# Clustering of Instruments in Carnatic Music for Content based Information Retrieval

Surendra Shetty
Department of MCA, NMAMIT., Nitte
Udupi, Karnataka
India
hsshetty4u@yahoo.com

Sarika Hegde
Department of MCA, NMAMIT., Nitte
Udupi, Karnataka
India
sarika.hegde@yahoo.in

***Abstract*— Music Information Retrieval (MIR) focuses on retrieving useful information from collection of music. The objective of research work in this paper is to explore clustering approaches which can be useful in automatically mining the content from Carnatic instrumental music. The content to be retrieved is the instrument that is primarily used to play the song. Carnatic music songs with ten different instruments namely, *Flute, Harmonium, Mandolin, Nagaswara, Santoor, Saxophone, Sitar, Shehnai, Veena and Violin* are considered as input. Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coefficients (LPC) features are used for representing music information. In the first step, visualization technique is used to explore the capability of different features in distinguishing Carnatic music with different instruments. Then different clustering techniques are used for understanding natural way of grouping among this instrumental music. A discussion on the comparison of instrument clustering results with different algorithms, combined with various features is also presented.**

***Index Terms*—Instrument Detection, Music Information Retrieval, Clustering**

## I. INTRODUCTION

Research in MIR helps in automating the process of labeling the songs with metadata and also search the music based on the actual content. The content can be regarding, genre of music, singer, instruments etc. We can find lot of research works focusing the techniques, methodology for efficient content retrieval of Western music. But, research in Indian music is started recently mainly studying on *Raaga* identification from music. In our paper, we are analyzing the clustering of instrument which is a part of MIR for *Carnatic* music. The types of instruments used in *Carnatic* music are different from those used in western music. Instrumental songs are also considered as most melodic and there are many who are interested in listening and practicing music in a particular type of instrument. There is a lot of interest in searching or accessing a song based on a type of instrument. In this regard, our paper discusses the grouping of carnatic music based on instruments used to play the music. The various features and approaches are compared, which gives an idea of suitability of a particular audio feature in representing/grouping music based on instruments. The objectives of the research work discussed is converting instrumental music into Mel Frequency Cepstral Coefficients, Linear Predictive Coefficients features and then visualization of these audio features showing the suitability of features in distinguishing various instruments. After which clustering of audio features using K-means, Fuzzy c-means, Hierarchical and GMM clustering algorithms is done.

Rest of the paper content is organized as follows. In the second section, we describe the background of related work. Feature extraction and clustering techniques used in our research work is described in third section. The method of conducting the experiments and the results are discussed in fourth section.

## II. BACKGROUND

Applications of pattern recognition technique in classification of musical instruments are described in Martin & Kim [1]. They have used the different features which are computed based on source excitation and resonance structure. These features are measured from the output of an auditory model. Methods have been proposed for automatic singer identification in [2]. Here authors have used two step algorithms. Vocal segments present in a song are identified with an untrained algorithm in the first step. The vocal segments are then given as input to the second step of singer identification algorithm. This part of the algorithm is trained with the data to identify the artists of the song. Similarly a system of singer identification is proposed by Zhang [3]. Here the algorithm attempts to distinguish singing voice from instrumental sounds in the given song. The designed statistical algorithm is trained with data for each singer's voice. Mesaros and Astola [4] have analyzed the contribution of MFCC feature for singer identification and tested it. Features like, LPC is used with Support Vector Machine (SVM) algorithm for *6* instruments and is classified successfully by Chetry & Sandier [5]. Musical instrument identification is also performed for large scale database with branch and bound technique used to select the feature set [6]. The problem musical instruments identification for monophonic musical signals is investigated with help of MFCC over several time

978-1-4673-8286-1/16 $31.00 © 2016 IEEE
DOI 10.1109/IACC.2016.33

scales [7]. Zlatintsi & Maragos [8] have proposed non-linear methods using fractal theory by analyzing the structure of musical signals at multiple time scale and the technique is applied for instrument classification. A method for musical instrument recognition in polyphonic signals is proposed based on local features and missing feature techniques by Giannoulis & Klapuri [9]. In their paper, Diment, Heittola and Virtanen [10], have explored semi-supervised learning techniques for musical instrument recognition. The technique for identification of instruments with an input given polyphonic music is proposed in [11]. In this approach two algorithm is used. One algorithm does the transcription of polyphonic music automatically which also does the job of instrument assignment. Another algorithm is used for recognition of instruments based on missing theory feature. HG & Sreenivas [12] have proposed a solution for detection of instruments in a polyphonic music using Factorial Gaussian Mixture-HMM (F-GM-HMM) algorithm.

