Indian J.Sci.Res. 17(2): 381-386, 2018 ISSN: 0976-2876 (Print)
ISSN: 2250-0138(Online)

AUTOMATIC FRAMEWORK OF MUSIC RINGTONE EXTRACTION FROM TOLLYWOOD SONGS

[1] Ramadevi Y ,[2] Prathima T ,[3] Apsar Sh
[1]Professor, Department. of Computer Science and Engineering, CBIT, Hyderabad, Telangana
[2]Research Scholar, Department. of Computer Science, JNTU Hyderabad, Telangana
[3] PG student, Department. of Computer Science and Engineering, CBIT, Hyderabad, Telangana

***Abstract**:* An automatic framework is used to extract the ringtones from music automatically. In this, song is considered as the grouping of segments of music such as intro, chorus, verse, bridge, outro. Mostly the ringtone will be the 'chorus' or 'intro' segments of music. The process of manually checking each song and cropping specific parts of the song is a tedious process. Western music and Bollywood songs are widely used for ringtone extraction. The accuracy is not stable for different genres of the songs such as hip-hop, ghazal etc work, for automatic extraction of ringtone, beat tracking is done by using Simon Dixon BeatRoot followed by feature extraction process as the audio data lies within beats. Songs from Tollywood (regional) were used for experimentation. SVM and Naïve Bayes classifiers are used for comparisons. The class labels are predicted based on training samples. The accuracy gained by SVM is 62.9% with 11093 beat data and the Naïve Bayes classifier gained 75% accuracy with the same beat data. In the two datasets of experimentation Naïve Bayes performed better than SVM.

***Index Terms* -** BeatRoot, Feature Extraction, Classification, Segment Boundary Detection.

I. Introduction

Automatic music extraction is very useful in significant fields. In this a song is taken as input and divided into segments which are considered as meaningful regions such as verse or chorus. The structure of song is usually divided into intro, verse, chorus, outro, etc. Ringtone is an audio file played on mobile phones to indicate an incoming call. Ringtones are popular because in a crowd of people with cellular handsets it is easy to identify easy whose phone is ringing.

Ringtones and ring-music bring more fun when people make calls and it remains as labor intensive work, people need to listen each and every song to set the starting point and ending point for a clip with in audio file, then extract the segment [1]. In this paper our main goal is to extract the ringtone automatically by detecting the boundaries of segments correctly with good accuracy.

Song forms are made up of a number of sections that may or may not be repeated within the same song. Some of the popular song structures are strophic (AAA) form, AAB (12 bar blues) form, AABA song form, AB or verse/chorus song form, ABC song form or verse/chorus/bridge song form. South Indian music song forms are very similar to western music forms.

A. Genres of Telugu songs

In music genre refers to musical style. Some of the popular genres of Indian music are [2]:

- Classical: The composition of classical music is based on ragas, which are the scales of seven basic notes such as sa, re, ga, ma, pa, dha and ni. The commonly played musical instruments of this genre includes sitar, surbahar, sarod, sarangi, santoor, bansuri, pakhavaj and tabla.
- Ghazal: According to Arabic dictionary the word ghazal means 'talking about woman', it is generally a poem consisting of five to fifteen couplets known as 'shers'. The ghazals became a part of the Indian music with the invasion of Mughals.
- Pop- Indipop music is a hybrid of Indian and western musical traditions.
- Devotional: Bhakti or devotion, constitutes an important part of Hindu religious practice. The broad sweep of devotional music includes chants and readings of scriptures such as the Vishwasahasranam, Shivamahimmah stotra, Bhagavad Gita and holy mantras, such as Om Namah Shivaya.
- Folk: India folk music owes its origins to the villages, which represents the folklore and lives of the villagers
- Tribal: Indian tribal music is originated from the inhabitants of the hilly regions and they are composed among the tribals of northeast India and southern states.

Folk and tribal music was composed and performed in order to celebrate a particular festival or to deliver a message.

B. Structure of Indian song

[1]Corresponding Author

Telugu is the one of the most spoken language of South India. Though the Telugu songs from the movies are very popular, but the fact is that their rich heritage lies in Carnatic renderings. According to music connoisseurs, there is a sweetness and melody in Telugu songs that is prevalent in every genre of compositions. Compared with the structure of western music and English songs, the structure of Telugu music is different.

The structure of Telugu music consists of segments such as intro, pallavi, anupallavi, bridge, charanam, outro. The segment anupallavi and Charanam (similar to chorus of Western music) are separated by the Bridge [3].

