal x (t) is calculated as [9]:

$$T_x(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} x(t)\varphi(\frac{t-b}{a})dt \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.6)$$

where a is the scaling parameter or scale, b is shifting parameter or shift and $\varphi(t)$ is the wavelet basis function. Foomany and Umapathy [9] have first calculated the contribution of each scale in CWT to the total energy and then have captured these features from the signal:

1. Scale Distribution Width (SDW): It has been defined [9] as the difference between the scales, say a1 and a2, at which the CWT of the signal reaches half of its maximum of average energy of each scale over all possible shift values.
2. Log SDW: It has been defined [9] as:

$$log\ SDW = \log a1 - \log a2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.7)$$

   where a1 and a2 are the scales as mentioned in point 1 above.
3. Dominant Scale: It has been defined as 'the scale at which the energy of CWT coefficients (averaged over a frame) is maximum'. [9]
4. Time Variance of Dominant Scale: It has been defined as 'the standard deviation of values of dominant scale at each shift'. [9]
5. Wavelet Mean of Inverse Scale (WMIS): It has been defined as [9]:

$$WMIS = \sum_a \sum_b \frac{|T_x(a,b)|}{\sum_a |T_x(a,b)|} \times \frac{1}{a} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.8)$$

The above stated features have been used to capture the following three traits in the signal [9]:

1. Wavelet based bandwidth characteristics
2. Dominant wavelet scale
3. Wavelet-based temporal variation of dominant scale

For classification task, Foomany and Umapathy [9] have used linear discriminant analysis (LDA).

3. PITCH ESTIMATION

'The pitch of a musical instrument note is primarily determined by its fundamental frequency of oscillation as perceived by a human'. [10]

A formula to convert f hertz into m mels is defined as [10]:

$$m = 1127 \log_e\left(1 + \frac{f}{700}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

Singh and Kumar [13] have used four pitch detection algorithms:

1. Autocorrelation function (ACF) is defined as [10] :

$$A_c(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau)\,, \ldots\ldots\ldots\ldots\ldots\ldots (3.2)$$

where $x(n)$ is the input signal and τ is the lag (delay) value $\tau = 0, \pm 1, \ldots$. The peaks in $A_c(\tau)$ have been used to estimate the pitch period and hence the pitch.

2. Average Magnitude Difference Function (AMDF), is defined as [10] :

$$x(m) = \frac{1}{N-m-1}\sum_{n=0}^{N-m-1} s(n+m) - s(n) \ldots\ldots\ldots (3.3)$$

where $s(n)$ is the input signal and $0 \leq m \leq N$. The values of m for which $x(m)$ becomes minimum has been the pitch period.

3. Cepstrum of a signal is defined as [10]:

$$c[n] = F^{-1}\{\log F\{s(n)\}\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.4)$$

where F is the DFT and F^{-1} is the IDFT. The frequency domain in spectrum is called quefrequency domain in the cepstrum. Peaks have been searched in the cepstrum and further procedure used to find the fundamental frequency has been the same as the autocorrelation method.

4. An all-pole filter can be used to model sound signals as [10]:

$$H(Z) = \frac{S(Z)}{E(Z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.5)$$

Linear predictive coding (LPC) method has been used 'to predict the current sample as a linear combination of its past p samples'. [10] (taking inverse Z transform of $H(Z)$ in equation 3.5) :

$$s(n) = \sum_{k=1}^{p} a_k s[n-k] + e(n) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.6)$$

where $s(n)$ is the input signal, $e(n)$ is the error, a_k is a parameter and p is called the order of LPC. LPC based method can detect formant frequencies, which are different from fundamental frequency.