.

## III. METHODOLOGY AND TECHNIQUES

Any type of data mining involves the following steps, data preprocessing, feature extraction, classification model construction and evaluation, and decision logic. The type of data taken as input in our case is audio, and so the processing and feature extraction method involves Digital Signal Processing (DSP) technique. In this section we describe about the data and the various techniques used in our work.

### 3.1 Dataset

Here the dataset is a collection of Carnatic instrumental songs played with various instruments like, *Flute, Harmonium, Mandolin, Nagaswara, Santoor, Saxophone, Sitar, Shehnai, Veena and Violin* (Monophonic audio clips collection from CD's). We have cut the first portion of audio and used it for collection of songs in .wav format. Fifteen songs of each instrument type are used totally constituting 150 songs in the dataset collection.

### 3.2 Audio data preprocessing and feature extraction

Feature extraction is one of the very important steps in audio data mining which involves converting the musical audio into representative numbers called features which we also refer as audio feature. Feature extraction technique becomes crucial as it transforms the way of representing the audio data and also any further processing is conducted on extracted features. Digital Signal Processing (DSP) technique is used for audio preprocessing and feature extraction. Instead of applying feature extraction for the entire audio data, framing is done. Framing is a process of breaking the set of samples of an entire audio file into smaller chunks called as frame. For each of the frame $f_i$, containing $L$ samples of the audio file, feature extraction technique is applied and it converts the frame into a feature which may be a single value or a vector containing sequence of values/coefficients represented as $FV = \{fv_1, fv_2 \cdots, fv_d\}$ where '$d$' is the total number of coefficients/attributes of single feature. We now describe Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coefficients (LPC) features.

***a. Mel Frequency Cepstral Coefficients (MFCC):***

The Mel-Frequency Cepstral Coefficient (MFCC) is a feature extracted by applying more than one Fourier Transform to the original signal in sequence [13, 14]. The first step is preprocessing which consists of framing and windowing of the signal. For each of the frame $f_i$, DFT coefficients are calculated by applying FFT algorithm. After applying the DFT, the resulting power spectrum is transformed to Mel-frequency scale. This is done by using a filter bank consisting of triangular filters, spaced uniformly on the Mel-scale. Finally, the cepstral coefficients are calculated from the Mel-spectrum by taking the Discrete Cosine Transform (DCT) of the logarithm of the Mel-spectrum. This is given by

$$C_i = \sum_{k=1}^{K} (logS_k).\, cos\left(\frac{i\pi}{K}\left(k - \frac{1}{2}\right)\right) \quad i = 1,2,\ldots,K \qquad (1)$$

where $C_i$ is the $i^{th}$ MFCC, $S_k$ is the output of $k^{th}$ filterbank channel (i.e. the weighted sum of the power spectrum bins on that channel), $K$ is the number of filterbanks [15].

The more MFCCs, the more precise the approximation of the signal's spectrum and it also means more variability of the data [16]. As the interest is only in the spectral envelopes, not in the finer, faster details like pitch, a large number of MFCC coefficients may not be appropriate. Pachet & Aucouturier [16] have used *eight* coefficients. Following this, we have used less than eight coefficients and for the implementation Auditory Toolbox is used [14]. Out of the 13 coefficients computed, we have considered the first *six* coefficients.

***b. Linear Predictive Coefficients (LPC):***

Linear predictive analysis is one of the most commonly used speech analysis tool. The basic idea behind linear prediction is that the next signal sample is predicted from a weighted sum of $p$ previous samples, given as follows: [17].

$$\hat{s}(n) = \sum_{i=1}^{P} a_i\, s(n-i) \qquad (2)$$

where the set $\{a_i\}$ is the set of prediction coefficients and $s(n-i)$ is a sample at time instant $(n-i)$. In other words, each sample of a signal is modeled as a linear combination of previous samples. The prediction coefficients are determined by minimizing the mean squared error between the actual sample and the predicted samples. We have considered the first *six* LPC coefficients out of *thirteen* coefficients computed. For the analysis, we have used the number of LPC coefficients same as number of MFCC coefficients. For each frame $f_i$, the sequence of *six* LPC coefficients is given as, $LP = (lp_1, lp_2, lp_3, lp_4, lp_5, lp_6)$ . LP is used as the feature vector $(FV)$ for discriminating different classes of audio signal.

