Study of Indian instruments for designing a speech/music classifier

[1]Arvind Kumar, [2]S.S Solanki, [3]Mahesh Chandra
Department of Electronics and Communication Engineering
Birla Institute of Technology, Mesra, Ranchi, India
[1]arvind9835@gmail.com, [2]sssolanki@bitmesra.ac.in, [3]shrotiya69@rediffmail.com,

***Abstract*: - Automatic speech/music classification is the art of labeling incoming audio samples into either of two classes i.e. speech and music, using multimedia signal processing technique. This work focuses on building a system model to classify speech and music segments. Speech samples were obtained from standard S&S database and music samples were obtained by playing different Indian musical instruments. Different state-of-the-art features for audio samples proposed in literature were extracted and variation of classification accuracies of these models built on these feature vectors was observed. Various experiments were conducted with Support Vector Machines (SVM) and Random Forest (RF) Classifiers. Best mean classification accuracy of 99.62% and 99.13% is observed for Mel Frequency Cepstral Coefficients (MFCC) feature with SVM and RF classifier respectively.**

Keyword: Speech/Music Classification, Indian Musical Instrument, SVM, RF

I. INTRODUCTION

Although independently both speech and music signal processing are active areas of research for several past decades, many researchers have also given attention in designing a front-end processor to discriminate speech and music samples. Speech/Music discrimination (SMD) can be used in variety of task such as content monitoring of broadcast information, tagging speech samples for effective speech recognition, automatic tuning of radio station etc. In future, SMD along with music information retrieval (MIR) can be used to provide unique experience to music lovers by generating playlist based on mood, singer and genre.

The first reported work in this field was by Saunders [1]. He utilized energy and zero crossing rate based features to classify speech and music samples. Features like fundamental frequency, average ZC rate and spectral peaks tracks were used with an accuracy of more than 90% for audio classification and 95% for audio segmentation. Although this technique of using ZCR was successful for pure speech and music, performance of this feature for discrimination of music with background music was not equally satisfactory. Scheirer and Slaney [2] explored Power Spectral Density in their work and various features like spectral flux, spectral roll-off and spectral centroid were proposed. Accuracy of 98.2% for 2.4s segments where observed for these features. This feature captures the statistical behavior of audio signal. Alexandre et. al. [3] used Spectral based features along with Mel Frequency Cepstral Coefficients (MFCCs) and high zero crossing rate ratio with Fisher Linear Discriminant classifier and k-Nearest Neighbour.

Most of the reported works in speech/music classification have evaluated their work on either of S&S database, GTZAN database or MUSAN database. Although these databases cover wide range of musical instruments, most of the music samples have western origin. Hence, it was an inspiration to evaluate the performance of existing features for music samples derived from Indian musical instruments. Moreover, observation from this study would also be helpful in studying speech/music segmentation for music concert based on Indian instrument. Fig. 1 displays a live concert where a musician alternatively speaks and plays resulting in both speech and music samples at different instances. Results from this study would be helpful in designing a speech/music classifier for such audio recordings.

Fig.1 Musician during live concert

80 samples from 16 different instruments were collected to match the number of speech samples from S&S database. Out of these, 60 samples were used for training and rest 20 samples were used for testing. Rest of the paper is organized as follows: Section II describes the process of database preparation, Section III highlights steps for feature extraction, Section IV discusses the classifiers in brief and Section V discusses experimental results.

978-1-7281-9180-5/20/$31.00 ©2020 IEEE

II. DATABASE PREPARATION

Experiments are carried out using samples of melody obtained from different musical instruments. The instruments selected for the experiment are Been, Dhol, Dholak, Esraj, Bansuri, Ghatam, Gopichand, Harmonium, Mridangam, Santoor, Sarod, Tabla, Tanpura, Tumbi, Taus and Veena. 60 samples of audio samples were used for training the network and the remaining samples in the database, are used for testing. The samples used are 16-bit mono samples recorded at 16 KHz using Philips SBCMD110/01 corded microphone. Each sample recorded is taken to be of 15 sec duration and recorded using WavePad sound editing software. Most of the samples are collected from musical institutes while some recordings were prepared from "Indian Musical Instrument" App downloaded from playstore. No voice is present in these segments. All samples are recorded in the same acoustic environment. Fig. 1 illustrates the instrument used in this study.

