International Journal of Hybrid Information Technology
Vol. 10, No.5 (2017), pp. 29-40
http://dx.doi.org/10.14257/ijhit.2017.10.5.03

MFCC, Spectral and Temporal Feature based Emotion Identification in Songs

Sarfaraz Masood[1*], Jeevan Singh Nayal[2], Ravi Kumar Jain[3], M. N. Doja[4] and Musheer Ahmad[5]

[1,2,3,4,5]Department of Computer Engineering, Jamia Millia Islamia
New Delhi, India
[1]smasood@jmi.ac.in, [2]jeevan.capricorn@gmail.com, [3]ravi92.jmi@gmail.com, [4]mdoja@jmi.ac.in, [5]musheer.cse@gmail.com

Abstract

This work aims for the solution of one of the challenging yet evolving problem of music information retrieval field i.e. identification of emotions in songs. A collection of four datasets of four separate song sample sizes were selected for the purpose of experiments. For each experiment features such as MFCC, spectral and temporal were extracted for each sample of the dataset. A multilayered sigmoidal feed-forward neural network was trained for construction of a model by using error back propagation algorithm. This helped in recognition of four emotion categories (sad, happy, peaceful and angry) from the song samples. The results obtained at the end of these experiments strongly suggest that this trained model was successfully able to identify the emotions in the selected song samples with an average class accuracy of 88.65%.

***Keywords**: Artificial Neural Network; Back propagation training, Thayer's Emotion Model; MFCC; Temporal Features, Spectral Features*

1. Introduction

In the modern day's technologically advanced age, there has been a significant expansion in digital music data at a fast pace, hence posing a challenge for the listeners to find a song from vast libraries for a particular context. Due to these of this challenge many songs are ignored as users tend to listen to a specific subset of songs which they prefer. To remove this anomaly music recommender systems based on particular context are required which can generate playlist based on mood, singer, genre *etc.*

Emotion identification from songs is one of the challenging fields in the music information retrieval (MIR) domain with vast applications. It can be defined as identifying the dominant mood expressed in a particular sound sample. Last.fm [15] is a popular platform which provides classification of songs based on emotions but it depends on manual annotation of emotions of songs given by each user. It categorizes a song in a particular category on the basis of the number of votes given by users to it. Sometimes, this also leads to false classification of songs as different listeners can perceive a song's emotions in a different way. On contrary to this, work presented in this paper is an automatic identification system of emotions which attempts to recognize dominant emotion of a particular song based on its musical features.

The next section of this paper discusses about the various research works which have been done in this field of mood or emotion classification of songs, which is followed by brief account of the problem statement. In the third section the various musical features selected for this work have been explained. This is followed by the detailed description of the design of this work in the fourth section. The last section shows the results achieved which is followed by the conclusion of this work.

ISSN: 1738-9968 IJHIT

2. Literature Review

Chien *et al.* [1] worked on the categorization of emotions in both song and speech collected through online resources. For the song samples, they used features like tempo, rhythm regularity and normalized intensity mean. Whereas, for the speech samples the fundamental features like pause frequency and the zero crossing rate were used. The song samples were divided into the main part and the refrain part and then performed classification on them separately using Gaussian mixture model. They were able to achieve accuracies of 55%, 60% and 80% for songs in the main part, refrain part and speech signals respectively. This work of emotion classification was performed only on 40 songs collected online.

Bin Zhu *et al.* [2] worked on emotion identification in music samples using a neurogenetic approach. A total of seven emotion related features were extracted for all the music samples. The dataset consisted of a total 203 samples of duration of 20-50 seconds each. They were able to get the identification accuracy of 83.33%.

Samira and Sameti [3] worked on emotion identification in music samples using support vector machines at two levels. Initially various features were extracted to describe each song and then they used an SVM classifier at two levels for training and testing. At level 1, they attempted to classify more distinctive features such as happy and angry. Whereas at level 2, some less distinct features like sad and peaceful were separated. In this work a significantly large set of 176 features were extracted from song samples for the purpose of classification and the dataset had only 280 samples.

Aathreya S. Bhat *et al.* [4] worked on identification of emotions in western and Hindi songs using multi-layered artificial neural networks. Various classes of audio features such as intensity, timbre, rhythm and pitch, except MFCC, were extracted. This work was performed on a relatively smaller sample set of 100 songs with duration of 45 seconds for each sample.

The detailed review of literature in [1] - [4] and [10]-[13] shows that most of the works were performed using a small dataset and also most of the works extracted a large number of features for performing classification. In [3], as much as 176 features were extracted making the system largely complex. In [2] and [4] large song sample durations were 45 and 50 seconds respectively. In this paper, training and testing is performed on larger dataset and the song samples, without removing their background music, were classified with correct emotions. This work shows comparatively better results for 30 seconds sample data. This work builds a relatively simpler system for classification of four emotion categories *i.e. Angry, Happy, Peaceful and Sad* based on the MFCC and spectral temporal features and which is based on training in multi-layered feed forward artificial neural network with backpropagation training algorithm.