II. Related Work

Fengyan Wu, Shutao Sun, Weiyao Xue [1] proposed a framework to extract ringtones from music automatically by using the theory of musical structure analysis and machine learning algorithms. They used SVM for the segment boundary detection and random forest classification to detect chorus segments. By conducting experiments they concluded that Chroma feature has better effect than MFCC. The combination of two features also did not enhance the effect. Precision in Boundary detection by SVM affects the process of ring tone extraction.

Sher Muhammad Doudpota, Sumanta Guha [4] proposed a system to automatically locate and extract full songs from digitized movies. Their work is focused on Bollywood movies. Their test results indicated high precision and high recall. They used SVM classifier to classify each frame as either music or non-music based on audio features.

Brian McFee, Daniel P.W. Ellis [5] describes a supervised learning algorithm to predict the segment boundaries. They also developed a latent structural repetition feature to detect repetitive sections of song with self similarity matrix. Experimental results demonstrate that the method efficiently integrates heterogeneous features, and improves segmentation accuracy. All these methods are implemented in Python with the Librosa package 2. All signals were down sampled to 22KHz mono, and analyzed with a 93ms window and 3ms hop. MFCCs are generated from 128 Mel bands with an 8KHz cutoff. They took 32 MFCCs and 12 log-magnitude chroma bins; repetition features are calculated with 2pt nearest neighbors under a standardized Euclidean metric, median-filtered with a window of width 7, and reduced to 32 dimensions each. Including the four time-stamp features, the combined representation has dimension D = 112. Beats were detected by the median-percussive method. They evaluated predicted segmentations on two publicly available datasets: Beatles-ISO 179 songs by the Beatles and SALAMI-free 253 songs from the SALAMI dataset which are freely available on the Internet Archive. They used Ordinal Linear Discriminant Analysis (OLDA) technique to learn a feature transformation which is optimized for musical segmentation, or more generally time series clustering. The proposed methods achieved higher accuracy for boundary detection at 0.5s resolution, which can be attributed to the increased accuracy afforded by the beat-synchronous and median-filtered repetition features.

Jouni paulus, Meinard Muller, Ansii klapuri [6] gives an overview of state-of-the art methods for computational music structure analysis. For segmenting and structuring music audio they identified three conceptually different approaches such as repetition-based, novelty-based, and homogeneity based approaches. They extracted the Mel Frequency Cepstral Coefficient features by using Discrete Cosine Transform, Chroma and rhythmogram features. They addressed different issues in the context of music structure analysis, while discussing and categorizing the most relevant and recent articles in this field. They used Self Distance Matrix (SDA) in order to detect the self similarities. In this system, they mainly focused on Western popular music.

Xi Shao, Namunu C Maddage, Changsheng Xu, Mohan S Kankanhalli [7] presented an automatic music summarization based on music structure analysis. From the audio signal, they extracted the note onset that represents the time tempo of the song and they performed music structure analysis based on the tempo information. To detect the repeated sections of song they used similarity analysis methods.

Antti Eronen[8] presented a computationally efficient method for detecting a chorus section in popular and rock music. The method utilizes a distance matrix representation that is obtained by summing two separate distance matrices calculated using the mel-frequency cepstral coefficient and pitch chroma features. The benefit of computing two separate distance matrices is that different enhancement operations can be applied on each. An enhancement operation is found beneficial only for the chroma distance matrix. This is followed by detection of the off-diagonal segments of small distance from the distance matrix. These detected segments, are processed using image processing filters in a neighborhood of the distance matrix surrounding the initial chorus section. The final position and length of the chorus is selected based on the filtering results.

Saurabh H. Deshmukh, Dr. S.G.Bhirud [9] analysed majority of the contributions and proposed the best suitable audio feature descriptor and the classifiers to be used for the problem of Singer identification in North Indian classical music. This type of music requires special attention and careful selection of feature extractors because of the involvement of accompanying instruments and melodic structure of the raga. They reported 70% efficiency when they combined RMS energy, Brightness and Fundamental Frequency by using K-means clustering.

Theodoros Theodorou, Iosif Mporas, Mporas, Nikos Fakotakis[10] presented an overview of automatic audio segmentation. They made the detection of silence and breathing noise by using Gaussian Mixture Models (GMM). For the segmentation stage they used two main approaches a. distance based techniques and b. model based techniques. They used Mel Frequency Cepstral Coefficients and Zero Crossing Rate (ZCR) for feature extraction with the frame length of 10 msec to 25 msec and frame step between successive overlapping frames is 50% of the frame length. They introduced hybrid technique, which results in resegmentation algorithm performing operations on both distance based and model based. They used different types of datasets in which they achieved higher accuracy of 98% for detecting the male and female speakers.

