Genre Classification of Songs Using Neural Network

Anshuman Goel,Mohd. Sheezan
Department of Computer Engg.
Jamia Millia Islamia
New Delhi, India
leocrimson@rediffmail.com, sheezanjmi@yahoo.in

Sarfaraz Masood, Aadam Saleem
Department of Computer Engg.
Jamia Millia Islamia
New Delhi, India
sarfarazmasood2002@yahoo.com,
aadam.saleem@gmail.com

Abstract—**The objective here is to eliminate the manual work of classifying genres of song in each song. With this startup work songs can be classified in real-time and proposed parallel architecture can be implemented on the multi-processing system as well. In this paper a set of features are obtained like beats/tempo, energy, loudness, speechiness, valence, danceability, acousticness, discrete wavelet transform etc., using Echonest libraries and are fed into the Parallel Multi-Layer Perceptron Network to obtain the genres of the song. The proposed scheme has an accuracy of 85% when used to classify two genres of songs that are Sufi and Classical.**

Keywords—genre; classification; songs; echonest; multi-layered perceptron

I. Introduction

A large amount of human labor and time is required when someone starts classifying songs in his music library. Or suppose these days' people around the globe likes and prefer to listen to the songs of a particular genre. Many online music commercial sites are providing classification based on genre but still this task is done manually tilled now over these websites. The tedious task of first listening to a song and then deciding the song to which genre it belongs to doesn't suit the 21st century. To eliminate this, we here have tried to propose a new scheme for genre classification of songs.

The previous work done on the genre classification of songs by Lindasalwa Muda [1]and Carlos N. Silla Jr. [5] was based upon Mel-frequency Cepstral Coefficients (MFCC). The highest percentage accuracy acquired is nearly about 65% which is not suitable for deployment in real time classification. This may be because MFCC are mainly used for speech recognition. MFCC can be said as an artificial human ear as it was designed in consideration of the human auditory system.The work by Masato Miyoshi is based upon the mood of the song [7] i.e., valence was proposed which is an interesting feature that can be extracted from a song but what about a sad Rock and Pop song. The classifier based on the mood will surely put both these two different genre songs under the same title which will not be acceptable to a genre classifier as mood is only detecting what type of song i.e., happy, sad, cheerful, etc., it belongs to. But for e.g., Rock songs are considered to be cheerful the genre classification based upon mood can work up to a certain limit but not always. Benyamin Matityaho proposed a scheme based upon the Fast Fourier Transform (FFT) [2]. It also proposes a method to provide 100% accuracybut this is generally not possible because some genres like Classical and Sufi look like similar but are very distinct. Moreover, there is no practical evidence that 100% accuracy can be achieved by this process. Alessandro L. Koerich [3] proposes to divide the songs into the segments and then extracting feature finds the accuracy in between 73% to 95% for different segments. Here the approach of dividing the songs into segments and then work on these segments is extolled. But here accuracy based on some segments is taken into consideration. It may be possible that an important section is missed by not taking it into consideration. Thus a measure which takes into account the overall song is needed.Zhouyu Fu, Guojun Lu [6] using Naïve Bayes Classifier produces an accuracy of 77% which cannot be deployed on commercial scale as for the commercial reasons. Vikramjit Mitra and Chia J. Wang [4] proposes a set of features for classing songs genre with an accuracy of 84% using features beat information, MFCC, DWT, energy, and power. Although the results are appreciable, the number of inputs that are there in neural network is very large. Therefore system will not be as fast as needed.

Extending [4]and using a different Neural Architecture and new sets of features like acousticness and valence with beats, speechiness, energy, loudness, danceability and discrete wavelet transform, we not only reduced the feature vector length to 34 (as opposed to 174 in [4]) which in turn will reduce the calculation overheadwith a great extent. While deciding features a deep thought on every feature, along with the previous work has been taken into account. The systemdesigned will bemuch faster and reliable while working in a distributed environment because of the classifier used. The neural network designed is also somewhat dissimilar to the commonly used one. The structure of neural network is based upon the number of experiments and observations done while implementing the classifier. The classifier is based upon the Parallel Multi-Layer Perceptron model. Moreover, we are able to increase the accuracy to 85% which atleast establish a ground for further or future work on a commercial or industrial scale.

978-1-4799-6758-2/14/$31.00 ©2014 IEEE

II. Feature Extraction

A total of 8 features are used for the classification of songs that are number of beats per minute (bpm), loudness, energy, danceability, speechiness, valence, acouticness and discrete wavelet transform are used. Some of these feaures are dervied mathematically while some are derived by using Echonest API.