### 3.3 Clustering

Clustering is a technique which is quite similar to classification approach. The difference in classification and clustering technique lies in whether the categories of different classes to which the given data belong are known in prior or not. In classification we have a set of labeled data where we know the class labels for each the observation and the classification algorithm is trained to learn about this.

A clustering algorithm is not given with any information about the class labels of the data. Grouping of data is done purely based on the similarities among the observations. A set of observations which are nearer to each other are made as one group in clustering process. These groups are called as clusters. Mainly distance calculation technique is used to compute the similarity and dissimilarity measures. The main objective of clustering is to detect the underlying, natural grouping of data which can be further used for classification and pattern recognition etc.

***a. Hierarchical Clustering*:** The algorithm produces a hierarchy of clustering. At each step, the two clusters with the shortest distance are merged into one, creating a new cluster and correspondingly the distance of the new cluster with the rest is calculated. This process is repeated until we have $c$ clusters

***b. K-means clustering:*** K-means clustering is partitioning method of clustering. Here we specify $K$ a user specified parameter, namely, the number of clusters desired. We first chose, $K$ observations from dataset randomly and assign it as the $K$ initial centroids. For each observation, we compute Euclidian distance between the observation and all the centroids. Then, each observation is assigned to the centroid with the smallest distance [18].

***c. Fuzzy C-means Clustering:*** This is actually a fuzzy version of K-means algorithm, which is called fuzzy c-means. It is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership weight, $w_{ij} \in [0,1], j = 1, \dots, m,$ which indicates the value of membership for $i^{th}$ observation in $j^{th}$ cluster. For an observation $FV^i$, $\sum_{j=1}^{m} w_{ij} = 1$

d. ***Gaussian Mixture Model Clustering***: In this method it is assumed that the data has been generated from a statistical model and then the data is described by finding the statistical model that best fits the data. At a higher level this process involves deciding on a statistical model for the data and estimating the parameters for that distribution. Mixture models view the dataset as a set of observations from a mixture of different probability distributions [19]. Assume that there are $m$ distributions and $k$ observations, $X = \{x_i, i = 1, \dots, k\}$

Let the $j^{th}$ distribution have parameters $\theta_j$, and let be the set of all parameters, i.e. $\Theta \equiv [\theta_1, \theta_2 \cdots, \theta_m]$. Then, $prob(x_i|\theta_j)$ is the probability of $i^{th}$ observation coming from the $j^{th}$ distribution. The probability that $j^{th}$ distribution is chosen to generate an observation is given by the weight $w_j$ $1 \le j \le m$, where these weights (probabilities) are subject to the constraint $\sum_{j=1}^{m} w_j = 1$. Then, the probability of an observation $x_i$ is given by,

$$prob(x_i|\Theta) = \sum_{j=1}^{m} w_j \, prob(x_i|\theta_j) \quad (3)$$

Since the observations are generated in an independent manner, then the probability of the entire dataset is just the product of the probabilities of each individual $x_i$.

$$prob(X|\Theta) = \prod_{i=1}^{k} prob(x_i|\Theta) \quad (4)$$

The probability density function for Gaussian distribution at a point $x_i$ is,

$$prob(x_i|\Theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}, \; -\infty < x_i < \infty, \sigma > 0 \quad (5)$$

The parameters of the Gaussian distribution are given by, $\Theta = (\mu, \sigma)$, where $\mu$ is the mean of the distribution and $\sigma$ is the standard deviation. We can estimate the parameters of these distributions from the data and describe these distributions (clusters).

### 3.4 Cluster Evaluation

We have used a supervised method of cluster evaluation where the usual procedure is to measure the degree of correspondence between the cluster labels and the class labels. There are two kinds of approaches here. The first set of techniques called as classification-oriented, such as purity. These measures evaluate the extent to which a cluster contains objects of a single class. Purity is computed as,

$$purity = \sum_{i=1}^{C} \frac{m_i}{k} p_i \quad (6)$$

where, C is the number of clusters, where $m_i$, is the number of observations in cluster $i$ and k is the total number of observations.

