Scientific Journal of Impact Factor(SJIF): 3.134
e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 2,Issue 6, June -2015

MSqueezer: Web Crawler Based On Lyrics & Track Extraction From Hindi Songs

Miss Darshita S. Pathak[1],Prof.Girish H. Mulchandani[2]
[1] Computer Engineering,V.V.P.Engineering College, Rajkot
[2] Asst. Prof.Computer Engineering,V.V.P.Engineering College, Rajkot
Gujarat Technological University

Abstract *— Web Crawler works on keyword based searching, same way searching of song is based on Title of song only. With Music Information Retrieval MSqueezer will extract out Lyrics from Mp3 Song Of Hindi Songs. The technique to crawl lyrics will be based on mainly the natural language processing where lyrics from the song will be extracted by using novel string matching algorithm so the retrieval of lyrics from the song will be easier and efficient. Each song falls within one category that define as Genre Of music. By listening song we can categorize that song falls within which genre ex Pop or Jazz. But At time of crawling it again based on search based for song that generally contains word of genre of webpage, but not categorize song by its spectral information. The technique of classification of harmonic wave for Genre is combination of support vector machine where regression analysis with variable stored already on pitch level repository is to be matched & of information retrieval to crawl to match that pitch and the result will be retrieved. These both research will helpful in area of music information retrieval research as well as at commercial and real time application fields too.*

Keywords- *Music Information Retrieval, MFCC , Lyrics , Metadata, Genre, Information Retrieval, Crawling.*

I.INTRODUCTION

Music information retrieval (MIR) is the interdisciplinary science of retrieving information from music. MIR is a small but growing field of research with many real-world applications. Those involved in MIR may have a background in musicology, psychology, academic music study, signal processing, machine learning or some combination of these. Music Information Retrieval (MIR) has been defined by Stephen Downie as 'a multidisciplinary research endeavor that strives to develop innovative content-based searching schemes, novel interfaces, and evolving networked delivery mechanisms in an effort to make the world's vast store of music accessible to all'. And this MIR mainly deals with Music Content, Music Similarity and Music Psychology areas[1]

Generally when we search song in any system either it will be in music player or in any crawler we generally retrieve the song if our search keyword matches with the information available in webpage that is crawled by search engine. This search mechanism is generally based on the system to crawl Title and at the next to metadata (if it stored in song, but if not stored it will not crawl it even).

By this system the problem arise when we know the wording of song that appears in between the song that is either in "Antras" of the song we will not retrieve the proper song that we are looking forward to find. And if we get the results that will be Lyrics of song that is stored in the webpage, we do not get the song in mp3 format. Also sometimes problem of specific track based songs we look to find is not retrieved i.e. if I look to find only Hip-Hop Genre based songs the searching to that kind of genre songs are not so efficient thus it leads to thesis motivation for us.

II. OBJECTIVES

By the thesis motivation we came to know that if make crawler where searching can be based on the lyrics of song and track extraction can be based on classification of harmonic waves, then we can lead to solve the problem. The technique to crawl lyrics will be based on mainly the natural language processing where lyrics from the song will be extracted by speech to text conversion method with adaption various techniques to smooth the pitch and voice tone of different singer so the retrieval of lyrics from the song will be easier and efficient. The technique of classification of harmonic waves is combination of support vector machine where regression analysis with variable stored already on pitch level repository is to be matched & of information retrieval to crawl to match that pitch and the result will be retrieved. Thus finally to make the extractor to extract out same as like the juicer extract the main ark from fruit ,thus same way the proposed system will extract out main ark of song that is Lyrics and Genre hence the name of system is

given unique name as MSqueezer : Web Crawler Based on Lyrics and Track Extraction From Hindi Songs.Main Objectives are as follows :

1) MSqueezer will be useful to find song by its "**Lyrics**".
2) Finding particular "**Genre/Track type**" songs will be easier by its music only , because it will result to "**Classify**" songs to different genre type.
3) Will improve efficiency of "**Searching Song**" because it's totally based on lyrics so which song you look to find will get easily

III.PROPOSED SYSTEM

Module 1: Lyrics Extraction

Step- 1:User simply fires the query that contains words appear in the Hindi song.
Step- 2:The request in the query form will be given to crawler means to MSqueezer
Step- 3:The MSqueezer will apply the text mining to crawl the exact song .
Step- 4 :As the song is to be find out ,from the repository it will be given as result to the user in mp3 format.

