Singer Identification in Indian Hindi Songs using MFCC and Spectral Features

Sarfaraz Masood[1], Jeevan Singh Nayal[2] and Ravi Kumar Jain[3]
[1,2,3]Department of Computer Engineering, Jamia Millia Islamia, New Delhi, India
E-mail: [1]smasood@jmi.ac.in, [2]jeevan.capricorn@gmail.com, [3]ravi92.jmi@gmail.com

***Abstract*—In this paper, there is an attempt to give a simple and efficient solution to the problem of singer identification in Indian Hindi songs using audio samples. The experiment is done using five well known singers. This work uses MFCC and some spectral features which are extracted for each audio sample taken of size 2 seconds. After a set of pre-processing steps, supervised training is done using multilayer feed-forward artificial neural network and back propagation algorithm. The results obtained strongly suggest that the final model is able to perform the task of singer identification in Indian Hindi songs with good accuracy.**

Keywords—Singer Identification; Artificial Neural Network; MFCC; Spectral Features

I. Introduction

With the enhancement in the volume of musical data on internet, the need for organising the metadata in databases, querying and retrieving information about particular singer, song, genre etc. has increased significantly. Extensive research has been done in the area of music information retrieval (MIR) to build such robust categorical and retrieval system. Singer identification (SID) is one of the subfield of MIR that has also grown rapidly in recent times. SID means identifying which singer has sung a particular song or part of song (if sung by multiple singers).

Identification of singer can be done easily by human brain with very little training. Human brain is an intelligent machine which organizes the audio inputs in such a way that they may be recognised with simplicity. However in machine learning domain SID is tough to achieve because of reasons like presence of background instrumental music (considered noise as it will distort the singer voice), audio of other singer(s) in case of multiple singers singing in chorus. Also singing voice is quite unlike from spoken voice.

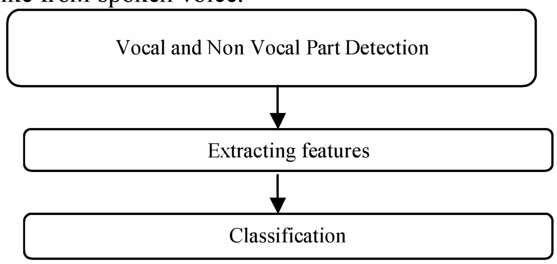


Fig. 1: Flowchart ofSID Process

Overall, a SID system has three steps: 1) Vocal and non vocal part detection, 2) Extracting features and 3) Classification. Classification will contain two phase i.e. the training phase, in which machine is trained by building classifier models using data given, and the testing phase, in which unknown data will be given as input on trained classifier model and the model will classify data into different classes. Portion of the testing data correctly classified will be measure of performance.

In the next section of this paper, research done in the field of singer identification will be discussed briefly followed by description of our problem statement. In the third section features selected will be explained. Fourth section explains about the design of our work, describing all the steps undertaken. Results obtained are discussed in the fifth section which is followed by the conclusion of the paper.

II. Literature Review

Tong Zhang [1] has worked on the singer identification where he extracted audio features namely MFCC and LPC on the song samples of 10-30 seconds from the starting point of singing in the song. It is one of the earliest works in the field of singer identification. Accuracy rate of 80% was achieved in his work. However samples duration considered to achieve this result were comparatively large.

Tsai *et al.* [2] identified singers by mixing the singer's voice with the instrumentals obtained from karaoke VCDs. This work was able to identify regions with no instrumental part in songs. Tsai and lee, [3] worked on the singer identification problem by analysing the spoken data (vocals) instead of the singing data as a whole. Not always, the spoken data of a singer is easily available and it fails in the case where only singing samples are available. However they used only MFCC features and ignored other spectral features of the voice.

S. Deshmukh *et al.*, in [4] proposed a hybrid method to find audio descriptors. They specifically considered the North Indian Classical Singers and focussed on the variations and style of classical music like ragas. They reduced the audio descriptors by taking few stronger descriptors in forward pass and backward pass and removed other unimportant features.

978-1-4673-8587-9/16/$31.00 ©2016 IEEE

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY BOMBAY. Downloaded on October 17,2024 at 07:38:35 UTC from IEEE Xplore. Restrictions apply.

