North Indian Classical Music's Singer Identification by Timbre Recognition using MIR Toolbox

Saurabh H. Deshmukh
Info Technology Department,
G.H.Raisoni COE&M, Wagholi, Pune, India

S.G. Bhirud, Ph.D
Computer Engineering Department,
Veermata Jijabai Technological Institute,
Mumbai, India.

ABSTRACT

Timbre recognition has always been associated with musical instrument identification. In music information retrieval (MIR) community there has always been increasing interest in identifying singers and identifying the musical instruments. There can be unending debate on, whether to consider human throat as a kind of musical instrument generating sound through air pressure and vibration or not. Any musical sound possesses 'Timbre', an unidentified and undefined entity which is yet non-tangible that uniquely defines the sound. A lot of research has been done on catching this fuzzy term in terms of its synonyms related to light or texture or so. MIR toolbox of MatLab provides some strong techniques to extract variety of audio attributes/characteristics from an audio file. These attributes are called as 'Audio Descriptors'. In this paper, we have studied such audio descriptors that fall under the Timbre category and found out the most prominent audio descriptors that plays substantial role in the identification of a singer from North Indian classical music. In this system we have considered Noisy data, Noise filtered data without background instruments and noise free studio recordings with only Tanpura instrument in the background as input. We have found that roll off, brightness, roughness and irregularity are the four strong audio descriptors designated under the Timbre category of MIR Toolbox, plays vital role in the identification of a singer from North Indian Classical Music giving accuracy of identification of 96.66% for three singers trained and tested simultaneously on studio recordings containing Tanpura(a Supportive Musical instrument of the singer) each of 5 sec duration and sampling rate of 11,025Hz, 16 bits with Pulse Code Modulation (PCM) uncompressed file format. The results indicate that a singer can also be treated as a wind type of musical instrument and be successfully identified by recognizing its Timbre.

Keywords

Music Information Retrieval, North Indian classical Music, MIR Toolbox, Timbre, Singer identification.

1. INTRODUCTION

Singer identification has been one of the most popular and important research areas of signal processing in early decades. Later, the problem of identifying a singer has been well taken by first, content based audio retrieval (CBAR) community and then further by music information retrieval (MIR) community. These communities worked toward identifying various features of a singing voice by different feature extraction techniques and then to classify the singer using a classifier. There has been vast research on extraction of useful musical and singing voice information from a piece of an audio file. Researchers have used these features in different ways. The taxonomy of such audio features has also been changed with the approach towards the features and methods of extracting the features.

The audio features can be arranged in hierarchical structure as per similarities in their characteristics [1]. *Steadiness or Dynamicity* of the feature, *Global or instantaneous* representation of the feature, *abstractness* of the feature, *method of extraction* of the feature and so on are the ways in which taxonomy of the audio features of an audio file can be represented [2]. More or less each research that does automatic singer identification, work on this kind of taxonomy of the audio features and its approach towards looking at the audio features being extracted. Also, even after classifying these audio features (also called as audio descriptors) there remain some descriptors, such as pitch or fundamental frequency, which cannot be classified into any category.

The presented work elaborates our approach towards identifying a singer, by identifying its timbre. The entire literature of singer identification is based on typical speaker recognition systems that use Mel Frequency Cepstral Coefficients (MFCC) and its variants or Linear Predictive Coding Coefficients (LPCC) and its variants as feature extractors. Very rarely other feature extraction methods are used for identifying the singer from North Indian Classical Music.

This manuscript is arranged in the following way. First, the concept of "Sound Timbre" and the literature survey on various definitions and usage of sound timbre have been explained. Then, the taxonomy of audio descriptors that fall under concept of timber as per Music information Retrieval Community (MIR) that we have followed has been given. For the input of North Indian classical vocal, which is homophonic [3](and not monophonic or polyphonic), some necessary aspects are to be studied before we use the audio samples. These special aspects of audio files containing North Indian Classical vocal are explained further. Finally, the experiments, results and the conclusions are discussed.