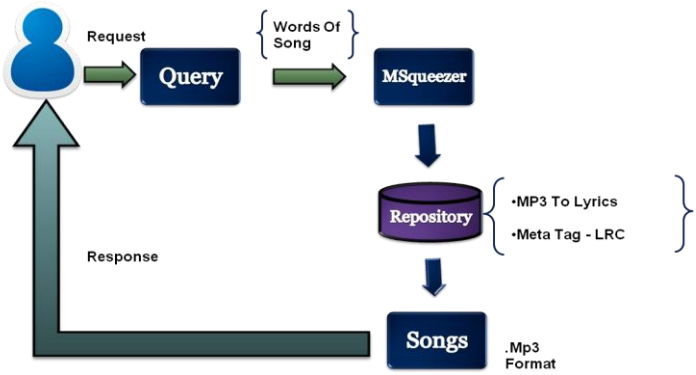

Figure 1 Proposed System for Lyrics Extraction

Module 2: Track Extraction

Step -1 : Hindi Song in form of mp3 is given as input, Java works good with lose less WAV format, thus MP3 to WAV conversion is performed.
Step -2 : The WAV format song is processed under extraction of audio feature, mainly MFCC- Mel Frequency Cepstral Co-efficient is extracted and stored in XML File.
Step -3 : WAV song's overall MFCC value is retrieved and on it Derivative Sum of that MFCC value is calculated.
Step -4 : Based on MFCC value, Range Classifier is applied which will assign that song to particular track type.
Step -5 : Now when user look for any particular genre, based on available genre and song stored in repository, search is matched and gives output songs that is relevant to search track type.

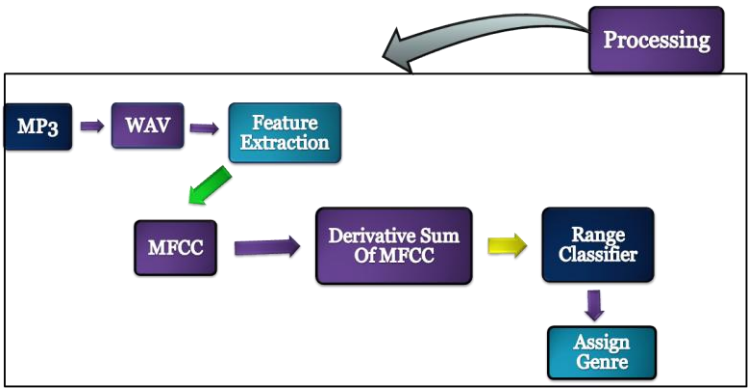

Figure 2 Proposed System for Track Extraction

IV.DATASET

Dataset that is used for implementation of track classification is been created by me, because of as such its novel approach and reference paper[13] contains GTZAN dataset which is of Roman and Latin Language songs, thus of 200 Hindi song dataset is created by me which contains, title of song, movie name, artist of song,mfcc value and year of song. Hindi Songs includes all age of filmy song starting from time of Dilip Kumar to latest ,all kind of age's songs are covered in dataset, also as far as bhajans which are of Hindi language is also included in dataset.