The pitch determination task has been approached as a two-step process by Su et al [11]:

1. pitch candidate selection using GMM and CNN
2. pitch tracking by generating a continuous pitch contour

Pitch states have been created which have been calculated as [11]:

$$s = \left\lfloor \log_2\left(\frac{p}{60}\right) \times 24 \right\rfloor \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.7)$$

Using equation (3.7), the pitch range p has been divided into 60 parts with 59 pitch states s, each state having 24 frequency bins in an octave. The input signal has been split into frames, then the frames have been grouped into windows and finally the windows have been given as an input to the convolutional neural network (CNN). The output of CNN has been used to predict the prominent pitch state, which is then used to calculate the following probability distribution function (pdf) on pitch values [11]:

$$p(z) = \sum_{k=1}^{K} \alpha_k N(z; \mu_k, \sigma_k^2) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.8)$$

where,

- $p(z)$ is the pdf for a Gaussian mixture model (GMM).
- $N(z; \mu_k, \sigma_k^2)$ is the Gaussian distribution that is used to model each pitch state.
- μ_k is the mean of the Gaussian distribution, is the centre frequency of the pitch state modeled by the Gaussian distribution.
- σ_k is the standard deviation of the Gaussian distribution is half the bandwidth of the pitch state modeled by the Gaussian distribution.
- α_k is a coefficient such that $\sum_{k=1}^{K} \alpha_k = 1, \alpha_k \geq 0$.

The pitch probability values have been tracked from frame to frame to generate a pitch contour on the basis of temporal continuity of pitch which has been modeled using Laplacian distribution [11]:

$$p_t(\Delta) = \frac{1}{2\sigma} \exp\left(-\frac{|\Delta - \mu|}{\sigma}\right) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (3.9)$$

- Δ is the change in pitch period from one frame to the next.
- μ is the location parameter
- σ is the scale parameter.

Bellur and Murthy [12] have defined tonic pitch as 'the reference note established by the lead performer, relative to which other notes in a melody are structured'. [12] For identifying tonic pitch, the signal has been split into frames. Then DFT has been applied to form an amplitude spectrum. Cepstrum has been obtained by taking logarithm of the amplitude spectrum. Finally inverse DFT (IDFT) has been applied to cepstrum to bring the signal back to time domain. Tonic pitch has been obtained using the knowledge of peak locations in the time domain signal and sampling rate. [12]

In this process some low energy frames may get masked by other frames. Bellur and Murthy [12] have proposed measuring short term energy after DFT calculation to detect these low energy frames. Then cepstrum and IDFT of these low energy frames have been calculated and peaks have been used to estimate the tonic pitch. [12]

In another method proposed by Bellur and Murthy [12], Euclidean NMF (ENMF) has been applied to the low energy frames. Only the spectral basis vectors component obtained from ENMF have been used for cepstrum and IDFT calculation and then for estimation of the tonic pitch. [12]

4. TEMPO ESTIMATION

'Tempo is indicated by beats per minute (BPM) or is expressed as tempo annotation in words such as "fast"'. [13]

Note onsets have been used for tempo estimation by methods like measuring inter-onset interval, calculating ACF of onset detection function (ODF), calculating DFT of ODF or a combination/variation of these methods. Note onsets are detected by abrupt increase in spectral amplitude of the signal. The spectral amplitude is defined as 'the sum of the spectral bins at each instant in time' [14]:

$$SA[n] = \sum_{k=0}^{\frac{N}{2}-1} |X[n,k]| \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4.1)$$

To detect abrupt changes in the spectral amplitude, a bi-phase filter $h[n]$ is used. Detecting abrupt changes in spectral amplitude is used to detect note onsets [14]:

$$ODF[n] = SA[n] * h[n] \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4.2)$$

Another method to detect note onsets is using the Kullback-Leibler (KL) divergence [14]:

$$\tilde{X}[n,k] = \sum_{i=n}^{n+4} |X[i,k]|\,;$$

$$\tilde{X}[n,k] = \frac{\tilde{X}[n,k]}{\sum_{k=0}^{\frac{N}{2}-1} \tilde{X}[n,k]} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.3)$$

$$ODF_{KL}[n] = \sum_{k=0}^{\frac{N}{2}-1} \tilde{X}[n+1,k] \log\left(\frac{\tilde{X}[n+1,k]}{\tilde{X}[n,k]}\right) \ldots\ldots\ldots (4.4)$$