3. Features Selection

A total of thirteen features were considered for this work among various spectral and temporal features described below.

MFCC: Mel-frequency cepstrum coefficients (MFCCs) create a mel-frequency cepstrum by which power spectrum of a sound signal may be represented. These coefficients provide the information of spectral shape of the sound sample which is similar to the scale of human hearing. A total of thirteen coefficients were evaluated by using mirmfcc function of the mirtoolbox [9] for the given sound samples.

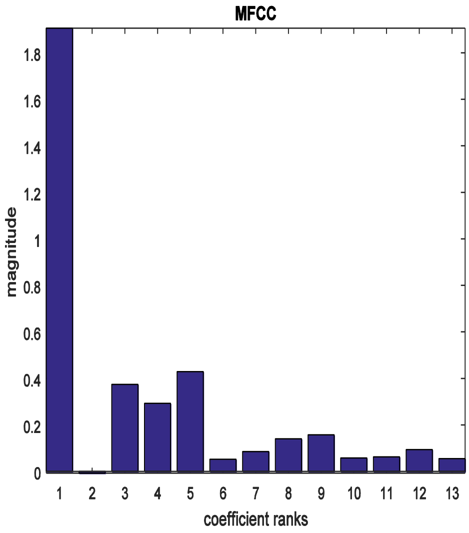


Figure 1a. MFCC for 'Angry' Sample

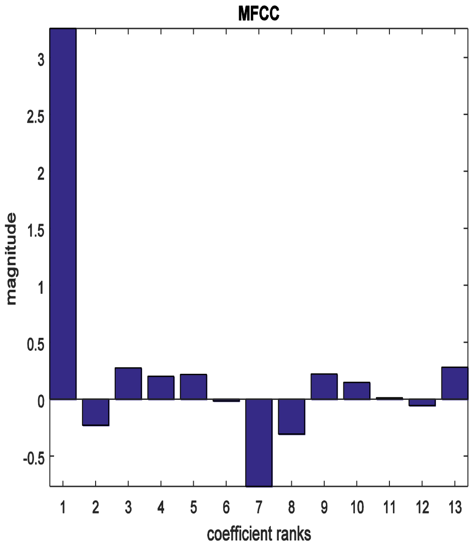


Figure 1b. MFCC for 'Sad' Sample

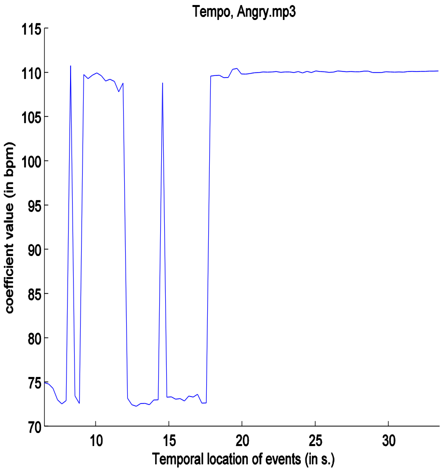
RMS Energy: Root mean square energy (x_{rms}) of any given sound sample is calculated as the square root of the average of the square of its amplitude (x_i) at any given ith frame. This average is for the total of '*n*' frames of the sound sample. The default size of a frame is of 50 milliseconds [9].

$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2+x_2^2+\cdots+x_n^2}{n}} \qquad (1)$$

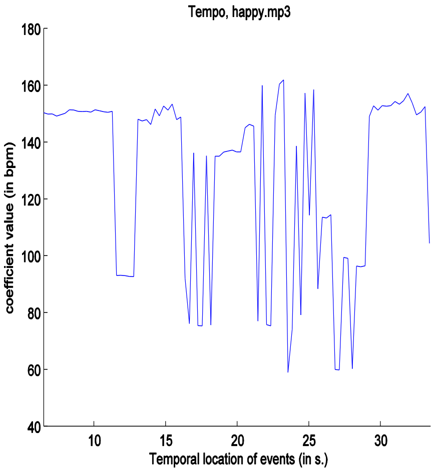
***Mode*:** It is the estimate of the modality of a sound signal which can be either major or minor. The terms Major and Minor describe the musical composition, chord, key, and scale of the sound sample. Its numerical value lies in the range of -1 and +1.