IV.I.MODULE-TRACK EXTRACTION

- **MFCC**

The Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. The name Mel comes from the word melody to indicate that the scale is based on pitch comparisons.[15]

MFCCs is a normalized energy parameter of audio.A popular formula to convert f hertz into m Mel is:[10]

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

MFCCs are commonly derived as follows:[11][12]

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

Figure 3 Calculation Of MFCC value

V.IMPLEMENTATION

Implementation of Track extraction module is carried out with help of Java API- jAudio that is useful mainly to extract out feature of MFCC. Dataset of 200 Hindi songs that includes almost every age of filmy songs and Hindi Bhajans too. Some kind of constraints that are used for implementation are taken as following values, DCT=13,window sample size=16bit(44.1 kHz,CD quality of song[14]),Song Format= .WAV.

Calculation :

- "Jumme Ki Raat" – Kick 2014 – 3:51 min = 231 seconds
- DCT=13
- Thus **231/13=17.76 ≈18** sec (13 Block Frame Of Size 18 sec each)

- On this 18 sec Apply Steps to calculate MFCC

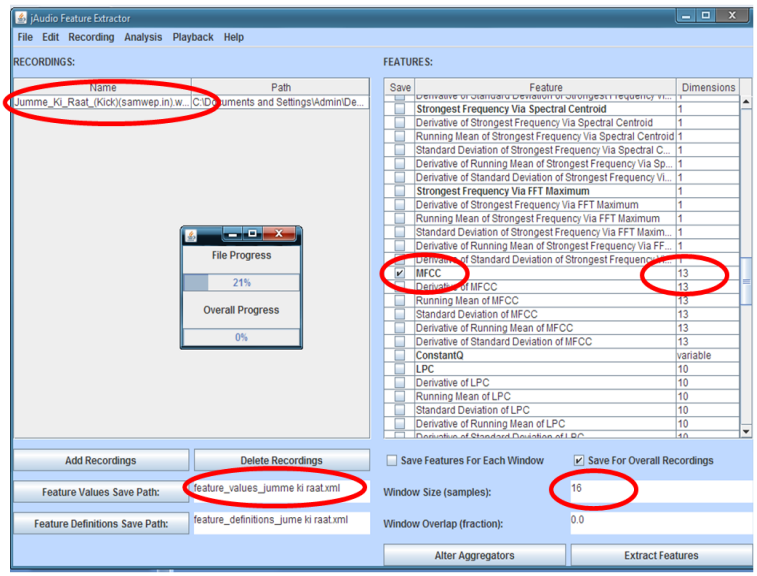

Figure 4 GUI of jAudio

song_title	movie_name	mfcc	genre_type	year
Tum Kya Jano	Hum Kisi Se Kam Nahi	45.7	Pop	1977
Aisi Deewanagi	Deewana	47.6	Pop	1992
Maria Maria	Partner	48.3	Pop	2007
Dil Ka Bhavar	Tere Ghar Ke Samane	48.6	Filmi Sugam	1963
Rim Zim Gire Sawan	Manzil	62.4	Tarana	1979
Meri Kahani	Meri Kahani	62.5	Classical	2008
Tum ho	Rockstar	62.8	Tarana	2011
Subhanallah	Yeh Jawaani Hai Deewani	63	Tarana	2013
Jumme Ki Rat Hai	Kick	64.7	Club	2014
Tu Meri	Bang Bang	74.3	Club	2014
Tamnchey Pe Disco	Bullet Raja	75.1	Disco	2014
Golmaal	Golmaal	87.4	Jazz	2006
Sooraj dooba hain	ROY	87.7	Jazz	2015
Tuje Dekh Dekh	Kaliyug	93	Sufi	2005

Table 4. DataSet Of Hindi Song & Its Relevant Genre

MFCC of all songs are carried out & based on value group of cluster is made by using classifier that classified 10 genres that are taken in this thesis. Results of particular song and its genre kind is mentioned in table.

V.MODULE-LYRICS EXTRACTION

V.I. LYRICS

Lyrics are a set of words that make up a song, usually consisting of verses and choruses. The writer of lyrics is a lyricist. The words to an extended musical composition such as an opera. LRC is a computer file format that synchronizes song lyrics with an audio file, such as MP3, Vorbis or MIDI[2,7].