The second types of methods are similarity-oriented, such as Jaccard measure and Rand statistic. These approaches measure the extent to which two observations that are in the same class are in the same cluster and vice-versa. The two matrices i.e. ideal cluster similarity matrix and ideal class similarity matrix are compared. The ideal cluster similarity matrix is defined as below,

$$cluster_{matrix}[i,j] = \begin{cases} 1 \; if \; i^{th} \; and \; j^{th} observations \; are \; in \; the \; same \; cluster \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases}$$

Ideal class similarity matrix is:

$$class_{matrix}[i,j] = \begin{cases} 1 \; if \; i^{th} \; and \; j^{th} observations \; are \; in \; the \; same \; class \\ 0 \qquad\qquad\qquad\qquad otherwise \end{cases}$$

To compute the binary similarity between the two matrices we compute the following four quantities.

$f_{00}$=number of pairs of observations having a different class and a different cluster

$f_{01}$=number of pairs of observations having a different class and the same cluster

$f_{10}$=number of pairs of observations having the same class and a different cluster

$f_{11}$=number of pairs of observations having the same class and the same cluster

After computing the above parameters, the simple matching coefficients known as the Rand statistic and the Jaccard coefficient, two of the most frequently used cluster validity measures are computed as follows

$$Rand\ Statistic = \frac{f_{00}+f_{11}}{f_{00}+f_{01}+f_{10}+f_{11}} \quad (7)$$

$$Jaccard\ coeffecient = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} \quad (8)$$

In addition to comparing the algorithms, this technique can also be used to determine the how well the clusters are formed in a given set of data, can be used to estimated the number of classes, and also can be used to compare the results of natural grouping through clusters with the given external information like classes.

We explain the results of applying different clustering techniques i.e. K-means clustering, Fuzzy c-means clustering, Hierarchical clustering and GMM clustering to the dataset. We also analyze the results of clustering with cluster evaluation measures. Each experiment is repeated for *ten* times and we have found that values of evaluation measures are almost stable in these trials. The average of these values are computed and shown in the results of all the following experiments.

## IV. EXPERIMENTAL RESULTS AND DISCUSSION

Both MFCC and LPC features are extracted from each of the songs in the dataset. We consider feature vector of each frame as an observation. In all the further experiments, we randomly select a sample of observations. For visualization, we have randomly selected 25 feature observations from each of the *ten* types of musical instruments and plotted the first three MFCC coefficients $(fv_{i1}, fv_{i2}, fv_{i3})$ out of *six* from each feature vector in 3D-space as shown in Fig 1.

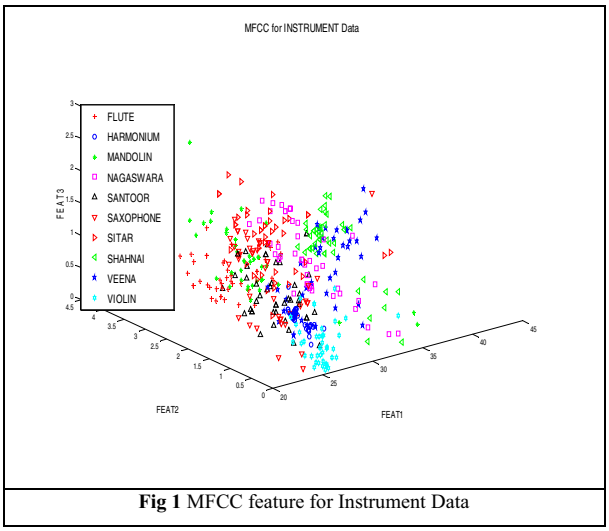


**Fig 1** MFCC feature for Instrument Data

The feature vectors are computed for the frame size of 100ms. We observe from the plot that the observations belonging to same type of instruments (indicated with same symbols) are forming a grouped as they are nearer to each other compared to those of other types. Even in visualization with LPC coefficients, it is observed that the observations belonging to same instrument type are grouped together. Similar types of instruments like *Saxophone* and *Nagaswara* groups are nearer to each other. This shows the possibility of using the LPC coefficients as one of the most promising feature for discriminating the different instrument classes.