2. CONCEPT OF SOUND TIMBRE

Most popular and may be first definition of timbre has been "the psycho acoustician's multidimensional waste-basket category for everything that cannot be labeled pitch or loudness." [4]. The timbre usually refers to be a feature of an audio that allows us to differentiate two sounds that are of same pitch, loudness and duration [5]. Timbre is a loosely defined term representing many components together in a sound. In literature, no complete definition of timbre is available [6], neither any unit of timbre has been yet defined. A lot of researchers have given names synonymous to timber. Timbre has always remained a non-tangible entity.

2.1 Can Timbre be used for singer identification?

Musical instruments are identified by timbre identification [6]. There has been a continuous controversy and debate on whether human throat can be considered as a kind of musical instrument or not? We do not want to indulge in this debate of right or wrong. A lot of researchers believe that timbre is to be related only to the sound generated through instruments or objects. While, on the other side if human throat is assumed to be a kind of wind instrument(like Saxophone) in which sound is produced because of the vibration occur in the vocal tract by movement of air passing through it then, we may consider timbre aspects of the sound generated through human throat.

Also, there are many aspects of timbre of an instrument if, we consider the *timbre shapers* used. These are bridges, cotton threads etc. that are used to sharpen or smoothen the sound generated through the instruments as in case of a Tanpura or Sitar instrument, for example. Ideally, in such situation, no two musical instruments can be said to have same or similar timbre of sound. Thus, even if we identify two musical instruments as, for example, a sitar and a violin then still we are not able to identify which sitar and which violin if there are more than two such instruments.

Timbre similarity methodologies to identify a singer or an instrument have shown increasing interest in the music information retrieval community [7]. We have believed in this research that a singer's voice does contain timbre aspects and we can use them in identifying the singer uniquely.

2.2 Why Timber is Complex to define?

In North Indian Classical Music the timbre is referred to the CAST of the voice. It is also termed as 'Tonal Quality' [8], some researchers have given it a term related to color [9] also some has used it as "tone color" [10] . To summarize the timbre attributes that makes it complex [4] are (i) It is perceptual and subjective. (ii) It is multidimensional in nature. (iii) There are no scales to judge or measure the timbre (iv) Study of timbre is itself interdisciplinary (v) There are no standard set of sound examples that can be compared to check if we have modeled the timbre correctly or not.

There are more than twenty definitions yet continuing about what exactly timbre is? And how it can be extracted from sound file? Some researchers have added 'log of raise time' and 'irregularity' as components of timbre attribute which are substituted by [11] into 'spectral irregularity with 'spectral flux'. Overall it depends on the researchers view point as to which low level audio descriptors, out of 52 audio descriptors specified by MPEG community [12], fall under category of Sound Timbre.

3. MIRTOOLBOX AND TIMBRE

The MIRtoolbox [13], [14], [15] designed to be used with MATLAB has tremendous strength for various ways in which audio features can be extracted and analyzed. The functions used to extract the audio features are so powerful that every time a new researcher doesn't have to recode the typical operations of re sampling, filtering, feature extraction etc. The functions can be grouped together to perform operations that are interdependent or can also be performed standalone. This toolbox is to be best used with Signal Processing Toolbox of MATLAB.

MIRtoolbox has its own taxonomy to include following audio descriptors that describe timbre. The toolbox has included timbre extractor *as attacktime, slope, zerocross, rolloff, brightness, mfcc, roughness* and *irregularity*. There are many systems designed till today that make use of MFCC (Mel Frequency Cepstral Coefficients) as primary technique to extract the audio features of a singer / speaker in terms of cepstral coefficients (usually 13 coefficents) and then use these features to generate a singer/ speaker model which can later be trained and classified. Olivier Lartillot and his team have cleverly included *mfcc* into the group of timbre extractors. This has surely helped us in identifying the singer with more accuracy. Although no literature as far as our knowledge declares *mfcc* as one of the components of timbre rather it is a representation of short term power spectrum of a sound, based on a linear cosine transform of log power spectrum on a non linear mel scale of frequency [16]. In following section we have explained the experiments, the methods used to extract the features, the classifier used and the results