Implementation part of lyrics is being worked with part of Metadata retrieval of song like name of singer,album,title,and name of song.Based on Metadata of Hindi mp3 song the retrieval procedure of Lyrics will be carried out by Metadata Extraction Of Hindi Song.

V.II.LYRICS FETCHER

Various Plug-in module are available that works Online with providing library update information, but problem exist that with that is if it's library containing that song Lyrics then and then it will work and it is fully based on Title of Song. Thus problem of lyrics retrieval by searching of "*Antra* " word still exist. I am working out on procedure of retrieval of song by crawling full Lyrics information that will avoid the problem of normal online player lyrics plug in module contain.

This novel approach is based on Metadata extraction procedure and LRC file information, part of Hindi Corpus and search query word 's problem is eliminated by using Hindi Dictionary.

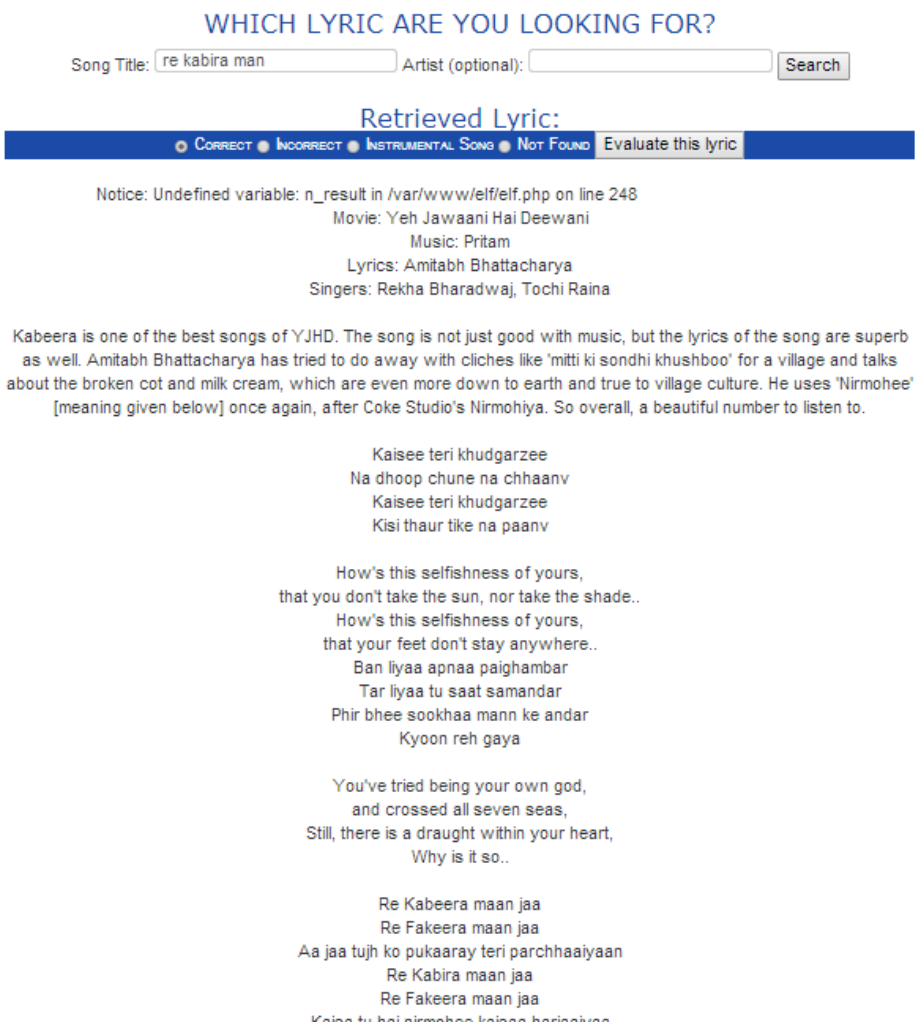

Figure 6 Lyrics Retrieval Of Re Kabira- "Yeh Jawani Hai Deewani-2013"