A. Beats Periodicity

The first thing that comes to every music lover is the beats of a particular song. The beats complexity and periodicity gives the song its beauty. Moreover and technically, it has been observed that most of the songs in a particular genre have nearly the same beat periodicity. That is why first feature included in the set is beats per minute. Along with beats per minute (bpm) average, its variance and skewness are also determined so that entire model of a song can be approximated well.

$$\text{bpm} = \frac{\text{Number of beats in the song}}{\text{Duration of song in minutes}} \quad (1)$$

B. Loudness

Loudness is an important factor if we look at the songs of different genres. Comparing a classical or a jazz song with that of the Rap or Pop, the later is found to have greater loudness then the earlier one. The loudness is measured in decibels (dB) and it can be correlated to the amplitude of the sound wave. Here loudness is determined by combining each segment obtained from the json file created using Echonest API and then determining the average and variance of the whole song considering each and every segment of the song.

C. Energy

Energy, a measure of intensity can play a sound role in determining the genre of a song. For example, if a Classical song is compared with a Rock song, the Classical song will always have lower energy levels then a Rock song. It can be even seen that in most songs the distribution of energy is also even across the genres as well over the song itself. This can be easily seen by looking at the energy spectrogram of the songs of a particular genre. Taken into account this fact, energy is also determined based on the general entropy, timbre and loudness as perceived from the song [8]. The energy here it is determined by using Echonest API.

D. Speechiness

Speechiness is defined as the voice content in the song. The songs with high voice content like Rap songs have high value then compared to other genres. This speechiness factor can also be linked to MFCC coefficients which also gives the information about the vocal content in the song. Mel-frequency Cepstral Coefficient (MFCC) was developed to represent a human ear in the technology world. MFCC were mainly used in the area of speech recognition, but it can be usefully employed for genre classification of songs. It is so because of the speech patterns found in the songs of a particular genre are similar in many ways. For example, if we consider Indian classical songs, a frequent use of ragas. With the identification of ragas we can easily classify these songs as Classical. In MFCC, the signal is divided into frames about 20 ms and passed through the mel-filter bank which uses the mel-scale. After that de-convolution of the source s(n) and impulse filter h(n) is done logarithmically.

$$s(n) \times h(n) = x(n) \quad (2)$$

Equation (2), becomes

$$\log(x(n)) = \log(s(n)) + \log(h(n)) \quad (3)$$

After (3), discrete cosine transform is applied and from the result 13 coefficients are retained [9]. To reduce the dimension of the matrix obtained from each window linear prediction is applied to obtain the value. Thus, the values while comparing a Rap song and a Classical song the speechiness of rap song is always found to higher which is also evident because Rap songs are much more like a speech then a Classical one.

E. Acousticness

Acousticness differentiates a natural sound with that of the electrical sound. Natural sound of drum sticks, harmonium etc., are found in the Classical or Sufi songs while in the Rock songs electrical guitar sound is more prevalent. Acousticness is calculated from timbre value obtained from the json file generated by the Echonest API. It is calculated over all the segments of the song generated by the API. Timbre is basically used for differentiating different musical instruments from each other [8]. More the value near to the numerical one means more the natural sound is found in the song. The idea behind using acousticness is to differentiate Classical or Sufi songs with that of Rock or Rap songs as Classical and Sufi songs are recorded via natural sound of drum, sticks, sitar, harmonium etc.

F. Valence

In [7] mood of the song plays a good role in the classification of songs. To make our model more accurate valence i.e., mood of the song is also derived. It determines the positivity displayed by the song i.e., cheerful, happiness, sad, depressed or angry. While calculating these factors as stated in [7] certain assumptions are made. The assumptions that a high energy frame represent a cheerful or brightness while a low energy frame is supposed to be a quite frame. With these assumptions and taking the log spectrum sum of the power spectrum the value determined as in [7] is used to determine the valence of a song.

G. Danceability

Danceability is defined as the ease with which a person can dance on that particular song. It is very difficult for someone to dance on a sad song as compared to cheerful Rock song. It is much easier for a person to dance on Rock or Pop song as compared to the Sufi song. Here this value is obtained by using the json or xml file generated by using Echonest API. It is mentioned that danceability can be calculated as measure of tempo, beat strength or many more parameters [10]. Here also, more the value nearer to one the more easily for a person to dance on it. And it is well known that to dance on a classical song is much more difficult than on a pop song, especially in South-East Asian countries.