Visualization gives some idea about the separability of instruments with MFCC and LPC features. Clustering experiment is done to analyze whether the MFCC/LPC features are able to group together the observations of same class using clustering techniques. To start with, a small dataset is considered by randomly selecting the *50* observations from the dataset constructed from *three* instruments, *Flute, Harmonium* and *Santoor* with MFCC feature. We applied all clustering techniques, with number of clusters, $k = 3$ as input.

**Table 1: Purity (3 Clusters), Rand Statistic and Jaccard Coefficient Measures for clustering of Instrument data**

| Clustering Algorithm | Feature | Purity | | | Rand Statistic | Jaccard Coefficient |
|---|---|---|---|---|---|---|
| | | Cluster1 | Cluster2 | Cluster3 | | |
| K-means Algorithm | MFCC | 0.648 | 0.593 | 1.000 | 0.781 | 0.510 |
| | LPC | 0.548 | 0.680 | 0.857 | 0.681 | 0.398 |
| Fuzzy C-means Algorithm | MFCC | 0.903 | 0.909 | 0.907 | 0.887 | 0.708 |
| | LPC | 0.857 | 0.680 | 0.548 | 0.681 | 0.398 |
| Hierarchical Algorithm | MFCC | 0.980 | 0.942 | 1.000 | 0.964 | 0.898 |
| | LPC | 0.863 | 0.720 | 0.505 | 0.686 | 0.410 |
| GMM Algorithm | MFCC | 0.979 | 1.000 | 0.942 | 0.965 | 0.900 |
| | LPC | 0.823 | 0.648 | 0.822 | 0.746 | 0.444 |

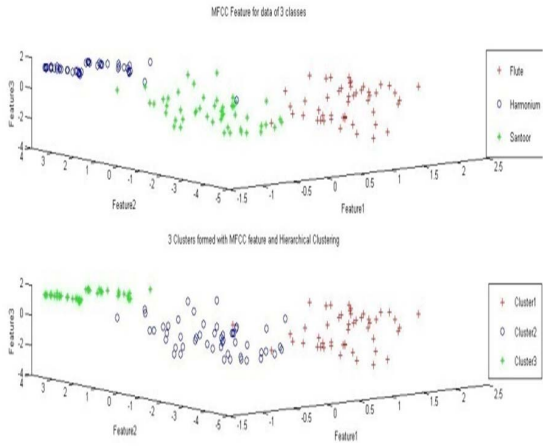


**Fig 2** Instrument data with MFCC feature and results of Hierarchical clustering

Since the Hierarchical and GMM clustering algorithms with MFCC feature perform better, we are showing the graphical views of these results in figure 2 and figure 3. In figure 2, the class *flute* (plus symbol) is clustered into Cluster 1(plus symbol). Class *Harmonium* (circle symbol) is mapped onto Cluster 3(star symbol) and class *Santoor* (star symbol) is mapped onto Cluster 2(circle symbol). In both figures, we can observe that most of the observations in same class are also found to be in same cluster. This point is also supported by the evaluation measure values, with purity value more than 90% for each individual cluster (Table 1, 5th and 7th row).

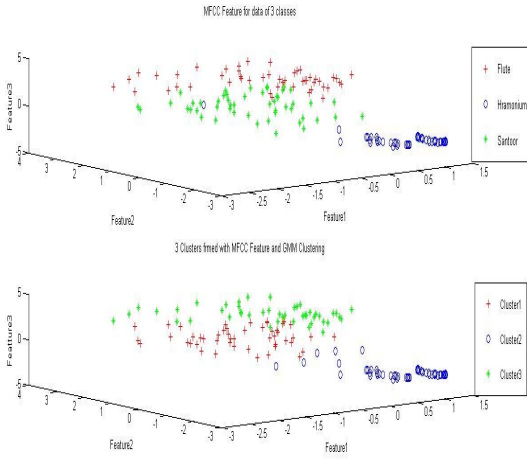


**Fig 3** Instrument dataset with MFCC features and results of GMM clustering

The results of the variations in rand statistic and the purity value for MFCC features are shown in figure 4(a) and figure 4(b) respectively. In those figures, it can be observed that GMM algorithm performance is slightly better compared to others. The GMM clustering result is highest for sample size of 50, 150 and 200. For the sample size of 100, fuzzy c-means is better compared to others. Here, Hierarchical and K-means algorithms perform poorly compared to GMM and Fuzzy c-means algorithms. The highest rand statistic value for the 200 observations is 0.379 and the highest purity achieved for the clustering of instrument data with MFCC feature is 59%.