4. SINGER IDENTIFICATION PROCESS AND EXPERIMENTS

The singer identification process uses three modules viz. an Input module, a Query Module and a Retrieval Module [17]. The algorithm of singer identification uses audio data filtering and standardization as first step in the Input module. By making use of MIRtoolbox functions the audio features are extracted in the feature extraction process. These feature vectors are then given to K- means Clustering algorithm to generate the codebooks in Query module. The test samples are fed to the retrieval module where K-Means clustering is used to identify the singer.

4.1 Input Database

We have used total three databases. All of them are filtered using inverse comb filtering technique and then are re-sampled to 11,025Hz with Pulse Code Modulation (PCM) uncompressed file format, 16 bits, with duration of 5 sec each and mono channel. The three databases DB1, DB2 and DB3 differ with respect to presence of noise and background music of accompanying instruments.

DB1 contains total 9 singers with 8 samples per singer (72 Files), recorded in a little noisy environment from North Indian classical Music's singers, through a hand held audio recorder. The purpose of using this kind of noisy audios was two folded. As per [18] the timbre has five major attributes. Out of the rest the first has been that, the timbre is a range between tonal and noise like character.

Theoretically, if outside noise is added to this database then all noise including some portion of timbre should get filtered and then the system should not identify the singer. Second aspect of this is to generalize the system performance and make a robust system susceptible to normal noise that comes along with input in the audio recordings.

DB2 contains Total 9 singers with 8 samples per singer (72 files). This database has been recorded in studio like environment containing very less noise as compared to DB1.

DB3 contains high quality audio recordings of total 10 singers with 10 samples per singer (100 Files), from popular North Indian classical vocalists' recordings. These recordings does not contain any rhythm instrument but they posses continuous background music of Tanpura and Violin/Harmonium/Flute etc. This database has been created this way to understand behavior of the system for background accompanying instruments generating homophonic music (North Indian Classical Music) in contrast to monophonic or polyphonic music (Western Classical Music).

4.2 Feature Extraction

The process of feature extraction has been very keenly formulated and systematically carried out so as to understand and find out effect of each audio descriptor, described under the title of timber in MIRtoolbox, on the overall singer identification process. The features are extracted by the algorithm proposed in [17]. In the foreword pass, one at a time each audio descriptor has been extracted and its effect on the final efficiency of singer identification has been recorded.

In the backward pass all less efficient are removed and kept only the prominent. Then again combinations of most effective audio descriptors were extracted and their effect on the efficiency of singer identification has been noted and so on. It is to be noted here that, for MFCC, which is one kind of critical band filters that is used to model usually the human auditory system behavior, there was a question of how many coefficients to be considered? As per literature 13 coefficients are usually extracted and used as feature vector. We have done all combinations of selecting first 10, 20 and then 30 coefficients and noted the effect on singer identification process and concluded that there is not much remarkable difference in the efficiency of singer identification process. Hence an average value of 20 coefficients has been considered.