***Tempo*:** Tempo, also known as beats per minute, describes the pace or speed of a given sound signal. It can be evaluated by identifying periodicities using the onset detection curve. Tempo may be defined as follows:

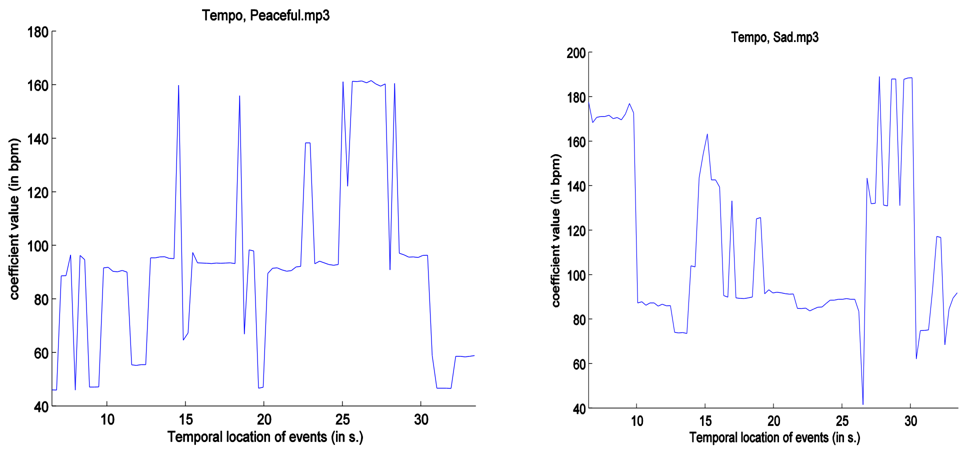
$$Tempo = \frac{Number\ of\ beats\ in\ song}{Duration\ of\ the\ song} \qquad (2)$$



(a)



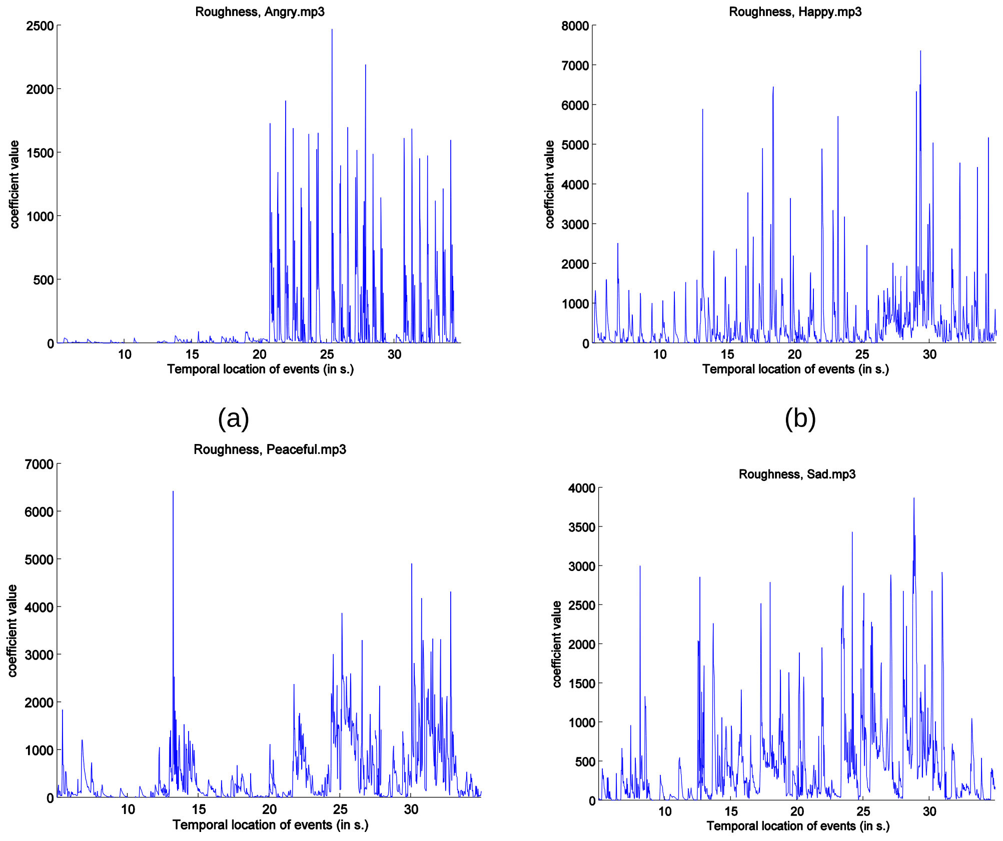
(b)



(c) (d)

Figure 2(a)-(d). Tempo for Angry, Happy, Peaceful and Sad Samples

Roughness: Roughness of a sound sample is estimated using the beating phenomenon where the pair of sinusoids is closed in a frequency. The value of roughness is calculated by computing the mean of dissonance between all pair of peaks of the spectrum [9].