Ying Hu *et al.* [5] worked on automatic singer identification using missing features methods-reconstruction and marginalization. The best results obtained by them under reconstruction, 128 channels was 69.91%. They have worked on clean singing voice obtained from the accompanied singing channel and accompaniment channel using vocal extractor software. But the accompaniment channel is not always easily available in Indian songs. Also they used maximum duration of 4 seconds for audio samples which can be reduced.

D. Dharini *et al.* [6] worked on singer identification by extracting PLP (Perceptual Linear Prediction) features and K-means clustering. They worked on song samples of 20 seconds which is quite long. Spectral features were also not considered for their work. They achieved an overall accuracy of 55.56%.

Purushotam G. Radadia *et al.* in [7] worked on the problem of SID using Mel Frequency Cepstral Coefficients (MFCC) and Cepstral Mean Subtracted (CMS) features and 2ndorder polynomial classifier was employed in their work. Duration of audio samples taken in their work were very large, at least 60 seconds and even concatenation was done to make audio samples up to 300 seconds to enhance performance. Accuracy obtained for MFCC and CMSMFCC was found to be 75.75% and 84.5%. Spectral features were also not considered in their work. Purushotam G Radadia *et al.* extended their work done in [7] in [8] as classification was done both by 3rdorder polynomial classifier (instead of 2ndorder classifier) and Gaussian mixture model (GMM).Rest other parameters were kept same as in [7]. Results obtained with 3rd order polynomial classifier was better with accuracy of 78% and 89.5% for MFCC and CMSMFCC, respectively.

A brief review of literature suggests that very less work has been done in SID on Indian Hindi songs. Also the song sample duration taken in previous work were quite long ranging from 10-60 seconds. Ying Hu *et al.* [5] have worked on samples of 4 seconds for their work. This work is presented on Indian Hindi song samples of 2 seconds. This small duration was chosen by believing that in 2 seconds, we humans also can easily identify a particular singer from a song.

III. Feature Selection

Eight features were selected for our work among others considered. These are mel-frequency cepstral coefficients (MFCCs), root mean square energy, brightness, roughness, spectral rolloff, skewness, flatness, and spectral centroid.

A. MFCC

The mel-frequency cepstralcoefficients (MFCCs) are the set of coefficients that gives the idea of spectral shape of sound. MFCCs are considered to be very effective in music analysis domain. A total of thirteen coefficients were considered for our work. Mirmfcc function is used to compute the value of MFCCs[9].

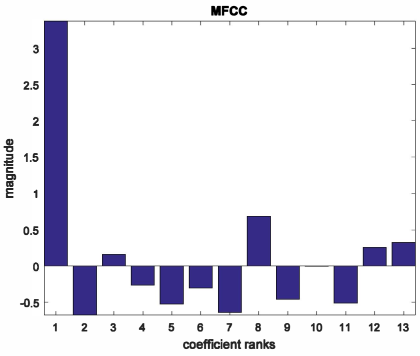


Fig. 2: MFCCs Graph for Song Sample of Arijit Singh

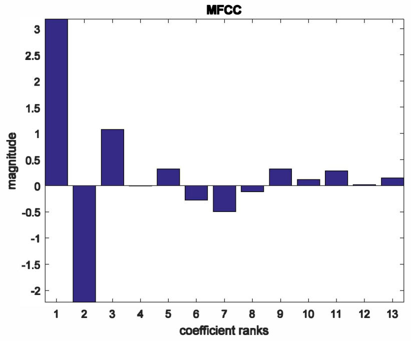


Fig. 3: MFCCs Graph for Song Sample of Kishore Kumar

B. Root Mean Square Energy

Root mean square energy or global energy of a signal (x_{rms}) is calculated by taking square root of average of square of the amplitude of the signal. Mirrms function is used to compute its value[9].

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

$x_i$is the amplitude of signal at ith frame and n is number of frames in the duration of sample. Default frame size is 50milliseconds.

C. Brightness

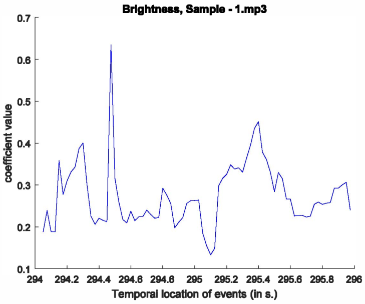


Fig. 4: Brightness Graph for Song Sample of Arijit Singh

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY BOMBAY. Downloaded on October 17,2024 at 07:38:35 UTC from IEEE Xplore. Restrictions apply.