Similar experiments are conducted for instrument data using LPC feature. The rand statistic and the purity value for varying number of observations are shown in figure 5(a) and 5(b). The highest purity achieved is 44% which is slightly less compared to the clustering with MFCC feature.

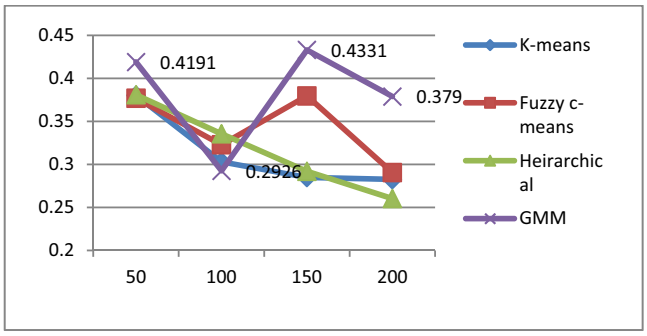


**Fig 4 (a)** Variation of Rand Statistic for Instrument data with MFCC

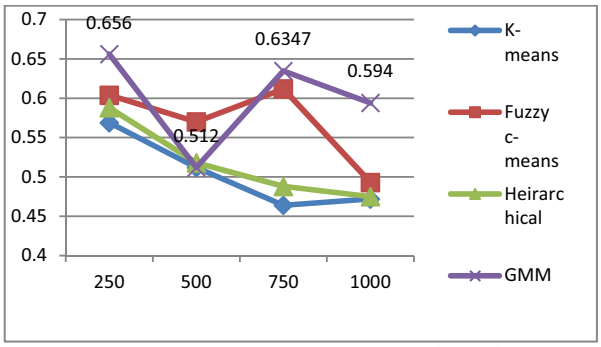


**Fig 4 (b)** Variation of Purity for Instrument data with MFCC

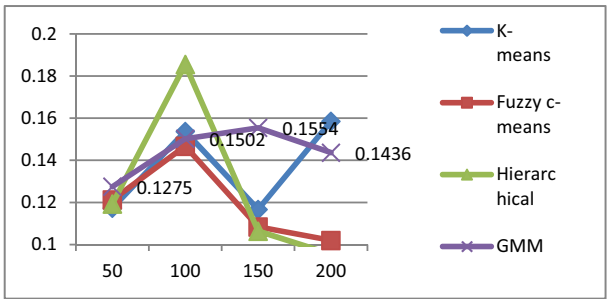


**Fig 5 (a)** Variation of Rand Statistic for Instrument data with LPC

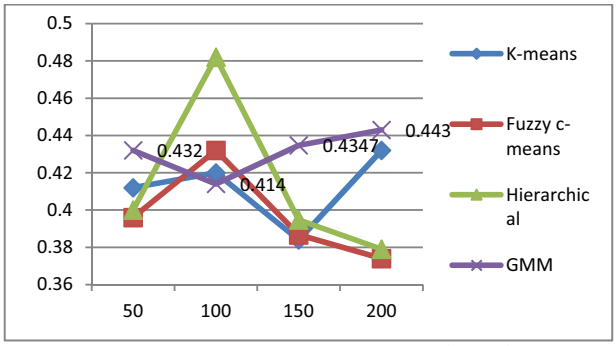


**Fig 5 (b)** Variation of Purity for Instrument data with LPC

For the small dataset of 50 observations, we found that the clusters are formed in exactly the way of classes. This is found to be true for all the dataset with all four clustering techniques. As we increased the number of observations to be clustered, the algorithms showed degradation in the performance. The performance of MFCC feature is slightly better compared to LPC for instrument data. The results show that there exists a natural structure within the data of same class in the given audio data. This is clear when we take small size data and apply clustering techniques. But as we increase the size of dataset, clustering algorithm could not detect the natural structure within the data very efficiently. This type of analysis would be helpful for comparing the results of clustering algorithm for audio datasets. It also determines that which kind of feature demonstrates natural structure better than others for a particular type of dataset.