4.3 Classifier

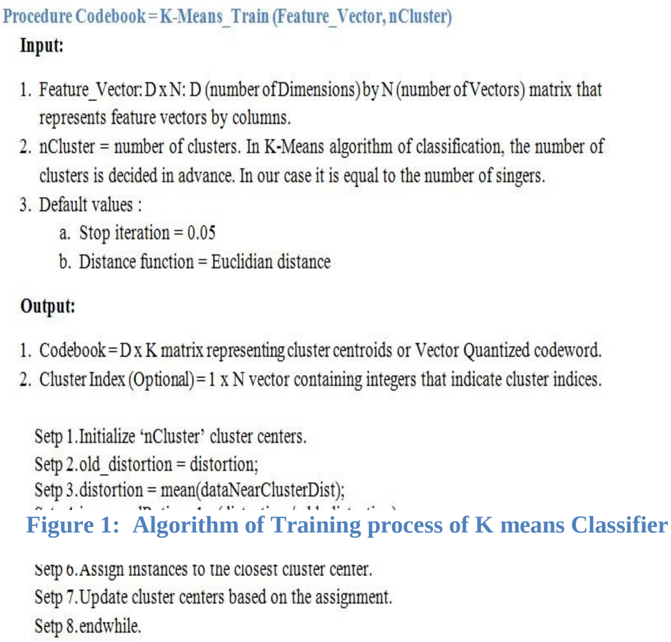
From the range of various classifiers such as K-Nearest Neighbor (KNN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Network (ANN) that can be used for singer identification, a very simple K-means classifier has been used since the aim was not to design a best singer identification system, rather was to test which audio descriptors under Timbre are helpful in successfully identifying a singer from North Indian Classical Music.

Algorithm: The K- means classifier algorithm, as explained in [19] is an iterative method that generates sequence of partitions (clusters) in the training dataset. It is supervised algorithm. Initially two inputs are given. These are a) dataset to be partitioned and b) the number of partitions (clusters) in which the data is to be allocated. Figure1 [20] indicates a generalized K-Means algorithm.

Initial partitions are generated by generating random centroids equal to the number of intended clusters. Then *Euclidean distances* of each data points from these centroids are calculated. On the basis of minimum distance these data points are assigned to the nearest centroids cluster. After finishing this process for all the dataset, again cluster centroids are recalculated and the process is repeated till there is no change in the centroids values generated in nth iteration and (n+1)th iteration.

4.4 Experimentations and Results

Each training and testing dataset was categorized into three classes Class A, Class B & Class C. Three input databases viz. DB1, DB2 and DB3 were fed to the system as shown in Table 1.



```
Procedure Codebook = K-Means_Train (Feature_Vector, nCluster)
Input:
1. Feature_Vector: D x N: D (number of Dimensions) by N (number of Vectors) matrix that
   represents feature vectors by columns.
2. nCluster = number of clusters. In K-Means algorithm of classification, the number of
   clusters is decided in advance. In our case it is equal to the number of singers.
3. Default values :
     a. Stop iteration = 0.05
     b. Distance function = Euclidian distance
Output:
1. Codebook = D x K matrix representing cluster centroids or Vector Quantized codeword.
2. Cluster Index (Optional) = 1 x N vector containing integers that indicate cluster indices.

  Setp 1.Initialize 'nCluster' cluster centers.
  Setp 2.old_distortion = distortion;
  Setp 3.distortion = mean(dataNearClusterDist);
  Setp 6.Assign instances to the closest cluster center.
  Setp 7.Update cluster centers based on the assignment.
  Setp 8.endwhile.
```

Figure 1: Algorithm of Training process of K means Classifier

Initially, only one audio descriptor had been extracted from each training dataset and efficiency of singer identification has been found out. When compared the results it was found that *rolloff, mfcc* and *irregularity* were giving better results for singer identification, hence other audio descriptors in combination with these were further tested. This process was repeated till we had exhausted all audio descriptors under Timbre category mentioned in MIRtoolbox. An average of known samples (Kn) and unknown samples (Un) efficiency has been calculated.