TABLE I. FEATURE VALUES OF A CLASSICAL AND A ROCK SONG

Features	Classical	Rock
Tempo	93.487	113.043
Loudness	-12.23	-6.403
Energy	0.470893641593	0.821718402183
Speechiness	0.0553177103039	0.0497143161169
Acousticness	0.709522213953	0.155634080369
Valence	0.68593790079	0.61588516209
Danceability	0.408699371909	0.758747465619

H. Discrete Wavelet Transform (DWT)

DWT is used to recognize various hidden patterns which are not seen by the naked eye. There are many wavelet transforms but here,only two i.e., Haar and Daubechies are used. The Haar wavelet's mother wavelet function ψ(t) can be described as

$$\psi(t) = \begin{cases} 1 & 0 \le t < \frac{1}{2}, \\ -1 & \frac{1}{2} \le t < 1, \\ 0 & otherwise \end{cases} \qquad (4)$$

While, Daubechies wavelet function is

$$c_i = g_0 s_{2i} + g_1 s_{2i+1} + g_2 s_{2i+2} + g_3 s_{2i+3} \quad (5)$$

where, g_0,g_1,g_2 and g_3 are coefficients and s_{2i+k} are the data values.

Like [4], the various coefficients derived are partitioned into six segments. On every partition mean and variance are calculate for both the wavelet transforms. Thus, generating a total of twenty four values which will be fed into the neural network as input along with the other above mentioned calculated feature values. It may seem that calculating this feature is not at all good idea. But after having a view of these coefficients of songs of different genres gives a clear view of the importance of the DWT Coefficients because songs of a particular genre have nearly have same values.

The values determined using some of the above feature is given in Table 1 for a clear idea about the same. The comparison is done on a classical versus a rock song.

III. SYSTEM ARCHITECTURE

The system architecture used here is the Parallel Multi-Layered Perceptron. It is works on Back Propagation Algorithm with momentum. A supervised learning is performed using sigmoidal activation function. The neural network model used here is based on the Parallel Multi-Layer Perceptron model [4]. This model can be said to be working similar to our brain. Like for the visual interpretation visual cortex is there and for audio, auditory cortex is there. Both work differently but in parallel. Similarly, in this model different cortexes are achieved based on number of experiments and observations. Because of which last 24 inputs are dealt separately. Likewise for beats, loudness and other features. Like the conjunction of visual and auditory in brain, here also it can be seen in layer 3 and layer 4 of the neural network model. The neural network model used is shown in Fig. 1.This model is contructed using Neuroph Studio. Neuroph Studio provides a Netbeans like structure allowing to create dynamic Neural Network links according to the need. It