A spike in the K-L divergence reflects a sudden change in the spectrum shape due to note onset. [14]

Yet another method is to calculate Euclidean distance between successive frames and then finding its derivative using bi-phasic filter to detect abrupt changes in Euclidean distance (ED), indicating note onsets [14]:

$$ED[n] = \sqrt{\left(\sum_{k=0}^{\frac{N}{2}-1} (|X[n+1,k]| - |X[n,k]|)^2\right)} \ldots\ldots\ldots\ldots (4.5)$$

$$ODF_{ED}[n] = ED[n] * h[n] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.6)$$

Wu [13] has obtained pairs of estimated tempo values. To do so, Wu [13] has passed the ODF through a Gaussian low pass filter and afterwards subtracted the local mean. Then STFT has been applied to the filtered signal, giving a tempogram, a plot of local tempo values in beats per minute at different time instants.

The tempogram has been used to obtain a long term periodicity (LTP) curve, which in turn has been used to select a pair of tempo candidates as local maxima in LTP curve. A probability mass function (pmf) has been obtained from LTP curve and probability of the tempo pair has been defined by Wu [13] as the sum of the probability of individual candidates in the pair. The pair has then been included in one of the four classes depending on whether one tempo in the pair is twice, thrice, 3/2 times, or has some other ratio with the second tempo in the pair. The highest probability pair from each of the four classes has been selected and these pairs have then been combined to form a four dimensional feature vector.

Then Wu [13] has extracted tempogram stripes around each tempo candidate within a pair and further has used them to obtain the LTP and then LTP has been used to obtain the pmf. Five statistics of the pmf $[\gamma_s, \kappa_s, \mu_s, \sigma_s, cv_s]$ have been calculated to decide the dominant among the two tempo candidates in the pair, where:

1. γ_s is the skewness
2. κ_s is the kurtosis
3. μ_s is the mean
4. $\sigma_s$is the standard deviation
5. cv_s is the coefficient of variation

Wu [13] has used these statistics to form a 'tempogram shape vector (tsv)' to be used by the k-Nearest neighbor [k-NN] and support vector machine (SVM) classifiers.

Wu and Jang [15] have applied low pass filtering (LPF) to the music signal before tempo estimation. Then the prominent tempo pair has been estimated . In the tempo pair estimation process, an onset detection function has been applied to the signal. The STFT has been applied to the output of the onset detection function which results in a tempogram. 'Tempogram represents the periodicity of onsets'. [15] Then the tempogram has been weighted to reflect the human perceptual aspect of the tempo. [15]

In tempo-pair generation, LTP has been calculated using the tempogram. The local maxima within a particular threshold in the LTP have been identified as tempo pair candidates. The tempo estimates from both the low pass filtered signal as well as non LPF signal have been considered. [15]

5. EXTRACTING AND IDENTIFYING NOTES

Musical instrument classification has been used as a first step by Zhu et al [16] in note onset detection. The musical instruments have been classified into the following four classes [16]:

1. Pitched percussive (PP) e.g. piano
2. Pitched non-percussive (PNP) e.g. bowed string instruments
3. Non-pitched percussive (NPP) e.g. drum
4. Wind instruments, which are non-pitched and non-percussive.

The class of an instrument has been predicted from the music piece using hidden Markov model (HMM) classifier and then the appropriate detection method has been chosen for finding the note onset times. [16] Five note-onset detection methods have been evaluated by Zhu et al [16]:

1. Energy method which has been found to be the most suitable for NPP instruments like drum. It is obtained by a first order Gaussian filter [16]:

$$(\acute{h}(n)) = -\left(n - \frac{N}{2}\right) e^{-\frac{\left(n-\frac{N}{2}\right)^2}{2\sigma^2}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5.1)$$

- n is an independent variable
- N is length of the filter

2. Complex phase method: The amplitude and phase of music signal have been obtained from its polar coordinate representation. When the phase is stable, the phase difference is uniform. So the phase of a particular frame has been obtained using the knowledge of phase value of its previous

frames and phase difference. The amplitude of that frame has been calculated using the amplitude of its previous frame.