Fig.2 Indian Instruments

III. FEATURE EXTRACTION

Musical Instruments are basically classified into four main categories which are also called families: string, brass, woodwind and percussion. Feature extraction is the process of processing audio signals to generate set of feature sequence. Features extracted in this work is mainly divided into two categories i.e. Temporal Domain and Spectral Domain.

A. Time Domain Features

Two of the most common temporal domain features used in research is short-term energy (STE) and zero-crossing rate (ZCR). STE measures the energy of a signal in short duration and ZCR measure the number of times the signal crosses zero. It gives the rate of sign change of a time-domain waveform. ZCR is an indicator for the noisiness of the signal. It is calculated by the number of times the signal crosses zero within a given window.

$$ZCR_t = \frac{1}{2}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \quad (1)$$

Where,
$(x_{n-1} < 0 \; and \; x_n > 0)$ or $(x_{n-1} > 0 \; and \; x_n < 0)$ or $(x_{n-1} \neq 0 \; and \; x_n = 0)$.

$$STE = \frac{1}{N}\sum_{n=0}^{N-1} x^2(n) \quad (2)$$

Where,
N= Number of samples in a frame x (n)

B. Spectral Domain Features

Over the years, different Short Time Fourier Transform (STFT) based features were proposed. Some of the popular features are spectral centroid, spectral roll-off, spectral-flux, spectral entropy, Chroma vectors and Mel-frequency cepstral coefficients. For a signal x(n) let $X_t(k)$ be the discrete fourier transform (DFT) of t th frame where k=0,1,2, ……., N-1, N being the frame length.

(i) Spectral Centroid: It measures the spectral shape of the audio samples with high values corresponding to "brighter" sounds. Spectral Centroid is also called "centre of gravity" for a spectrum. For a t^{th} frame, Spectral Centroid, C_t is defined as

$$C_t = \frac{\sum_{k=0}^{N-1}(k+1)X_t(k)}{\sum_{k=0}^{N-1}X_t(k)} \quad (3)$$

(ii) Spectral roll-off: It indicates the frequency below which 85-90% of the energy content of the spectrum is concentrated. Fig. 2, illustrate the concept graphically.

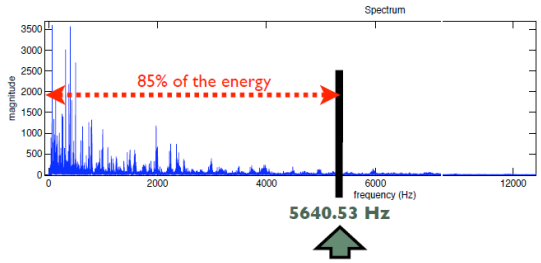


Fig.3 Spectral roll-off (fig. used from MIRTOOLBOX manual)

(iii) Spectral flux: It measures the change of spectral information between two successive frames. It is

978-1-7281-9180-5/20/$31.00 ©2020 IEEE

obtained by finding the squared difference of the normalized spectra of two frames.

$$SF_{t,t-1} = \sum_{k=0}^{N-1}(norm(X_t(k)) - norm(X_{t-1}(k))^2 \quad (4)$$

Where, $X_t(k)$ is kth DFT coefficient of t^{th} frame

(iv) Spectral entropy: It measures the spectral power distribution of a signal. Signal's normalized power distribution in the frequency domain is treated as probablity distribution and its Shannon Entropy is calculated.

$$H = -\sum_{i=0}^{L-1} n_i log_2(n_i)\, \pi r^2 \quad (5)$$

Where, n_i is the normalized spectral energy.

(v) Fundamental frequency: Both speech and music signal are harmonic signal, characterized by fundamental frequency. Fundamental frequency for music signal has comparatively larger variation than speech signal and has been explored for speech/music classification. However, tracking of fundamental frequency for both speech and music signal is a complicated task and numerous methods have been proposed in literature. In this work, mean of fundamental frequency is evaluated for short frames and is used as a feature vector.