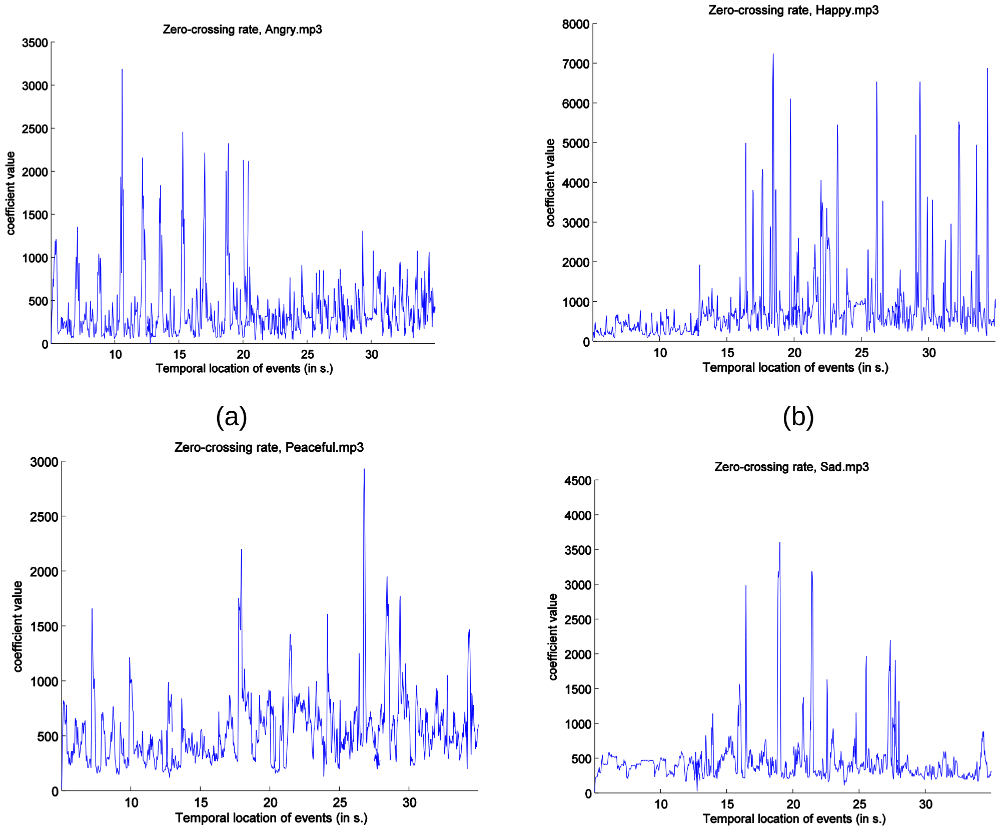


(a) (b)

(c) (d)

Figure 3(a)-(d). Roughness for Angry, Happy, Peaceful and Sad Samples

Zero Crossing Rate: Zero crossing rate is calculated as the number of times the X-Axis is crossed by the sound signal *i.e*. changes in sign of the signal per second. It chiefly indicates the sound sample's noisiness.



(a) (b)

(c) (d)

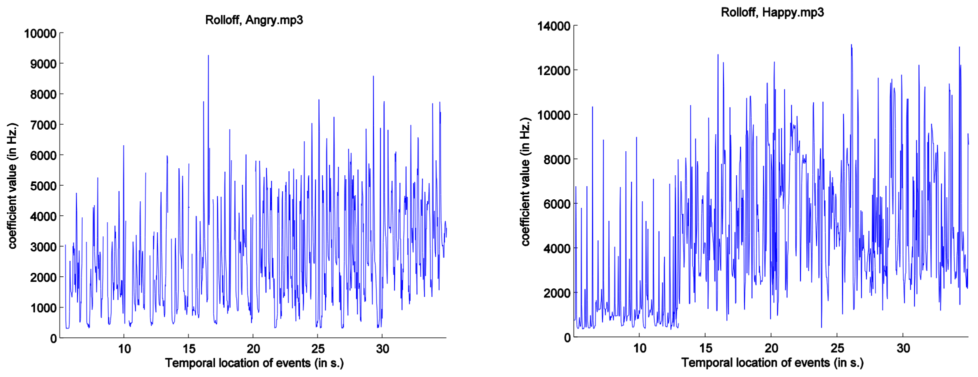
Figure 4(a)-(d). Zero Crossing rate for Angry, Happy, Peaceful & Sad samples

Attack Time: It is the time based duration of the attack phase. It is a subjective measure of time instant at which rhythmic emphasis is observed in a sound.