Following Table 2 shows the summary of the results obtained. It can be clearly observed that decreasing the number of singers for training and testing increases the efficiency of the system. Also the common audio descriptors giving best singer identification efficiency from all three datasets are *Roll off, Brightness and Irregularity.*

Table 1. Different types of databases and the number of training and testing files used

DB/Class	Class A	Class B	Class C	
DB1: Noisy data	9 Sg 5 Sp	5 Sg 5 Sp	3 Sg 5 Sp	Train
	9 Sg 8 Sp	5 Sg 8 sp	3 Sg 8 Sp	Test
DB2: Noise Filtered data without Background Music	9 Sg 5 Sp	5 Sg 5 Sp	3 Sg 5 Sp	Train
	9 Sg 8 Sp	5 Sg 8 sp	3 Sg 8 Sp	Test
DB3: Noise free Studio Recordings with Tanpura	10 Sg 7 Sp	5 Sg 7 Sp	3 Sg 7 Sp	Train
	10 Sg 10 Sp	5 Sg 10 sp	3 Sg 10 Sp	Test

Table 2. Summary of the results for all classes and all datasets. Highest efficiency is obtained for 3 singers in D3.

Audio Descriptors	No of Singers	Efficiency			Data Base
		Kn	Un	Avg	
ROLLOFF+MFCC+IRREGULARITY+ BRIGHTNESS	9	40.278	40.741	40.509	D1
	5	45	26.667	35.833	
	3	75	77.778	**76.389**	
ZCR+ROLLOFF+IRREGULARITY+BRIGHTNESS	9	61.111	66.665	63.889	D2
	5	60	60	60	
	3	70.833	66.667	**68.75**	
ROLLOFF+BRIGHTNESS+IRREGULARITY+ROUGHNESS	10	61	56.667	58.333	D3
	5	68	60	64	
	3	93.333	100	**96.667**	

The final best combination result is graphically shown in Figure2. The efficiency of three databases viz. D1 (Blue), D2 (Red) and D3 (Green), for all audio descriptor combinations, shows a growth in all three datasets, when number of singers were reduced for these combinations of audio descriptors.

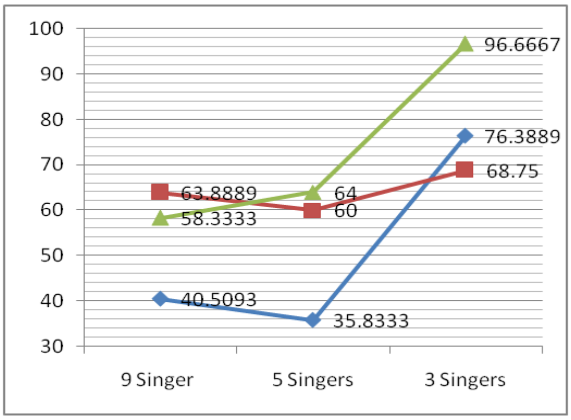


Figure 2. Graph showing highest singer identification efficiency as 96.66%, for 3 singers in Database D3.

5. CONCLUSION

In this paper, we have analyzed singer identification problem of North Indian Classical Music, with an angle of considering human singing voice as a musical instrument. The audio datasets of Noisy data, Noise filtered data without background instruments and noise free studio recordings with only Tanpura instrument in the background, was considered. We assumed that, human throat can also be imagined as a kind of wind musical instrument, where air passing through it generates sounds similar to a flute or a saxophone for example. K-means classifier which is supposed to be simplest method of data clustering had been used. The classification efficiency for noise free studio recordings along with musical instruments (in D3), for combination of *roll off, brightness, roughness* and *irregularity* clearly indicates that, these four audio descriptors grouped under Timbre category of MIRtoolbox has potential, to classify a singer efficiently without traditionally using *mfcc.* Moreover, the interference of background music, to singer identification process, has been filtered out automatically, because of the classification parameter aspects that are been considered. Also, as further research aspect, the common audio descriptors form all these results, giving efficiency in identifying a singer, were *roll off brightness* and *irregularity,* can be combined or integrated to generate a new feature vector and be used, in combination with statistical feature extractors to improve the efficiency of the problem of singer identification from North Indian Classical music. Finally, an average accuracy of singer identification of the entire system was found to be 80.59%.