(vi) Chroma Vector: It is the distribution of energy of an audio signal in 12 traditional pitch class of equal-tempered scale of western music and is computed using DFT coefficients. Sequence of chroma vectors is known as Chromagram. Fig. 3 illustrates a Chromagram.

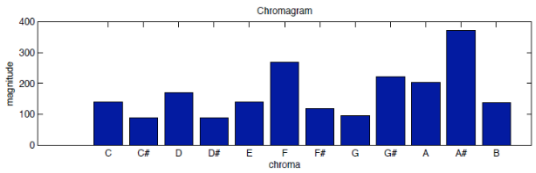


Fig.4 Chromagram of an audio sample

(vii) Mel-frequency cepstral coefficients: MFCC is the most used features for different speech processing application. DFT coefficients are mapped to triangular mel-filter bank to mimic the human ear. Mel Frequency Cepstral coefficients (MFCC) basically describe the spectral shape of an audio input. The MFCC feature is obtained by transforming a signal from frequency (hertz) scale to Mel-scale by using Eq.6.

$$mel(f) = 2595\ \log_{10}(1 + \frac{f}{700}) \quad (6)$$

The Mel-scale has 40 filter channels. The first filter bank measures the power of the signal and the remaining 12 linearly spaced outputs represent the spectral envelope. The harmonics of the signal is presented by the 27 log-spaced channels. Then Discrete Cosine transform is applied which converts the filter outputs to give the MFCCs. First 13 coefficients and their single and double derivative are used for classification.

IV. CLASSIFIERS

Speech/music discrimination is a task of binary classification. Researchers over the years have extensively used support vector machines (SVM) [4, 5, 6, 7, 8] and Gaussian mixtures models (GMM) [5, 6, 9, 10]. Although, in some of the works, Artificial Neural Network (ANN) [11], k-Nearest Neighbours (k-NN) [12], Naïve Bayes [13], Decision tree [11] and Dynamic Time Warping (DTW) [12] based classifiers had also been explored. This study evaluates the performance of models built on SVM and RF Classifier.

A. Support Vector Machines

It is a discriminative classifier which separates two different classes by a hyperplane for a given vector weight *w* and bias *b*. The distance between the closest data points and the hyperplane is called margin of separation. These points which are closest to the hyperplane are called support vectors. The algorithm tries to find a hyperplane which maximizes the margin of separation.

For a set of data with training vectors x_j and their categories y_j in some dimension d where $x \in R^d$ and $y_j = \pm 1$, the equation of hyperplane is

$$f(x) = x'w + b = 0 \quad (7)$$

Where, w is the weight vector and b is a bias.

B. Random Forest

Random Forest classifier works by embedding large number of decision trees. Each individual decision tree in random forest generates a class prediction. A class with maximum number of votes becomes the model prediction. Random Forest has great advantage over other classifiers as it has large number of uncorrelated models working together as a single committee outperforming individual models.

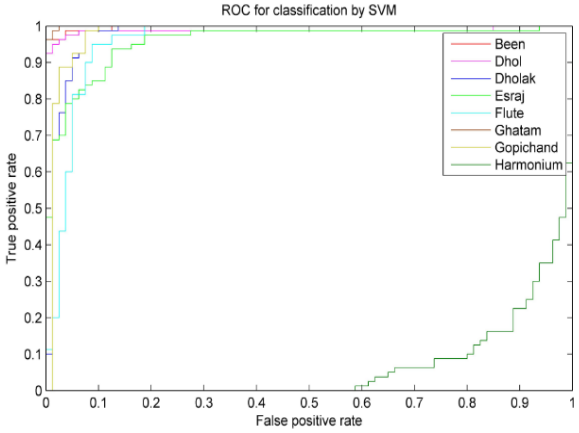
V. EXPERIMENT RESULTS

Under this section, we discuss various experimental works carried in this work. Initially, the efficiency of these features is discussed using ROC and Box plot followed by observation of the classification accuracy using two classifiers.