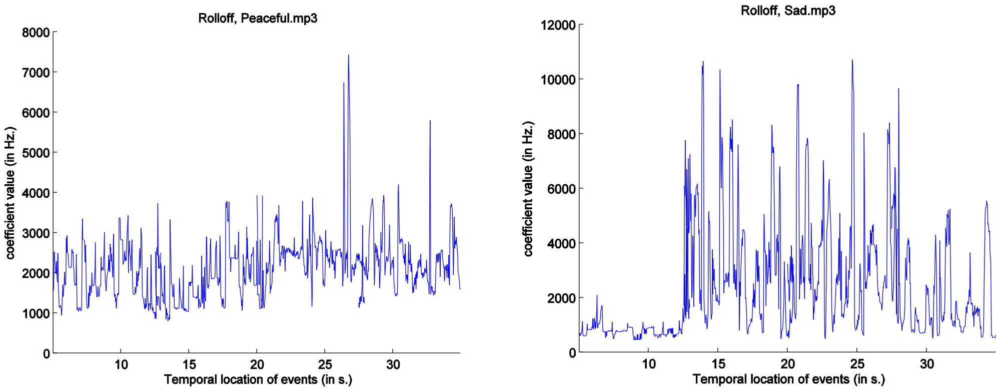
Fluctuation: Fluctuation of a sound sample indicates the rhythmic periodicities along the different auditory channels.

Inharmonicity : Inharmonicity can be defined as a measure of the amount of partials that are not multiples of the fundamental frequency.

Spectral Roll Off: Spectral roll off is the frequency below which some fraction of energy of the sound is contained. Its default value is set to 0.85.



(a) (b)



(c) (d)

Figure 5(a)-(d). Spectral Roll off for Angry, Happy, Peaceful and Sad samples

Low Energy Rate: Low energy rate is the percentage of frames having energy less than average energy in a sound signal. It can be used for assessment of temporal distribution of energy to get the idea whether the energy of all frames is constant throughout or some frames differ from others.

Event Density: Event density calculates the mean frequency of events *i.e.* how many notes onsets per second in the sound sample.

Spectral Centroid: It is used to characterize a spectrum and is defined as the geometric center of a sound distribution obtained by taking mean of weighted frequencies.

4. Experiment Design

A. Dataset Collection

For the purpose of this work, a list of 1000 song samples was selected from *Last.fm* [15] and then separate 10, 15, 20 and 30 seconds samples of those songs were obtained from [16] to create four separate datasets**.** These separate datasets were used to analyze and understand the effect of sample length on the emotion classification accuracy. At [15], the samples are tagged for emotions by their listeners. A song which is tagged by most of the listeners for a particular emotion was assigned that particular emotion class.

A total of 250 song samples were chosen for each of the four emotion categories *i.e. Angry, Happy, Peaceful and Sad*. Also in this work, a song's emotion was identified in sample without removing the background rhythm, beats and its tone as they also help in identifying emotions.

B. Emotions Selection

A total of 4 emotion labels of Thayer's 2D model [6] of arousal and space were chosen as most previous works had been done using them as the dominant classes.

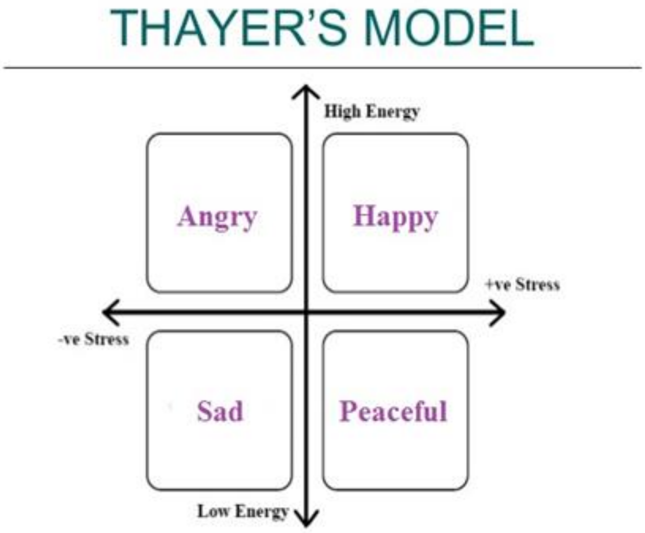


Figure 6. Thayer's 2-D Emotion Model

Instead of calculating the arousal and valence values separately for the samples, the four quadrants were selected as emotion classes as shown in Figure 6.