6. REFERENCES

[1] DALIBOR MITROVIC, MATTHIAS ZEPPELZAUER, and CHRISTIAN BREITENEDER, "Features for Content -Based Audio Retrieval," *Advances in Computers*, vol. 78, pp. 71-150, 2010.

[2] Geoffroy Peeters, "A large set of audio features for sound description (Similarity and Classification) in the CUIDADO Project," Ircam,Anlysis/Synthesis Team, 1Pl Igor, Stravinsky, 75001, Paris, France, Analysis report V 1.0, 23rd April,2004.

[3] ITC Infotech India Ltd. -FAQ. http://www.itcsra.org. [Online]. http://www.itcsra.org/sra_faq_index.html

[4] S. & Bregman, A. McAdams, "Hearing musical streams," *Computer Music Journal republished in C. Roads & J. Strawn (Eds.) The Foundations of Computer Music, MIT, Cambridge, Mass.*, vol. 3, no. 4, pp. in CMJ 26-43 and later in TF of CM 658-698, 1985.

[5] Hyoung-Gook Kim, Nicolas Moreau, and Thomas Sikora, *MPEG-7 AUDIO AND BEYOND*. west susscx, England : John Wiley & Sons, Ltd, 2005.

[6] Tae Hong Park, "Towards Automatic Musical Instrument Timbre Recognition," Princeton University , New Jersey,USA, Dissertation Report of PhD 2004.

[7] Jean-Julien aucouturier, Francois Pachet, and Mark Sandler, ""the way it sounds":Timbre models for Analysis and retrieval of Music Signals," *IEEE Transactions on Mulitimedia*, vol. 7, no. 6, December 2005.

[8] Hermann von Helmholtz, *On the Sensations of Tone*. New York: Dover, 1954.

[9] Wayne Slawson, *Sound Color*. Berkeley: University of California Press, 1985.

[10] Sigmund Lewarie, *A Study in Musical Acoustics*. Westport, Conn: Greenwood Press, 1981.

[11] S., J. W. Beauchamp, S. Meneguzzi McAdams, "Discrimination of Musical Instruments Sounds Resynthesized with Simplified Spectrotemporal Parameters," *JASA* , vol. 2, p. 104, 1999.

[12]MPEG-7 Audio. ((2005, October)) http://mpeg.chiariglione.org/standards/mpeg-7/audio. [Online]. MPEG-7 Audio. [Online].

[13] Olivier Lartillot, "MIRToolbox 1.4 User's Manual ," Finnish Centre of Exce!ence in Interdisciplinary Music Research, User Manual 30th May 2012.

[14] Petri Toiviainen Olivier Lartillot, "“A Matlab Toolbox for Musical Feature Extraction From Audio”," in *International Conference on Digital Audio Effects*, Bordeaux, 2007.

[15] Petri Toiviainen, Tuomas Eerola Olivier Lartillot, ""A Matlab Toolbox for Music Information Retrieval”," *Springer Data Analysis Machine Learning and Applications,Studies in Classification, Data Analysis, and Knowledge Organization*, 2008.

[16] Wikipedia. (2014, Jan) Mel-Frequency_cepstrum. [Online]. http://en.wikipedia.org/wiki/Mel-frequency_cepstrum

[17] Saurabh Deshmukh and Sunil Bhirud, ""A Hybrid Selection Method of Audio Descriptors for Singer Identification in North Indian Classical Music"," in *Emerging Trends in Engineering and Technology (ICETET)* , Himji, Japan, 2012, pp. 224 - 227.

[18] J. F. Schouten, *"Elusive attributes of timbre"*., 1968 page no. 42.

[19] Anil K Jain and Richard C Dubes, *Algorithms for Clustering Data*. New Jersey, United States of America : Prentice Hall, 1948,1988.

[20] H.-H. Bock, "Clustering methods:- a history of k-means algorithms," in *Selected Contributions in Data Analysis and Classification*. Aachen, Germany: Springer Berlin Heidelberg, 2007, pp. 161-172.