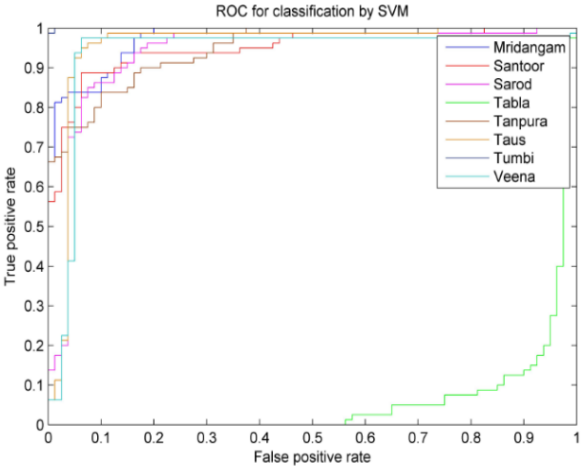
A. Feature Importance

978-1-7281-9180-5/20/$31.00 ©2020 IEEE

Fig. 5 illustrates the ROC plot for ZCR features for SVM classifier using speech sample from S&S dataset and music sample from one instrument at a time. Each of Fig. 5a and 5b illustrate the ROC plot for eight instruments. Curve with higher Area Under the Curve (AUC) indicates higher discriminatory evidence. Ghatam and Taus shows best ROC curve whereas Harmonium and Tabla shows worst performance for ZCR feature. Further, feature importance is studied through Box plot and is illustrated in Fig. 6. Box plot visualize the discriminatory evidence of these features for various Indian instruments. First column in the diagram plots the statistical characteristics of the extracted features for speech signal followed by sixteen Indian instruments. Minimum overlap amongst two classes indicates strong discriminatory evidence whereas maximum overlap indicates weak discriminatory evidence of the features. At the end, classification accuracy of these features is discussed for SVM and RF classifiers.



(a)



(b)

Fig.5 ROC plot for ZCR feature for different Indian instrument

B. Classification Accuracy

Each of the extracted features from different instruments is fed to the classifiers and the classification accuracy is evaluated over ten iterations.

TABLE I : PERFORMANCE OF DIFFERENT FEATURES WITH SVM CLASSIFIER(RBF)

Instruments	**Mean Classification Accuracy (%)**				
	F1	**F2**	**F3**	**F4**	**F5**
Been	100.0	93.75	84.06	80.62	100.0
Dhol	97.81	48.43	93.43	72.50	100.0
Dholak	96.87	52.18	93.75	70.00	99.68
Esraj	99.68	62.18	87.81	75.00	99.68
Flute	100.0	84.06	96.56	75.31	98.12
Ghatam	97.18	52.50	95.31	79.68	99.37
Gopichand	99.37	58.75	99.06	79.37	100.0
Haromonium	100.0	86.56	90.31	70.93	100.0
Mridangam	88.12	55.00	81.56	75.93	100.0
Santoor	99.06	69.37	88.53	79.37	98.75
Sarod	95.62	60.31	88.12	76.56	98.75
Tabla	96.56	64.06	73.43	84.68	100.0
Tanpura	100.0	55.93	91.56	86.25	100.0
Taus	99.37	60.00	92.81	70.93	99.68
Tumbi	99.68	55.62	79.06	82.18	100.0
Veena	97.50	71.25	84.06	71.56	100.0
Average	**97.92**	**64.37**	**88.71**	**76.92**	**99.62**
	F6	F7	F8	F9	
Been	84.68	68.12	91.25	97.50	
Dhol	78.43	65.62	91.56	95.31	
Dholak	83.75	99.06	99.37	92.81	
Esraj	82.18	91.56	85.93	56.26	
Flute	78.12	97.50	94.37	90.00	
Ghatam	71.25	60.31	88.43	98.43	
Gopichand	66.87	98.12	95.31	90.62	
Haromonium	58.75	93.75	94.37	51.56	
Mridangam	62.50	90.93	86.56	86.25	
Santoor	76.56	90.00	83.75	61.87	
Sarod	62.81	91.56	90.00	83.43	
Tabla	50.31	56.87	87.81	49.68	
Tanpura	73.43	82.50	88.12	69.68	
Taus	70.31	91.87	83.43	95.00	
Tumbi	87.18	71.25	90.00	99.06	
Veena	76.56	92.81	88.43	94.06	
Average	**72.73**	**83.86**	**89.91**	**81.97**	