Brightness can be defined as proportion of the energy above a certain frequency value in a sound signal. Default value of frequency value is 1500 Hz. Mir brightness function is used to compute its value[9].

D. *Roughness*

Theory of roughness or sensory dissonance in a sound signal is based on beating phenomenon where sinusoids are closed in frequency as proposed by Plomp and Levalt. It is computed by mirroughness function by taking the peaks of the spectrum and then doing the average of dissonance between all pair of peaks[9].

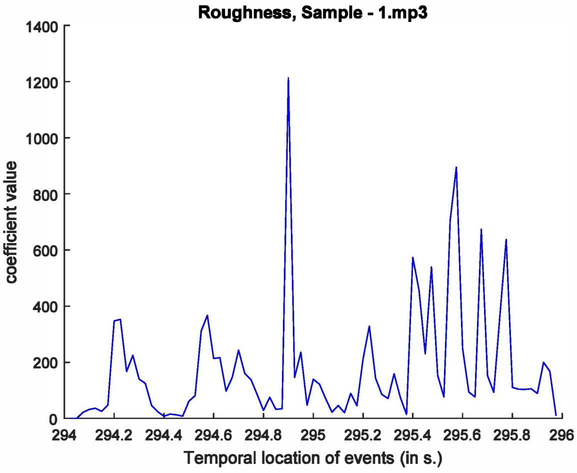


Fig. 5: Roughness Graph for Song Sample of Arijit Singh

E. *Spectral Rolloff*

Spectral rolloff gives the estimation of the amount of the high frequency present in a signal. It gives the frequency under which a certain amount of energy is confined. This amount is fixed to 85% by default. Mirrolloff function is used to compute its value[9].

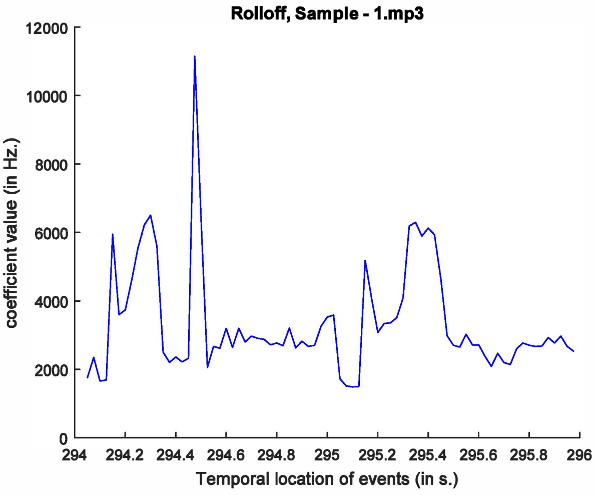


Fig. 6: Spectral Rolloff Graph for Song Sample of Arijit Singh

F. *Skewness*

Skewness gives the idea of the symmetry of the distribution around mean. Positive value of skewness i.e. positively skewed signal means few values are larger than the mean giving a long tail to right of spectrum. Similarly negative value would imply long tail to left. Zero skewness means signal is symmetrical [9].

$$Skew(x) = \mu_3 = \frac{1}{N}\sum_{j=1}^{N}(\frac{x_j - \bar{x}}{\delta})^3 \qquad (2)$$

Coefficient of skewness is calculated by mirskewness function. It is the ratio of skewness to standard deviation raised to third power[9].

$$Coefficient\ of\ Skewness = \frac{\mu_3}{\delta_3} \qquad (3)$$

G. *Flatness*

Flatness gives the idea of smoothness or spikiness of the distribution of the signal. It is computed by taking ratio of geometric mean to arithmetic mean [9].

$$Flatness(x) = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} \qquad (4)$$

Mirflatness function is used to compute flatness.

H. *Spectral Centroid*

Spectral centroid can be defined as geometric centre of a spectrum. It basically indicates about the “centre of mass” of the spectrum. It is calculated as the weighted mean of the frequencies present in the signal. Mircentroid function return the centroid of signal [9].