## V. CONCLUSIONS

The Instrument data from Western music have been analyzed very much but there is less focus given for Indian instruments. Features like MFCC, and LPC have been used for representing these data. The result of feature extraction shows that the MFCC and LPC coefficients are promising feature for this kind of data. Results of visualizations help us to decide what feature would be valid for a given type of data. As we increased the number of observations to be clustered, the algorithms showed degradation in the performance. For instrument data, GMM algorithm clustering results are better compared to others, which is followed by Fuzzy c-means algorithm. The performance of MFCC feature is slightly better compared to LPC for instrument data.

## REFERENCES


[1] Martin, K. D., and Kim, Y. E. (1998). 2pMU9: Musical Instrument Identification: A Pattern-Recognition Approach. *136th meeting of the ASA*.

[2] Kim, Y.K., and Brian, Y., (2002). Singer Identification in Popular Music Recordings Using Voice Coding Features. *In Proceedings of ISMIR.*

[3] Zhang, T., (2003). System and Method for Automatic Singer Identification. *In Proceedings of the IEEE Conference on Multimedia and Expo*, Vol. 1, pp. 33-36.

[4] Mesaros, A., and Astola, J., (2005). The Mel-Frequency Cepstral Coefficients in The Context of Singer Identification. *In Proceedings of the International Conference on Music Information Retrieval.*

[5] Chetry, N; and Sandier (2006).Linear Predictive Model for Musical Instrument Detection,; *IEEE Conference on Acoustics, Speech and Signal Processing*, Vol. 5, 2006.

[6] Benetos, E., Kotti, M., Kotropoulos, C., (2007). Large Scale Musical Instrument Identification. *Proceedings SMC'07, 4th Sound and Music Computing Conference*, Greece.

[7] Sturm, B. L., Morvidone, M., and Daudet, L.,(2010). Musical Instrument Identification using Multiscale Mel-frequency Cepstral Coefficients. *18th European Signal Processing Conference (EUSIPCO-2010).*

[8] Zlatintsi, A., & Maragos, P. (2013). Multiscale fractal analysis of musical instrument signals with application to recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, *21*(4), 737-748.

[9] Giannoulis, D., & Klapuri, A. (2013). Musical instrument recognition in polyphonic audio using missing feature approach. *Audio, Speech, and Language Processing, IEEE Transactions on*, *21*(9), 1805-1817.

[10] Diment, A., Heittola, T., & Virtanen, T. (2013, September). Semi-supervised learning for musical instrument recognition. In *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European* (pp. 1-5). IEEE.

[11] Giannoulis, D., Benetos, E., Klapuri, A., & Plumbley, M. D. (2014, May). Improving instrument recognition in polyphonic music through system integration. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 5222-5226). IEEE.

[12] HG, R., & Sreenivas, T. V. (2015, April). Multi-instrument detection in polyphonic music using Gaussian Mixture based factorial HMM. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (pp. 191-195). IEEE.

[13] Gold, B., and Morgan, N., (2006). Speech and Audio Signal Processing – Processing and Perception of Speech and Music. *Wiley India Pvt. Ltd.*, ISBN: 81-265-0822-1.

[14] Slaney, M., (1998). Auditory toolbox: A MATLAB Toolbox for auditory modeling work, *Tech. Rep. 1998-010, Interval Research Corporation, Palo Alto, Calif, USA,* 1998, Version 2.

[15] Rabiner, L., and Juang, B., (1993). Fundamentals of Speech Recognition. *Prentice Hall,* ISBN,-10:013051572.

[16] Aucouturier, JJ., and Pachet, F., (2004). Improving Timbre Similarity: How high's the sky? *Journal of Negative Results in Speech and Audio Science,* Vol 1, No 1, 2004.

[17] Peltonen, V., Tuomi, J., Klapuri, A, Huopaniemi, J., Sorsa, T., (2002). Computational Auditory Scene Recognition. *In IEEE International conference on Acoustics Speech and Signal Processing*, Vol 2, 2002.

[18]. Alpaydin, E., (2004). Introduction to Machine Learning, *PHI Publications*, ISBN-81-203-2791-8.

[19] Theodoridis, S., and Koutroumbas, K., (2009). Pattern Recognition, *Fourth* Edition, *Academic Press.*