F1=Chromagram, F2=Energy, F3=Spectral Entropy, F4=Spectral Flux, F5=MFCC, F6=Pitch, F7=Spectral Centroid, F8=Spectral Roll-off, F9=ZCR

Table 1 tabulates the mean classification accuracy for SVM classifier using Radial Bias Function (RBF) kernels for classifying speech and music samples from different Indian instruments. MFCC feature shows the best average mean classification accuracy of 99.62% whereas energy feature shows the worst average mean classification accuracy of 64.37%. Table 2 tabulates the average mean classification accuracy for RF classifiers with 50 decision trees. Analogous to the SVM classifier, MFCC has the best average mean classification accuracy of 99.13% and Energy has the worst average mean classification accuracy of 63.90%. Fig. 7 plots the variation of out-of-bag classification error with different number of

978-1-7281-9180-5/20/$31.00 ©2020 IEEE

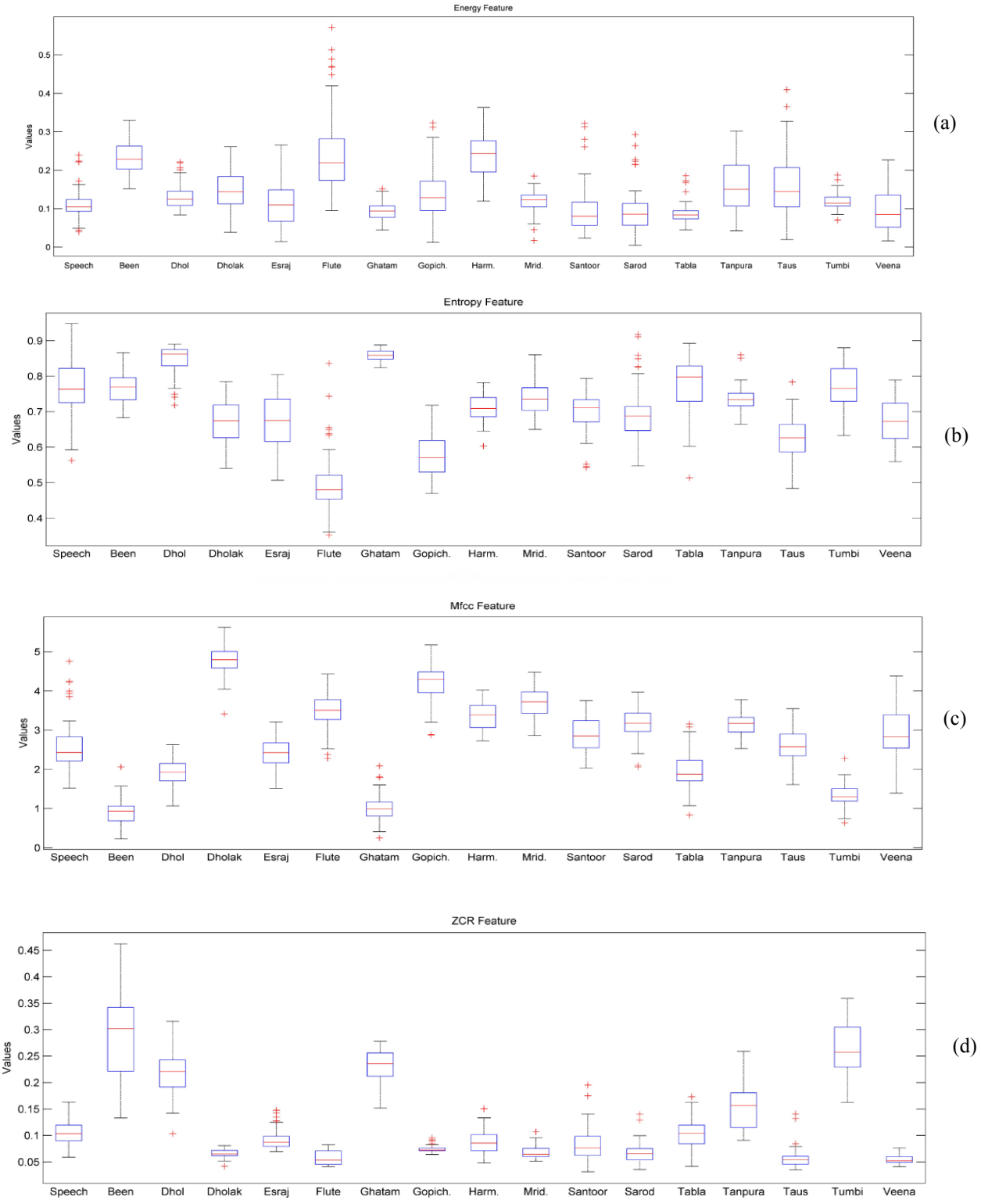


Fig.6 Box Plot for different features for speech and music signals from various Indian instruments

978-1-7281-9180-5/20/$31.00 ©2020 IEEE

decision trees used for ZCR features. Curve for Esraj and Harmonium shows poor convergence indicating low classification efficiency.

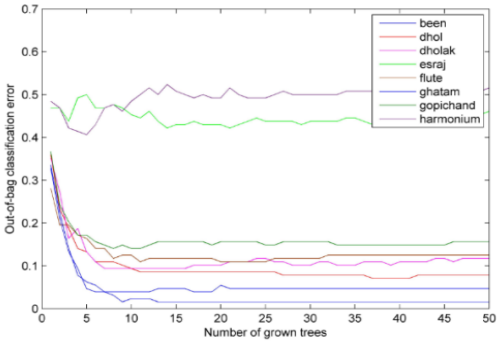


Fig.7 Out-of-bag classification error plot for ZCR feature for different Indian instrument