Mathew Thomas, Y. V. Srinivasa Murthy and Shashidhar G. Koolagudi [3] performed analysis on detecting largest possible repeated patterns in Indian audio songs which will be ringtones. They analyzed the structure of Indian songs into parts such as intro, pallavi, anupallavi, charanam, bridge, outro. For repeated pattern detection they used Dynamic Time Warping (DTW) method and spectral affinity using spectral features such as MFCCs and Modulation Spectral Features (MSFs). Jitter is also used for determining cycle to cycle pitch variation. They observed that DTW time complexity is very high for large music files. They found the accurate results with the combination of MFCCs, MSFs and Jitter.

Namunu C. Maddage [11] presented Automatic structure detection for popular music in order to detect the chord and the singing voice boundaries in music. Dataset of 50 popular English songs are used which are sampled at 44.1KHz with 16 bits per sample. In order to detect the chords they modeled 48 chords with Hidden Markov Models where 35 songs are for training and 15 songs are used for testing. They achieved the average frame accuracy of 80.87% for chord detection. SVM with RBF kernel is used to classify frames into vocal or instrumental class. 56.33% average accuracy is gained for chorus detection.

Beth Logan and Stephen Chu [12] addressed an approach for Music Summarization using key phrases in which they worked to extract the chorus section, the most repeated section and the most memorable section. They extracted the MFCC features from the song and to discover the song structure they used two techniques, one is clustering and the other is Hidden Markov Model(HMM). Bottom up clustering is used to group the similar features whereas unsupervised Baum-Welsh training is used to train HMM. Sampled the audio data at 16KHz by dividing the audio into 25.6ms windows at a 10ms frame rate. By using these, structure heuristics are applied to choose the key phrase. After evaluating 18 beatless songs by ten users they concluded that clustering achieved 95% confidence than random whereas HMM method was not better than random.

Feng Li, You You, Yuqin Lu, and YuQing Pan [13] proposed an automatic segmentation method by combining SVM classification and audio self similarity approach. They considered 100 Chinese and 100 English popular music segmented songs and manually assigned labels to each and every segment. Audio is sampled at 22050 KHz and the length of the frame at 46ms. They used MFCC, first order derivative (ΔMFCC) and Sub-frequency band spectral energy ratio(SFR) to train binary SVM classification model. They achieved 88.67% cross validation accuracy by considering gamma as 0.125.

Ewald Peiszer, Thomas Lidy, Andreas Rauber [14] presented an algorithmic approach for automatic audio segmentation. Their main aim is to extract the song structure, semantic labels like verse, chorus, bridge etc. They focused on segment boundary detection and structure detection by sampling the audio 22050 Hz with non overlapping frames. They used Corpus data set of 94 distinct songs and 15 songs for the test set which consists of different genres like pop, folk, rap, dance, etc. They introduced a new XML format for displaying ground truth annotations and segments like start and end times, labels and alternative labels. 40m MFCC coefficients are used with the frame size of 2^{14} frames which are beat synchronized for better accuracy. They noted that the proposed algorithm extracts a finer structure than the ground truth corpus.

Jouni Paulus, Anssi Klapuri[15] proposed a method for extracting the meaningful music labels such as “chorus” or “verse” by using Fitness measure and Greedy search algorithm. They used three different data sets such as TUT structure 07(557 radio-play pieces), UPF Beatles (174 songs), RWC pop(100 pieces with 80% of Japanese music and rest American chart hits). Feature extraction is carried out by calculating MFCC, chroma in 92.9ms frames with 50% frame overlap and rhythmogram uses frame length up to several seconds with step size of 46.4ms. After the evaluation they observed that F-measure is quite similar in all datasets, but the recall and precision differs greatly, the result is over segmented in UPF Beatles dataset where as it is under segmented in RWC pop. The optimization problem is solved with greedy search algorithm and the segments are assigned with meaningful labels such as verse or chorus.

III. Methodology

The methodology of automatic ringtone extraction is carried out as shown in figure 1.