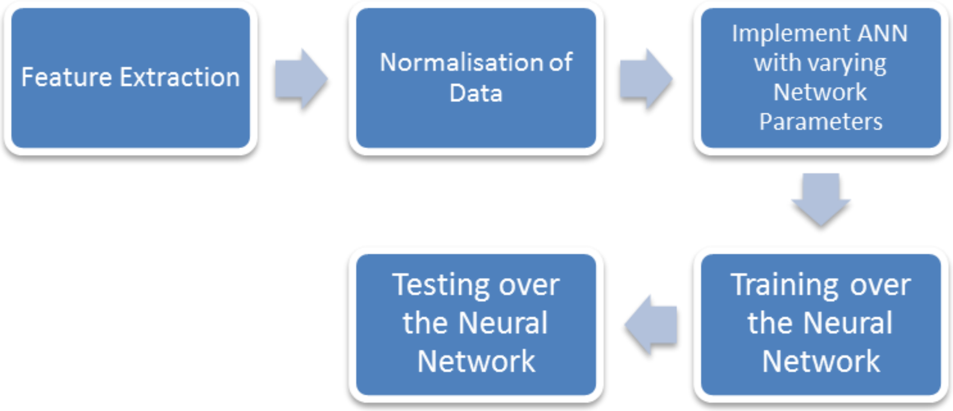

Feature Extraction
Normalisation of Data
Implement ANN with varying Network Parameters
Training over the Neural Network
Testing over the Neural Network

Figure 7. Proposed Framework of the Solution

The selected features were extracted from each song sample for each constructed dataset using MIR toolbox [9]. This toolbox has also been used in various works such as [7] and [8] and known to be very useful in classification problems for audio samples. Some preprocessing steps such as min-max normalization were also performed before the proceeding to the training phase. The proposed framework followed for the solution to the selected problem is shown in Figure 7.

A multi-layered feed forward ANN with a one hidden layer was implemented using Neural Network Toolbox in Matlab [14]. The constructed datasets were divided into a ratio of 80:20 for constructing separate training and the testing sets. For identification of suitable values of some network parameters their values were varied as shown in Table 1 for the exploration of suitable network architecture.

Table 1. Neural Network Parameters for Each Experiment

Parameter	ExpI	Exp II	Exp III	Exp IV
LearningRate	0.3	0.3	0.2	0.2
Epochs	450	450	450	450
HiddenNeurons	10	17	10	17
TrainingFunction	'trainlm'	'trainlm'	'trainscg'	'trainscg'
ActivationFunction	Sigmoidal	Sigmoidal	Sigmoidal	Sigmoidal

5. Results

The True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN), Accuracy, Recall, Precision and F-measure for each of the selected classes are shown in Tables 2, 3, 4 and 5. The values of accuracy obtained are highlighted in each table for easy comparison.

Table 2. Results Obtained on Varying Network Parameters for 10 Seconds Sample Dataset

	EXPERIMENT 1				EXPERIMENT 2				EXPERIMENT 3				EXPERIMENT 4			
	H	*A*	*S*	*P*	*H*	*A*	*S*	*P*	*H*	*A*	*S*	*P*	*H*	*A*	*S*	*P*
TP	42	25	34	43	42	29	33	40	40	29	32	42	38	31	33	46
TN	102	119	110	101	102	115	111	104	103	114	111	101	110	117	115	102
FP	16	13	8	18	16	15	10	14	12	16	9	19	11	14	5	21
FN	11	25	10	9	11	21	11	12	13	21	12	10	15	19	11	6
AC	***84.2***	***79.1***	***88.9***	***84.2***	***84.2***	***80.0***	***87.3***	***84.7***	***85.1***	***79.4***	***87.2***	***83.1***	***85.1***	***81.8***	***90.2***	***84.6***
PR	72.4	65.8	81.0	70.5	72.4	65.9	76.7	74.1	76.9	64.4	78.0	68.9	77.6	68.9	86.8	68.7
RC	79.2	50.0	77.3	82.7	79.2	58.0	75.0	76.9	75.5	58.0	72.7	80.8	71.7	62.0	75.0	88.5
FM	75.7	56.8	79.1	76.1	75.7	61.7	75.9	75.5	76.2	61.1	75.3	74.3	74.5	65.3	80.5	77.3

Table 3. Results Obtained on Varying Network Parameters for 15 Seconds Sample Dataset

	EXPERIMENT 1				EXPERIMENT 2				EXPERIMENT 3				EXPERIMENT 4			
	H	*A*	*S*	*P*	*H*	*A*	*S*	*P*	*H*	*A*	*S*	*P*	*H*	*A*	*S*	*P*
TP	43	42	27	32	46	41	29	33	32	27	40	45	30	27	43	46
TN	101	102	117	112	103	108	120	116	112	117	104	99	116	119	103	100
FP	18	11	18	8	18	10	15	7	7	18	10	20	7	20	11	15
FN	11	12	15	17	8	13	13	16	17	15	14	9	19	15	11	8