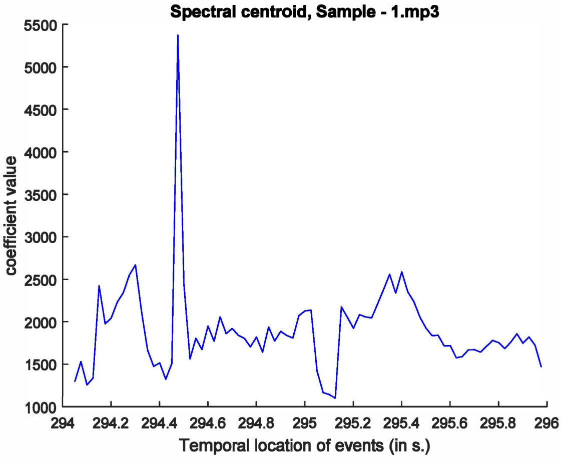


Fig. 7: Spectral Centroid Graph for Song Sample of Arijit Singh

IV. Experiment Design

In the starting phase of our work, Indian Hindi songs samples were used in which we attempted to separate singer voice and instrumental part. But when the singer’s voice and accompanying music gets mixed, it gets difficult to separate them. Also vocals can be separated purely from song only if instrumental version of song is available. Hence, due to the lack of availability of

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY BOMBAY. Downloaded on October 17,2024 at 07:38:35 UTC from IEEE Xplore. Restrictions apply.

instrumentals of most of the Indian Hindi songs, the samples were manually selected from regions within the song which has low intensity background music.

A total of five well known Indian Hindi singers were selected for this work–Arijit Singh, Kishore Kumar, Mukesh, Mahender Kapoor and Narendra Chanchal for this experiment. These singers were considered as this covers well the various genres on Hindi songs. Also these singers have varying voices and singing style from each other as perceived aurally. E.g. Kishore Kumar generally sings melodious songs whereas Narendra Chanchal generally sings songs in high tone.

A total of 66 songs i.e. 11 of Arijit Singh, 22 of Kishore Kumar, 15 of Mukesh, 10 of Mahender Kapoor and 8 of Narendra Chanchal were collected from various websites which provide free access to the songs. Then 50 samples for each singer were taken from their respective songs making a total of 250 samples for the work. After the selection of singers, audio samples and the significant feature vectors, the experiment is performed. The experiment goes through a series of stages as follows:

A. *Stage 1: Selection of Audio Samples*

The audio samples of songs were manually selected to be of 2 seconds due to the intuition that 2 second duration of vocals are enough for us as well to identify the singer.

B. *Stage 2: Extraction of Selected Features*

After the selection of appropriate samples of the said duration, extraction of selected features was done. For this purpose MIRtoolbox1.6.1 was used. MIR toolbox is designed for music information retrieval and to deal with audio data.

C. *Stage 3: Normalisation of Dataset*

After the extraction of selected features, the data was scaled to lie within a small range using simple min-max normalisation.

D. *Stage 4: Training, Validation and Testing of Dataset*

TABLE 1: ANN PARAMETERS FOR EACH EXPERIMENT

Parameter	Exp 1	Exp 2	Exp 3	Exp 4
Learning rate	0.3	0.3	0.3	0.3
Epochs	450	450	450	450
Hidden neurons	10	17	10	17
Training function	Scaled Conjugate Gradient (scg)	Scaled Conjugate Gradient (scg)	Levenberg Marquardt (LM)	Levenberg Marquardt (LM)
Activation Function	Sigmoidal	Sigmoidal	Sigmoidal	Sigmoidal

Multi-layered feed-forward artificial neural network was implemented using Neural Network Toolbox available in MATLAB Software. Of the collected 250 song samples 80% data i.e. 200 samples were used for training, 10 samples for validation, and rest 40 samples were used for testing the trained network.

A set of four separate experiments were conducted by varying the training algorithms and the number of hidden layer neurons while keeping rest of the parameters same. These experiments helped in identifying the most suitable ANN architecture for this problem.

V. RESULTS

ANN is trained in each experiment according to parameters in Table 1. The saved networks are tested on rest 40 samples and results are obtained. The true positive(TP), true negatives(TN), false positives(FP), false negatives(FN), accuracy(AC), precision(PR), recall(RC) and f-measure(FM) are calculated for each class (for each singer i.e. S1, S2, S3, S4, and S5) in each experiment and are shown in Table 2.