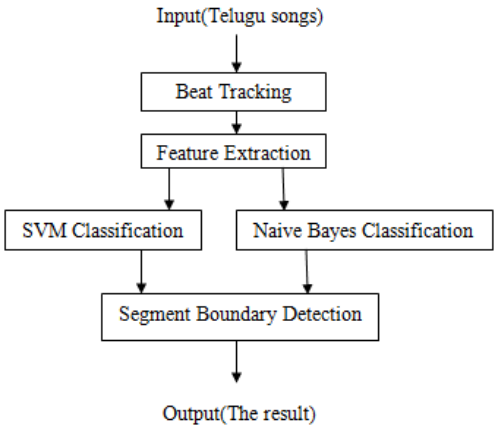

Fig. 1: Architecture of the proposed system

The proposed system for the automatic ringtone extraction consists of major steps such as beat tracking, feature extraction, classification and segment boundary detection. These processing steps are detailed below:

A. Beat Tracking

To extract the ringtones from the Telugu songs, two sample datasets are prepared. One is the dataset with full songs manually cropped into music and non music segments and the other dataset with music and non music pieces collected from different genre songs. These datasets are sampled at the rate of 22050 Hz. The input files which are taken from .mp3 (compressed) format, are converted into .wav (decompressed) format. To avoid the conflicts arising in the sound, the stereo signal is converted into mono signal. The .wav files are given as input to Simon Dixon BeatRoot [16] system to detect beats. These detected beat times are taken under the duration of 500 ms and the corresponding audio samples are extracted from the starting and ending time of each beat.

B. Feature extraction

The audio samples from each beat are considered for the feature extraction [17]. The window size is considered as 50 ms and a step size of 25 ms (50 percent frame overlap). In this the audio signal is divided into short term windows (frames). Mel Frequency Cepstral Coefficients (MFCC) and Chroma features are extracted from these audio sample beat data. MFCC is used for speech recognition whereas Chroma features reflect the harmonic and melodic information. The above mentioned 25 features i.e., 13 MFCC coefficients and 12 Chroma bins are extracted from the frames at each beat. Means of all the 25 features across frames are calculated. Averages of 25 dimensional data for the audio data at each beat are provided for classification. Corresponding labels of music and non music data are assigned per each beat.

C. SVM and Naive Bayes classification

The output from the feature extraction is taken as input for the classification. SVM classification [18-19] is done by dividing the two third of data for training and one third of data for testing. SVM classification is done using Radial Basis Function (RBF) kernel. Class label one is assigned for Music data and class label zero is assigned for non music data. The labeled data is also used to train Naive Bayes Classifier [20] in order to compare the performance of both classifiers. Hold out method is used to divide the dataset into training set and test set with 0.3.

D. Segment Boundary Detection

Labels are predicted after training the classifier based on test data. The continuous labels are combined as one segment and labeled as either music or non music based on the class label of the consecutive beats. If a label is detected in between the continuous music or non music labels it will be clubbed into that continuous label and considered as one segment.

IV. Results

Experiments are done based on Telugu songs from Tollywood movies. One dataset with 20 full songs is manually divided into music and non music pieces. These 20 full songs consists of 75 music pieces and 90 non music pieces. Second dataset consists of 39 music pieces and 29 non music pieces selected from the different songs. Different genres include pop, classical, ghazal, devotional, and folk. Non music is mixed with singing voice and musical instruments. SVM achieves an efficiency of 62.9% on the full songs dataset whereas it gains 62% accuracy on music pieces dataset. Naive Bayes gains the maximum accuracy of 75% on full songs dataset whereas it gains 76% accuracy on second dataset. Comparison of these two classifiers on both the datasets are shown in figure 2 and figure 3 respectively.

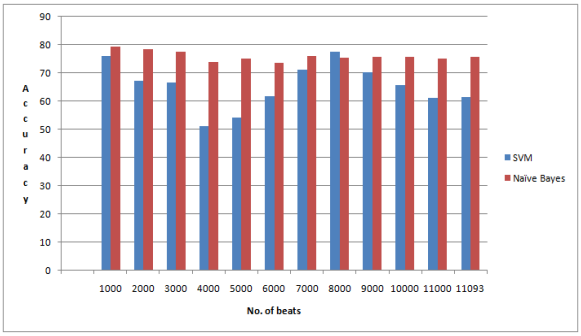

Fig 2: Performance comparison between SVM and Naive Bayes with full songs dataset

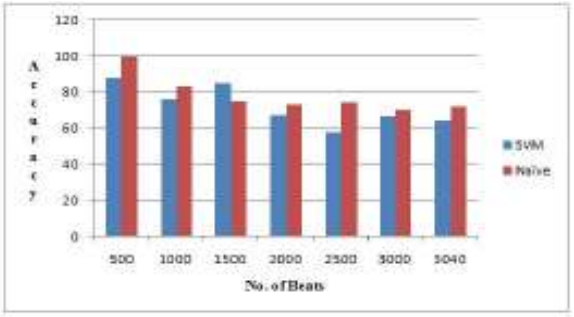

Fig 3: Comparison between SVM and Naive Bayes with second dataset

From these graphs the X-axis represents the number of beats whereas Y-axis represents the efficiency achieved. In figure 3, Naive Bayes shows 100% accuracy for 500 beat sample data as it results to overfitting problem. In the Naive Bayes hold out method is used which randomly selects the training samples and testing samples which results to the over fitting problem. The empirical results show Naive Bayes performs better than SVM.