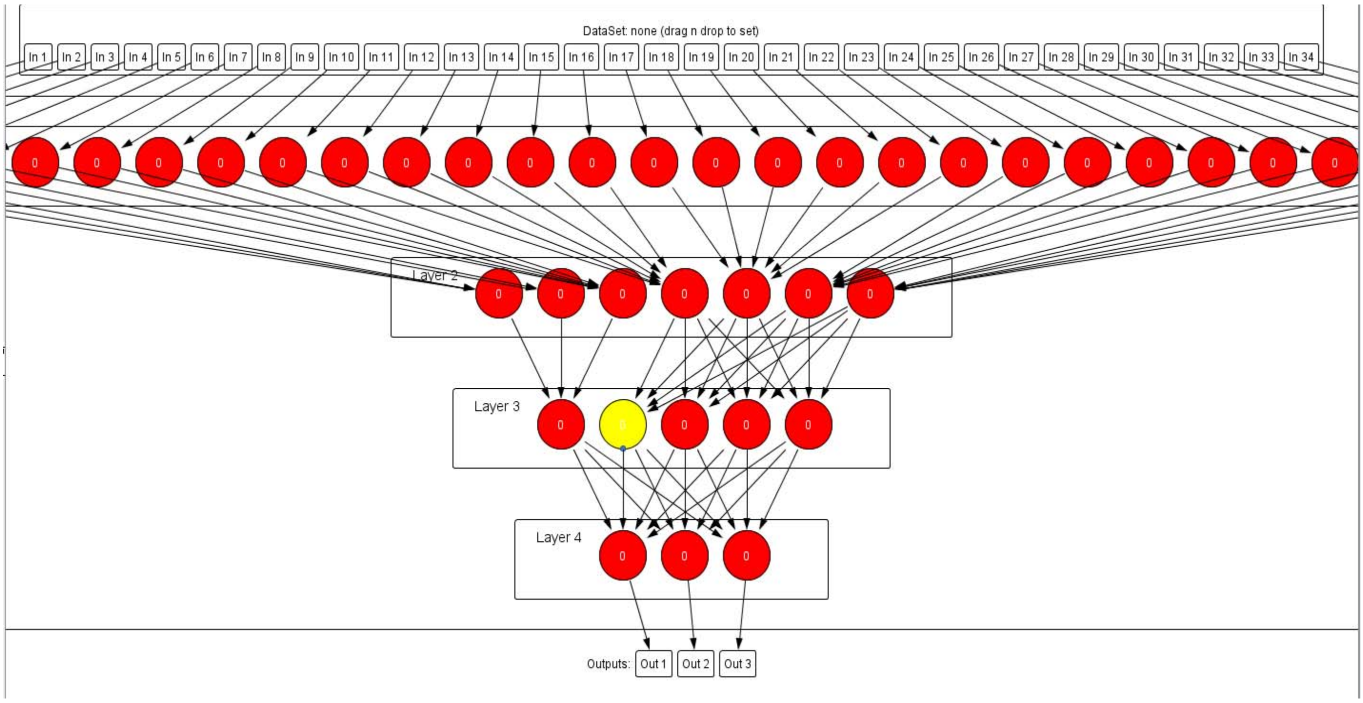

Fig. 1 Neural Network Architecture

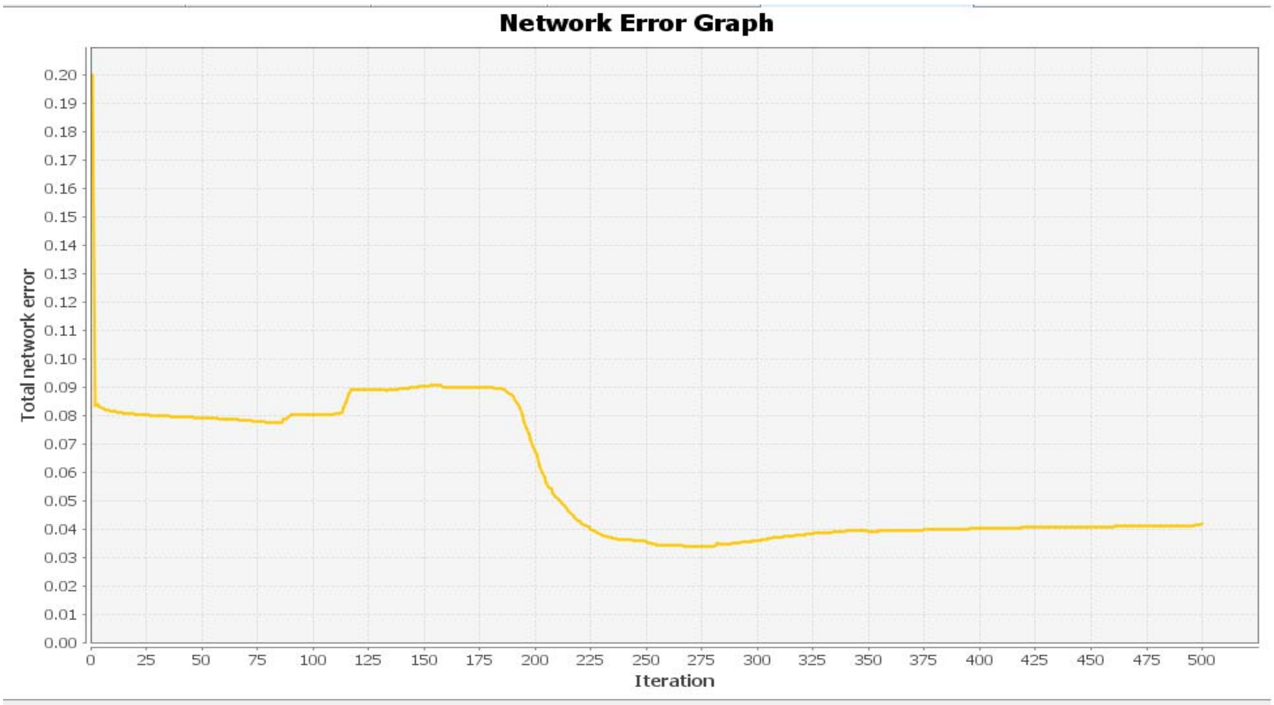

Fig. 2 Network Error Graph

uses JAVA Neuroph API to implement Neural Network functioning. Here are 34 input vectors (three for beats i.e., bpm, beats variance and beats skewness, then loudness mean and average, energy, speechiness, acousticness, valence, danceavility, twenty four from DWT Haar and Daubechies mean and variance. With this different architecture every type of feature is evaluated seperately according to its area. Like DWT coefficient at extreme right are evaluated in disjunction with other inputs so that they update their weights according to there area thus not affecvting the weights of other input areas if there is some robust input even. As supervised learning is adopted the output of this neural network is compared to targeted ouput and then according to it, weights are updated using Back-Propagation Algorithm.