Overall accuracy for each experiment is shown in Fig 8. By comparing the results obtained in each experiment, best accuracy of 92.5 is achieved with training algorithm Levenberg Marquardt (LM) and number of hidden layer neurons 17.

The classified samples are shown in the confusion matrix for experiment four in Fig. 9 as follows:

Confusion matrix shows how the classification results are distributed over the whole set of classes. Target class is the actual class of sample whereas output class is the class obtained after classification. Green blocks contain the correctly classified samples in their own class and reds are the samples classified into the wrong class. Samples are considered to be correctly classified if their target class and output class are same otherwise sample is said to be misclassified. The box in blue shows the classification accuracy of 92.5% and 7.5% misclassification for overall experiment. Overall accuracy is obtained by combining the results of correctly classified samples of each class. The grey box in the bottom shows the classification and misclassification accuracy of the samples whose target class is i^{th} column. For e.g. in second column, a total of 9 samples belong to target class two. Of which 7 are correctly classified by network and 2 are misclassified in class three. Thus percentage of correctly classified sample for class two among the total sample of class two is 77.8%. This percentage is also termed as recall value. Similarly the grey box in the right side shows the classification and misclassification accuracy of the samples whose output class is i^{th} row. For e.g. second row shows the samples classified in class two. Of which 7 samples actually belongs to class two and 1 sample is of another class misclassified in class two. Thus percentage of correctly classified samples among the total samples classified in class two is 87.5% which is also called as precision value for class two.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY BOMBAY. Downloaded on October 17,2024 at 07:38:35 UTC from IEEE Xplore. Restrictions apply.

TABLE 2: SIGNIFICANT RESULTS OBTAINED ON VARYING NETWORK PARAMETERS FOR EACH EXPERIMENT

	Experiment 1					Experiment 2					Experiment 3					Experiment 4				
	S1	*S2*	*S3*	*S4*	*S5*	*S1*	*S2*	*S3*	*S4*	*S5*	*S1*	*S2*	*S3*	*S4*	*S5*	*S1*	*S2*	*S3*	*S4*	*S5*
TP	11	7	5	6	6	10	6	5	7	7	11	6	6	8	3	9	7	5	7	9
TF	24	28	30	29	29	25	29	30	28	28	23	28	28	26	31	28	30	32	30	28
FP	0	1	3	0	1	1	0	3	0	1	1	1	2	1	1	0	1	2	0	0
FN	0	3	1	1	0	0	1	1	2	1	0	3	2	0	1	0	2	0	1	0
AC%	***100***	***89.7***	***89.7***	***97.2***	***97.2***	***97.2***	***97.2***	***89.7***	***94.6***	***94.6***	***97.1***	***89.5***	***89***	***97.1***	***94***	***100***	***92.5***	***94.9***	***97.3***	***100***
PR%	100	87.5	62.5	100	85.7	90.9	100	62.5	100	87.5	91.7	85.7	75	88.9	75	100	87.5	71.4	100	100
RC%	100	70	83.3	85.7	100	100	85.7	83.3	77.8	87.5	100	66.7	75	100	75	100	77.8	100	87.5	100
FM%	100	77.8	71.4	92.3	92.3	95.2	92.3	71.4	87.5	87.5	95.6	75	75	94.1	75	100	82.4	83.3	93.3	100