V. Conclusion and Future work

The proposed work is carried out on Telugu songs to extract the ringtones automatically. Two classifiers are trained and tested to compare the performance. This framework shows that Naive Bayes performs well than SVM. It is also observed that SVM consumes more time for classification than Naive Bayes. The drawback of this system is that the accuracy is not stable for different genres of songs.

Future work is intended to build ensemble classifier for multi genre Tollywood songs to enhance the accuracy across genres.

References

[1] Fengyan Wu, Shutao Sun, Weiyao Xue, IEEE. "Automatic Extraction of Popular Music Ringtones Based on Music Structure Analysis", IEEE ICIS 2016, June 26-29, 2016, Okayama, Japan.

[2] Artists Pages, www.artistspages.org/types _of_indian_music.htm.

[3] Mathew Thomas, Y. V. Srinivasa Murthy and Shashidhar G. Koolagudi, "Detection of Largest Possible Repeated Patterns in Indian Audio Songs using Spectral Features", 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), 978-1-4673-8721-7/16(c)2016IEEE.

[4] Sher Muhammad Doudpota, Sumanta guha, "mining movies to extract song sequences", MDMKDD'11, August 21, 2011, San Diego, CA, USA.

[5] Brain McFee, Daniel P.W.Ellis, "Learning to segment songs with ordinal linear discriminant analysis," 2014 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP).

[6] Jouni Paulus, Meinard Muller, Ansii klapuri, "Audio based music structure analysis", 2010 International Society for Music Information Retrieval, Utrecht.

[7] Xi shao, namunu Maddage, Chungsheng xu, mohan kankanhalli, "Automatic music summarization based on music structure analysis", 2012 , Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore.

[8] Antii Erronen, "Chorus detection by combined use of MFCC and chroma features and image processing filters", Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux, France, September 10-15, 2007.

[9] Saurabh H. Deshmukh, Dr. S.G.Bhirud, "Analysis and application of audio features extraction and classification method to be used for North Indian Classical Music's singer identification problem", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014 .

[10] Theodoros Theodorou, Iosif Mporas, Nikos Fakotakis,"An Overview of Automatic Audio Segmentation", International journal of Information Technology and Computer Science(IJITCS), vol.6, no.11, pp.1-9, 2014. DOI: 10.5815/ijitcs.2014.11.01

[11] Namunu C.Maddage, "Automatic Structure Detection for Popular Music", Institute for Infocomm Research, 1070-986X/06/$20.00 (c) 2006 IEEE.

[12] Beth Logan and Stephen Chu, "Music summarization using key phrases", International Conference on Acoustics, Speech and Signal Processing, Cambridge Research Laboratories, 2000.

[13] Feng Li, You You, Yuqin Lu, Yuqin Lu, YuqingPan," An Automatic Segmentation Method of Popular Music Based on SVM and Self-similarity ", Springer International Publishing Switzerland 2015 Q. Zu et al. (Eds.): HCC 2014, LNCS 8944, pp. 15–25, 2015.

[14] Ewald Peiszer, Thomas Lidy, Andreas Rauber, "Automatic Audio Segmentation: Segment Boundary and Structure Detection in Popular Music", Vienna University of Technology, Austria,

2007.

[15] Jouni Paulus, Anssi Klapuri, “Music Structure Analysis Using a Probabilistic Fitness Measure and a Greedy Search Algorithm”, Columbia University, IEEE transactions on audio, speech, and language processing, vol. 17, no. 6, august 2009, Las Vegas, USA.

[16] Simon Dixon, "Evaluation of the Audio Beat Tracking System BeatRoot" Centre for Digital Music Department of Electronic Engineering Queen Mary, August 2007 University of London Mile End Road, London E1 4NS, UK .

[17] Theodoros Giannakopoulos and Aggelos Pikrakis, “Introduction to Audio Analysis”, 2014 Elsevier Ltd.

[18] Wu Fengyan, “Singing Voice Detection of Popular Music Using Beat Tracking and SVM Classification”, International Conference on Computer and Information Science (ICIS 2015), June 28-July 1 2015.

[19] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, "A Practical Guide to Support Vector Classification", Department of Computer Science National Taiwan University, Taipei 106, Taiwan, May 19, 2016.

[20] Irina Rish, “An empirical study of the naive bayes classifier”, In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, T.J. Watson Research Center, 30 Saw Mill River Road, Hawthorne, 2001.