Table II : Performance of different features for RF Classifier

Instruments	Classification Accuracy (%)				
	F1	**F2**	**F3**	**F4**	**F5**
Been	99.37	90.93	76.56	80.31	100.0
Dhol	95.62	57.18	89.68	79.68	100.0
Dholak	97.50	63.75	90.93	75.62	98.12
Esraj	99.37	54.37	84.06	79.68	99.68
Flute	100.0	75.62	97.50	78.75	96.56
Ghatam	95.31	55.62	95.62	79.68	99.68
Gopichand	99.37	62.18	97.18	86.56	99.37
Haromonium	99.68	82.18	90.62	76.25	99.37
Mridangam	84.37	51.87	80.93	79.06	99.37
Santoor	95.93	59.68	82.50	80.93	98.12
Sarod	93.75	61.25	84.06	81.87	96.56
Tabla	87.50	61.87	74.37	82.81	100.0
Tanpura	100.0	55.00	90.00	90.31	100.0
Taus	98.12	64.68	91.87	74.68	99.37
Tumbi	100.0	57.18	76.87	86.25	100.0
Veena	98.43	69.06	82.81	73.75	100.0
Average	**96.52**	**63.90**	**86.59**	**80.38**	**99.13**
	F6	**F7**	**F8**	**F9**	
Been	77.81	68.75	93.43	97.18	
Dhol	68.12	61.25	87.81	90.62	
Dholak	68.75	99.68	99.06	89.68	
Esraj	74.06	84.68	83.12	56.56	
Flute	81.56	97.50	91.87	86.87	
Ghatam	69.37	55.00	85.00	98.43	
Gopichand	64.68	97.50	94.06	84.06	
Haromonium	52.50	92.18	87.18	51.25	
Mridangam	62.50	91.56	80.31	77.50	
Santoor	66.87	88.43	76.87	63.12	
Sarod	60.62	93.43	86.87	71.25	
Tabla	50.31	60.31	83.43	53.75	
Tanpura	67.50	78.75	90.62	70.62	
Taus	67.81	91.87	78.43	89.68	
Tumbi	86.87	67.18	91.56	98.12	
Veena	72.50	95.62	85.93	87.81	
Average	**68.23**	**82.73**	**87.22**	**79.15**	