IV. Results

The results are evaluated for two genres of songs that are Classical and Sufi songs. The database consists of 200 songs each of Classical and Sufi. 60% of the songs are used as training set while the rest are used for testing. The database is constructed upon the Indian songs. Echonest library is used for extraction of some features like acousticness, danceability, valence etc. The Echonest library return the URL of a json page which contains the basic information derived from every segment of the songs which are timbre, beats and loudness. From this basic info a number of features are evaluated that are bpm, variance, skewness, loudness etc. Using this, the values are extracted for the each song and are fed to the neural architecture created in Neuroph Studio as shown in Fig. 1. The results are evaluated using a learning rate of 0.1 and momentum as 0.5. The network error graph is shown in Fig 2. The graph is drawn for nearly 500 iterations and using training set data. This converges to an error of 0.04 with minimum error of nearly 0.035 around the 250th iteration.

After that the testing is performed using the test data which accounts for 40% of our dataset and results are evaluated with an accuracy of 85%.The accuracy of Classical songs is 87% while that of Sufi songs is 82%.

Comparing the results with the previous work, the accuracy acquired is quite high as compared to the work done in [1], [5] ,[7] and [6] but nearly same as in [4]. But as compared to [4] the feature set has been greatly reduced from 174 vectors to just 34 which clearly indicate that lesser amount of time, cost, effort and energy would be used to build this system.

V. Conclusion

The propose work uses a total of eight features and applied upon the two genres that are Classical and Sufi songs. The results obtained with an overall accuracy of 85%. The structure designed can be implemented on distributed or multi-processor environment to determine the results faster i.e., on real or run time only. A small in decrease in the accuracy can be because of the fact that Indian Sufi songs are much like Classical songs because of the similar use of natural songs and speechiness patterns. This work can be successfully implemented for other genres around the world like Jazz, Rock, Pop, Rap, etc. It can also be used be used in the music selling websites allowing them to reduce their overhead of classifying and increasing the success rate among their customer because manual work may still have confusion among the operator about the different genres and also require them to have prior knowledge about it. With this no prior knowledge is need. This work can also be extended for songs which are a mixture of two genres like Jazz-Rock, Rock-Rap,Classical-Rock etc.

Acknowledgment

The motivation of work is given by Mr. Sarfaraz Masood, Assistant Professor, D/o Computer Engg, F/o Engg. &

Technology, Jamia Millia Islamia, New Delhi, India. We also like to thank again Mr. Sarfaraz Masood and Dr. Bashir Alam, Assistant Professor, D/o Computer Engg, F/oEngg. & Technology, Jamia Millia Islamia, New Delhi, India for guiding us from the start to the successful completion of our project with this unique innovation which can help the music industry in a great way in future.

REFERENCES

[1] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques," JOURNAL OF COMPUTING, VOLUME 2, ISSUE 3, MARCH 2010, ISSN 2151-9617.

[2] Matityaho, Benyamin, and Miriam Furst. "Neural network based model for classification of music type." Electrical and Electronics Engineers in Israel, 1995., Eighteenth Convention of. IEEE, 1995.

[3] Koerich, Alessandro Lameiras, and Cleverson Poitevin. "Combination of homogeneous classifiers for musical genre classification." Systems, Man and Cybernetics, 2005 IEEE International Conference on. Vol. 1. IEEE, 2005.

[4] Vikramjit Mitra and Chia J. Wang, "A Neural Network based Audio Content Classification", Proceedings of International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007.

[5] Silla, C. N., Alessandro L. Koerich, and Celso AA Kaestner. "Feature selection in automatic music genre classification", Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on. IEEE, 2008.

[6] Fu, Zhouyu, et al. "Learning naive Bayes classifiers for music classification and retrieval", Pattern Recognition (ICPR), 2010 20th International Conference on. IEEE, 2010.

[7] Miyoshi, Masato, et al. "Feature selection method for music mood score detection", Modeling, Simulation and Applied Optimization (ICMSAO), 2011 4th International Conference on. IEEE, 2011.

[8] Tristan Jehan, "Analyzer Documentation, Echonest API", http://developer.echonest.com/docs/v4/_static/AnalyzeDocumentation.pdf

[9] AldebaroKlautau, "The MFCC", 11/22/05, http://www.cic.unb.br/ ~lamar/ te073/Aulas/mfcc.pdf

[10] Acoustic Attribute Overview, Echonest API Documentation, http://developer.echonest.com/acoustic-attributes.html