AC	***83.2***	***86.2***	***81.4***	***85.2***	***85.1***	***86.6***	***84.2***	***86.6***	***85.7***	***81.4***	***85.7***	***83.2***	***84.9***	***80.7***	***86.9***	***86.4***
PR	70.5	79.2	60.0	80.0	71.9	80.4	65.9	82.5	82.1	60.0	80.0	69.2	81.1	57.4	79.6	75.4
RC	79.6	77.8	64.3	65.3	85.2	75.9	69.0	67.4	65.3	64.3	74.1	83.3	61.2	64.3	79.6	85.2
FM	74.8	78.5	62.0	71.9	77.9	78.1	67.4	74.2	72.7	62.1	76.9	75.6	69.8	60.7	79.6	80.0

Table 4. Results Obtained on Varying Network Parameters for 20 Seconds Sample Dataset

	EXPERIMENT 1				**EXPERIMENT 2**				**EXPERIMENT 3**				**EXPERIMENT 4**			
	H	*A*	*S*	*P*	*H*	*A*	*S*	*P*	*H*	*A*	*S*	*P*	*H*	*A*	*S*	*P*
TP	35	44	32	32	37	41	24	44	33	34	43	35	30	36	43	38
TN	108	99	111	111	109	105	122	102	112	111	102	110	117	111	104	109
FP	13	12	23	8	13	8	9	23	7	21	13	13	5	24	9	14
FN	12	7	16	21	10	10	24	9	20	14	8	12	23	12	8	9
AC	***85.1***	***88.3***	***78.6***	***83.1***	***86.4***	***89.0***	***81.6***	***82.0***	***84.3***	***80.6***	***87.3***	***85.3***	***84.0***	***80.3***	***89.6***	***86.5***
PR	72.9	78.6	58.2	80.0	74.0	83.7	72.7	65.7	82.5	61.8	76.8	72.9	85.7	60.0	82.7	73.1
RC	74.5	86.3	66.7	60.4	78.7	80.4	50.0	83.0	62.3	70.8	84.3	74.5	56.6	75.0	84.3	80.9
FM	73.7	82.2	62.1	68.8	76.3	82.0	59.3	73.3	71.0	66.0	80.4	73.7	68.2	66.7	83.5	76.8

Table 5. Results Obtained on Varying Network Parameters for 30 Seconds Sample Dataset

	EXPERIMENT 1				**EXPERIMENT 2**				**EXPERIMENT 3**				**EXPERIMENT 4**			
	H	*A*	*S*	*P*	*H*	*A*	*S*	*P*	*H*	*A*	*S*	*P*	*H*	*A*	*S*	*P*
TP	42	42	24	44	35	44	33	40	33	34	43	40	40	34	42	42
TN	110	110	128	108	117	108	119	112	117	116	107	110	118	124	116	116
FP	17	4	8	18	14	11	15	7	5	20	7	17	6	14	4	17
FN	7	10	21	9	14	8	12	13	20	11	9	9	13	11	10	7
AC	***86.4***	***91.6***	***83.9***	***84.9***	***84.4***	***88.9***	***84.9***	***88.4***	***85.7***	***82.9***	***90.4***	***85.2***	***89.3***	***86.3***	***91.9***	***86.8***
PR	71.2	91.3	75.0	70.9	71.4	80.0	68.8	85.1	86.8	63.0	86.0	70.2	87.0	70.8	91.3	71.2
RC	85.9	80.8	53.3	83.0	71.4	84.6	73.3	75.5	62.3	75.6	82.7	81.6	75.5	75.6	80.8	85.7
FM	77.8	85.7	62.3	76.5	71.4	82.2	70.9	80.0	72.5	68.7	84.3	75.5	80.8	73.1	85.7	77.8

The cumulative comparison of values in the Tables 2,3,4 and 5 shows that 30 seconds samples help in achieving much higher class accuracies than 10, 15 and 20 second samples. Also, the neural network performs better training when the number hidden nodes are selected as 17 and the scaled conjugate gradient training function is used for learning. The best results were obtained in the Experiment 4 of the 30 second sample dataset with the best average accuracy of all the classes as 88.65%. A comparison of accuracies obtained across the varying network architectures and different song sample size datasets of 15, 20 and 30 seconds is shown in Figure 8.