3. Spectral flux method has been found to be suitable for PP instruments like piano and wind instruments [16]:

$$SF(k) = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} H(|S(\omega, k)| - |S(\omega, k-1)|) \dots \dots \dots (5.2)$$

where $H(x)$ is a half wave rectifier function.

4. Wrapping-compensation correlation detection has given best results in recognizing PNP instruments like bowed string [16]:

$$d(t_0) = \frac{1}{\sum_{f=t_0-\left[\frac{w}{2}\right]}^{t_0+\left[\frac{w}{2}\right]} \prod_{b=1}^{b_0} \left(\left(\frac{c_b(t_0, f)}{\sqrt{c_b(t_0, t_0) c_b(f, f)}} \right)^{W_b} \right)} \dots \dots (5.3)$$

where

- w represents the relevant frame number
- b_0 represents the number of sub-bands
- t_0 denotes the current frame
- c_b is the covariance of the two frames of the same sub-band
- W_b is the compensation coefficient to account for vibrato.

5. Conditional independent component analysis (ICA) has found useful for NPP instruments like drum. The entropy of the signal is the highest at the note onset. The signal is a linear synthesis of independent components as follows [16]:

$$x = As \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (5.4)$$

The entropy of the signal is calculated as [16]:

$$S(x) = \log(\det A) - \sum_{i=1}^{n} \log f_i(s_i) \dots \dots \dots \dots \dots \dots (5.5)$$

where

- n is the number of independent components
- f_i is the pdf of the ith independent component.

While on one hand the instrument class has been used in detecting note onsets, on the other hand the characteristics of notes have been used to classify the instruments by Hu and Liu [17]. Hu and Liu [17] have proposed a note onset detection function as [17]:

$$D(n) = \frac{1}{\sum_{m=n-M}^{m=n-1} \frac{c(n, m)}{\sqrt{c(n, n) c(m, m)}}} \dots \dots \dots \dots \dots \dots \dots (5.6)$$

M = total number of frames,
n = frame number
c(n,m) = covariance of the nth and mth frame.

Note onsets have been detected from the output of the above function convolved using first order Gaussian differentiator. [17]

Turrubiates et al [18] have converted the signal to the frequency domain using Fourier transform. Harmonic product spectrum (HPS) $Z[n]$ has been used to estimate the pitch and it has been calculated as follows [18]:

$$Z[n] = \prod_{m=1}^{N} Y[mn] \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (5.7)$$

$$n = 0, 1, 2, \dots$$

where $Y[n]$ is the power spectrum. $Z[n]$ is the product of all harmonics of the fundamental frequency and hence fundamental frequency amplitude rises in the HPS Z[n]. The harmonics have been calculated as follows [18] to reduce the effect of external factors like temperature:

$$Y[mn] = Y[n-1] \times Y[mn-1] + Y[n] \times Y[mn] + Y[n+1] \times Y[mn+1] \dots \dots \dots \dots \dots (5.8)$$

A variance threshold thr has been used for feature vector dimensionality reduction from a data set M of size n to a new data set $TrainSet$ of size m with $n \gg m$ [18]:

$$M = [\vec{f_1}, \vec{f_2}, \dots, \vec{f_n}] \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (5.9)$$

$$TrainSet = [\vec{\eta_1}, \vec{\eta_2}, \dots, \vec{\eta_m}] \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots (5.10)$$

where $\vec{\eta_1}$ are the vectors $\vec{f_2}$ whose variance $\sigma_i^2 \geq thr$. These feature vectors have then been given as an input to an MLP neural network trained to recognize the note being played. [18]

6. CONCLUSION

The above discussion has highlighted the approaches used in music analysis. It reviews a range of methods for musical instrument separation, musical instrument recognition, tempo and pitch estimation and extraction and recognition of musical notes. It can be concluded that the instrumental components of music, musical notes and some perceptual attributes of music, namely, the pitch and the tempo can be estimated using machine learning approaches, artificial neural networks, deep learning approaches, digital signal processing and other semi-automatic methods.