F1=Chromagram, F2=Energy, F3=Spectral Entropy, F4=Spectral Flux, F5=MFCC, F6=Pitch, F7=Spectral Centroid, F8=Spectral Roll-off, F9=ZCR

VI CONCLUSION

This study evaluates the efficiency of different temporal and spectral features for designing an effective speech/music classifier for speech samples derived from S&S database and music samples derived from playing Indian musical instruments. Experiments were conducted in combination of 16 musical instruments and 9 spectral and temporal features with SVM and RF classifier. Initially, feature importance is studied through Box-plot and ROC plot. MFCC feature outperforms all other features with an average mean classification accuracy of 99.62% and 99.13% with SVM and RF respectively. Both the classifier shows similar trend in classification accuracy for different features highlighting the importance of feature irrespective of the classifier being used.

REFERENCES

[1] Saunders, John. "Real-time discrimination of broadcast speech/music." In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 2, pp. 993-996. IEEE, 1996.

[2] Scheirer, Eric, and Malcolm Slaney. "Construction and evaluation of a robust multifeature speech/music discriminator." In 1997 IEEE international conference on acoustics, speech, and signal processing, vol. 2, pp. 1331-1334. IEEE, 1997.

[3] Alexandre-Cortizo, Enrique, Manuel Rosa-Zurera, and Francisco Lopez-Ferreras. "Application of fisher linear discriminant analysis to speech/music classification." In EUROCON 2005-The International Conference on" Computer as a Tool", vol. 2, pp. 1666-1669. IEEE, 2005.

[4] Khan, M. Kashif Saeed, and Wasfi G. Al-Khatib. "Machine-learning based classification of speech and music." Multimedia Systems 12, no. 1 (2006): 55-67.

[5] Khonglah, B.K. and Prasanna, S.M., Speech/music classification using speech-specific features. Digital Signal Processing, 48, pp.71-83, 2016

[6] Zhang, Hao, Xu-Kui Yang, Wei-Qiang Zhang, Wen-Lin Zhang, and Jia Liu. "Application of i-vector in speech and music classification." In 2016 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 1-5. IEEE, 2016.

[7] Lim, Chungsoo, and Joon-Hyuk Chang. "Efficient implementation techniques of an svm-based speech/music classifier in smv." Multimedia Tools and Applications 74, no. 15 (2015): 5375-5400.

[8] Tsipas, Nikolaos, Lazaros Vrysis, Charalampos Dimoulas, and George Papanikolaou. "Efficient audio-driven multimedia indexing through similarity-based speech/music discrimination." Multimedia Tools and Applications 76, no. 24 (2017): 25603-25621.

[9] Sell, Gregory, and Pascal Clark. "Music tonality features for speech/music discrimination." In 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 2489-2493. IEEE, 2014.

[10] Fuchs, Guillaume. "A robust speech/music discriminator for switched audio coding." In 2015 23rd European Signal

978-1-7281-9180-5/20/$31.00 ©2020 IEEE

Processing Conference (EUSIPCO), pp. 569-573. IEEE, 2015.
[11] Lavner, Yizhar, and Dima Ruinskiy. "A decision-tree-based algorithm for speech/music classification and segmentation." EURASIP Journal on Audio, Speech, and Music Processing 2009 (2009): 1-14.
[12] Pikrakis, Aggelos, Theodoros Giannakopoulos, and Sergios Theodoridis. "A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks." IEEE Transactions on Multimedia 10, no. 5 (2008): 846-857.
[13] Zhou, Huiyu, Abdul Sadka, and Richard M. Jiang. "Feature extraction for speech and music discrimination." In 2008 International Workshop on Content-Based Multimedia Indexing, pp. 170-173. IEEE, 2008.

978-1-7281-9180-5/20/$31.00 ©2020 IEEE