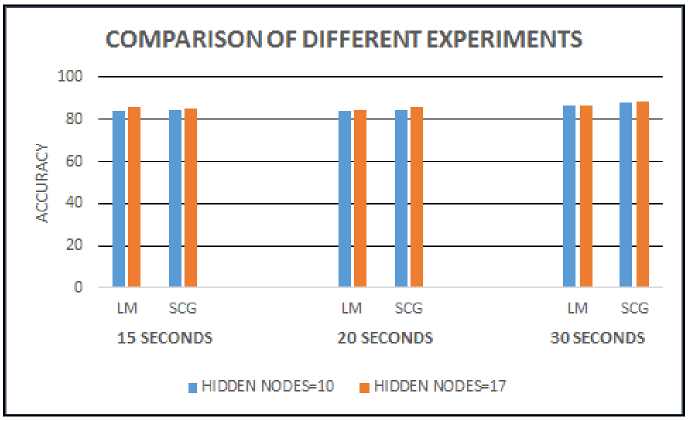


Figure 8. Comparison of Experiments on Different Datasets

These results show that a significant identification rate of 88.65**%** was achieved using 25 features, including 13 MFC Coefficients, 12 temporal and spectral features, over a comparatively larger dataset of 1000 samples. Whereas in [3] as many as 176 features were used to achieve 87.27% accuracy. Also [1] and [2] used 40 and 203 songs respectively and obtained accuracies of 80% and 83.3% respectively, which are lesser than obtained in this work. In [4], only 100 song samples were selected of 45 seconds duration each and 85.75% accuracy was obtained.

6. Conclusion

In this paper, a significant attempt is made to provide a solution to the problem of emotion identification in songs by extracting MFCC, temporal and spectral features and constructing an artificial neural network as the classifier, with an average class accuracy of 88.65%. These results strongly suggest that the selected MFCC, spectral features and temporal features play an important role in recognizing emotion content of songs even without removal of background music.

These results obtained were better as compared to results achieved in recent works attempted in the same area and considered the largest set song samples as well. Also the number features identified were the minimum as compared to other works. A logical extension to this work can be consideration of a larger set of similar emotions which can be challenging to classify.

References

[1] C. H. Chen, P. T. Lu and O. T. C. Chen, "Classification of four affective modes in online songs and speeches", The 19th Annual Wireless and Optical Communications Conference (WOCC 2010), Shanghai, **(2010)**, pp. 1-4. doi: 10.1109/WOCC.2010.5510629

[2] Bin Zhu and K. Zhang, "Music emotion recognition system based on improved GA-BP," Computer Design and Applications (ICCDA), 2010 International Conference on, Qinhuangdao, **(2010)**, pp. V2-409-V2-412. doi: 10.1109/ICCDA.2010.5541390

[3] S. Pouyanfar and H. Sameti, "Music emotion recognition using two level classification," Intelligent Systems (ICIS), 2014 Iranian Conference on, Bam, **(2014)**, pp. 1-6. doi:10.1109/ IranianCIS.2014. 6802519

[4] S. Bhat, V. S. Amith, N. S. Prasad and D. M. Mohan, "An Efficient Classification Algorithm for Music Mood Detection in Western and Hindi Music Using Audio Feature Extraction", Signal and Image Pro-

cessing (ICSIP), 2014 Fifth International Conference on, JejuIsland, **(2014)**, pp. 359-364. doi: 10.1109/ICSIP.2014.63

[5] Y. H. Chin, P. C. Lin, T. C. Tai and J. C. Wang, "Genre based emotion annotation for music in noisy environment," Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on, Xi'an, **(2015)**, pp. 863-866. doi: 10.1109/ACII.2015.7344675

[6] R. E. Thayer, The Biopsychology of Mood and Arousal, Oxford University Press, New York, **(1989)**.

[7] 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.doi: 10.1109/INDICON.2015.7443497

[8] M. Sheezan, Goel, S. Masood and A. Saleem, "Genre classification of songs using neural network," Computer and Communication Technology (ICCCT), 2014 International Conference on, Allahabad, 2014, pp. 285-289.doi: 10.1109/ICCCT.2014.7001506

[9] MIRtoolbox documentation: https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/ materials/mirtoolbox/ MIRtoolbox1.6.1guide).

[10] Pao, Tsang-Long, "Comparison between weighted d-knn and other classifiers for music emotion recognition." Innovative Computing Information and Control, 2008. ICICIC'08. 3rd International Conference on. IEEE, **(2008)**.

[11] Cheng, Heng-Tze, "Automatic chord recognition for music classification and retrieval." Multimedia and Expo, 2008 IEEE International Conference on. IEEE, **(2008)**.

[12] Ujlambkar, Aniruddha M., and Vahida Z. Attar. "Automatic mood classification model for indian popular music." Modelling Symposium (AMS), Sixth Asia. IEEE, **(2012)**.

[13] Su, Dan, Pascale Fung, and Nicolas Auguin. "Multimodal music emotion classification using AdaBoost with decision stumps." Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. IEEE, **(2013)**.

[14] MATLAB 8.0, The MathWorks, Inc., Natick, Massachusetts, United States

[15] www.last.fm

[16] www.shazam.com