ACKNOWLEDGEMENT

We thank Prof. Dr. V B Nikam, Head of Department of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute, Mumbai who provided valuable insights and expertise that greatly assisted the research.

REFERENCES

1. F. J. Rodriguez-Serrano, S. Ewert, P. Vera-Candeas and M. Sandler, "A score-informed shift-invariant extension of complex matrix factorization for improving the separation of overlapped partials in music recordings," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 61-65.

2. Y. Hu and G. Liu, "Separation of Singing Voice Using Nonnegative Matrix Partial Co-Factorization for Singer Identification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 4, pp. 643-653, April 2015.

3. S. Masood, S. Gupta and S. Khan, "Novel approach for musical instrument identification using neural network," 2015 Annual IEEE India Conference (INDICON), New Delhi, 2015, pp. 1-5.

4. V. Sharma, "A Deep Neural Network based approach for vocal extraction from songs", 2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, 2015, pp. 116-121.

5. S. Uhlich, F. Giron and Y. Mitsufuji, "Deep neural network based instrument extraction from music," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, 2015, pp. 2135-2139.

6. P. J. Donnelly and J. W. Sheppard, "Cross-Dataset Validation of Feature Sets in Musical Instrument Classification," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, 2015, pp. 94-101.

7. C. P. Singh and T. K. Kumar, "Efficient selection of rhythmic features for musical instrument recognition," 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, Ramanathapuram, 2014, pp. 1393-1397.

8. D. G. Bhalke, C. B. Rama Rao and D. S. Bormane, "Dynamic time warping technique for musical instrument recognition for isolated notes," 2011 International Conference on Emerging Trends in Electrical and Computer Technology, Tamil Nadu, 2011, pp. 768-771.

9. F. H. Foomany and K. Umapathy, "Classification of music instruments using wavelet-based time-scale features," 2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), San Jose, CA, 2013, pp. 1-4.

10. C. P. Singh and T. K. Kumar, "Efficient pitch detection algorithms for pitched musical instrument sounds: A comparative performance evaluation," 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi, 2014, pp. 1876-1880.

11. H. Su, H. Zhang, X. Zhang and G. Gao, "Convolutional neural network for robust pitch determination," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 579-583.

12. A Bellur and H. A. Murthy, "A cepstrum based approach for identifying tonic pitch in Indian classical music," 2013 National Conference on Communications (NCC), New Delhi, India, 2013, pp. 1-5.

13. F. H. F. Wu, "Musical tempo octave error reducing based on the statistics of tempogram," 2015 23rd Mediterranean Conference on Control and Automation (MED), Torremolinos, 2015, pp. 993-998.

14. T. P. Vinutha, S. Sankagiri and P. Rao, "Reliable tempo detection for structural segmentation in sarod concerts," 2016 Twenty Second National Conference on Communication (NCC), Guwahati, 2016, pp. 1-6.

15. F. H. F. Wu and J. S. R. Jang, "A supervised learning method for tempo estimation of musical audio," 22nd Mediterranean Conference on Control and Automation, Palermo, 2014, pp. 599-604.

16. B. Zhu, J. Gan, J. Cai, Y. Wang and H. Wang, "Adaptive onset detection based on instrument recognition," 2014 12th International Conference on Signal Processing (ICSP), Hangzhou, 2014, pp. 2416-2421.

17. Y. Hu and G. Liu, "Dynamic characteristics of musical note for musical instrument classification," 2011 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Xi'an, 2011, pp. 1-6.

18. J. d. J. Guerrero-Turrubiates, S. E. Gonzalez-Reyna, S. E. Ledesma-Orozco and J. G. Avina-Cervantes, "Pitch estimation for musical note recognition using Artificial Neural Networks," 2014 International Conference on Electronics, Communications and Computers (CONIELECOMP), Cholula, 2014, pp. 53-58.