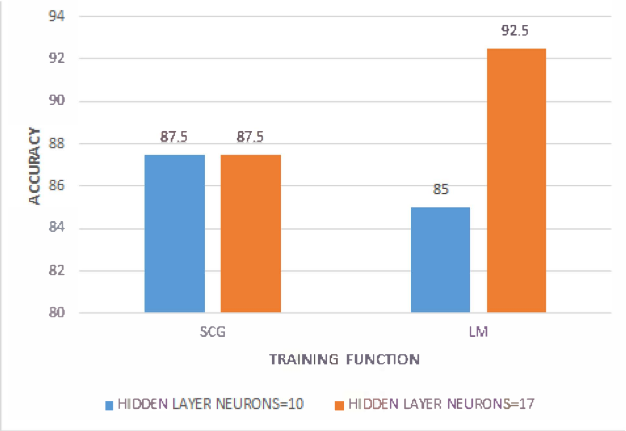
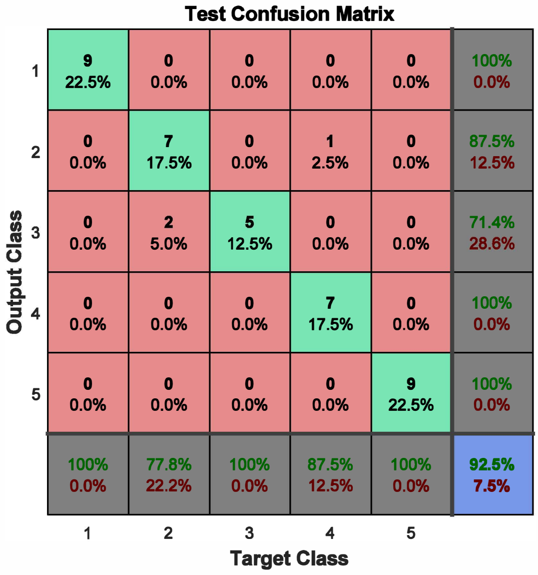


Fig. 8: Accuracies for Each Experiment



	1	2	3	4	5	
1	9 22.5%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
2	0 0.0%	7 17.5%	0 0.0%	1 2.5%	0 0.0%	87.5% 12.5%
3	0 0.0%	2 5.0%	5 12.5%	0 0.0%	0 0.0%	71.4% 28.6%
4	0 0.0%	0 0.0%	0 0.0%	7 17.5%	0 0.0%	100% 0.0%
5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	9 22.5%	100% 0.0%
	100% 0.0%	77.8% 22.2%	100% 0.0%	87.5% 12.5%	100% 0.0%	92.5% 7.5%

Fig. 9: Confusion Matrix for the Test Data in Exp 4

VI. CONCLUSION

By analysing the results obtained, we have significantly attempted to solve the problem of singer identification in Indian Hindi songs considering five well known singers with an overall 92.5% accuracy using multi-layered feed-forward artificial neural network. We achieved the results using 2 second samples. This result shows that MFCC and spectral features are proved to be significant in the task of singer identification.

The work can be extended to be performed on large set of song sample dataset and also for various regional singers. The model can also be used for auto playlist generator of songs based on identification of singers.

REFERENCES

[1] Tong Zhang, “Automatic singer identification, ” Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on, 2003, pp. I-33-6 vol.1.

[2] W. H. Tsai and H. P. Lin, “Background Music Removal Based on Cepstrum Transformation for Popular Singer Identification, ” in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1196-1205, July 2011.

[3] W. H. Tsai and H. C. Lee, “Singer Identification Based on Spoken Data in Voice Characterization, ” in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 8, pp. 2291-2300, Oct. 2012.

[4] S. Deshmukh and S. G. Bhirud, “A Hybrid Selection Method of Audio Descriptors for Singer Identification in North Indian Classical Music, ” Emerging Trends in Engineering and Technology (ICETET), 2012 Fifth International Conference on, Himeji, 2012, pp. 224-227.

[5] Y. Hu and G. Liu, “Automatic singer identification using missing feature methods, ” Multimedia and Expo (ICME), 2013 IEEE International Conference on, San Jose, CA, 2013, pp. 1-6.

[6] D. Dharini and A. Revathy, “Singer identification using clustering algorithm, ” Communications and Signal Processing (ICCSP), 2014 International Conference on, Melmaruvathur, 2014, pp. 1927-1931.

[7] H. A. Patil, P. G. Radadia and T. K. Basu, “Combining Evidences from Mel Cepstral Features and Cepstral Mean Subtracted Features for Singer Identification, ” Asian Language Processing (IALP), 2012 International Conference on, Hanoi, 2012, pp. 145-148.

[8] P. G. Radadia and H. A. Patil, “A Cepstral Mean Subtraction based features for Singer Identification, ” Asian Language Processing (IALP), 2014 International Conference on, Kuching, 2014, pp. 58-61.

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

[10] Matlab Documentation.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY BOMBAY. Downloaded on October 17,2024 at 07:38:35 UTC from IEEE Xplore. Restrictions apply.