VI.CONCLUSIONS

1. The proposed system drawn for this research is useful to fulfill its objective to make searching of song by its lyrics efficiently and also to classify track of song. Overall efficiency leads of track extraction in right direction that also will be beneficial for any future work where more number of genres are targeted.
2. Innovative approach for Lyrics Extraction by Metadata extraction and crawling is useful to built any music information retrieval related application.
3. This music crawler will surely be unique one in the field of search engine for music information retrieval and also beneficiary as point of commercial application.
4. Last but not the least, dissertation target to normal human interface who are fond to be known as music lover, MSqueezer is surely for them

ACKNOWLEDGEMENTS

I wish to express my gratitude to my guide Prof. Girish H. Mulchandani, Computer Engineering Department, for introducing me to the problem and providing valuable advice. He not only provided me help whenever needed, but also gives motivation for the same.

REFERENCES

1)Author A. Frans Wiering,Dept. IT,Utrecht University, Netherlands."Can Human Benefit From Music Information Retrieval"

2) Bigand, E., Poulin-Charronnat, B.: Are We "Experienced Listeners"? A Review of the Musical Capacities That Do Not Depend on Formal Musical Training. Cognition 100

3) "LEVERAGING REPETITION FOR IMPROVED AUTOMATIC LYRIC TRANSCRIPTION IN POPULAR MUSIC" Matt McVicar†Daniel PW Ellis Masataka Goto†- IEEE Signal Processing 2014

4)"Multiobjective Time Series Matching for Audio Classification and Retrieval" Philippe Esling and Carlos AgonIEEE Language Processing Oct,2013

5) Matthias Mauch, Hiromasa Fujihara, and Masataka Goto,IEEE Trans. On Audio and Speech, Language Processing, 2012 "Chord Information IntoHMM-Based Lyrics-to-Audio Alignment"

6) Cory McKay, John Ashley Burgoyne, Jason Hockman, Jordan B. L. Smith, Gabriel Vigliensoni and Ichiro FujinagaISMIR 2010 "Evaluating Genre Classification Performance Lyrical Features Relative to Audio,Symbolic and Cultutal Features

7)"Simple LRC Format", http://en.wikipedia.org/wiki/LRC_(file_format)

8) "Genre classification and the invariance of MFCC features to Key and Tempo" Tom LH. Li and Antoni B. Chan.International Conference on MultiMedia Modeling, Taipei, 2011.

9) Mining Hindi-English Transliteration Pairs from Online Hindi LKanika Gupta,Monojit Choudhury, Kalika BaliNI Systems (India) Pvt. Ltd. Bangalore, Microsoft Research Labs India

10) Douglas O'Shaughnessy (1987). Speech communication: human and machine. Addison-Wesley. p. 150. ISBN 978-0-201-16520-3.

11)Min Xu et al. (2004). "HMM-based audio keyword generation". In Kiyoharu Aizawa, Yuichi Nakamura, Shin'ichi Satoh. Advances in Multimedia Information Processing – PCM 2004: 5th Pacific Rim Conference on Multimedia. Springer. ISBN 3-540-23985-5.\

12) Sahidullah, Md.; Saha, Goutam (May 2012). "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition". Speech Communication 54 (4): 543–565. doi:10.1016/j.specom.2011.11.004

13) Tom LH. Li and Antoni B. Chan" Genre classification and the invariance of MFCC features to Key and Tempo" ,Apears InInternational Conference on MultiMedia Modeling, Taipei, 2011, Page 9

14) « WAV Calculation« http://www.audiomountain.com/tech/audio-file-size.html

15) "MFCC"http://en.wikipedia.org/wiki/Mel-frequency_cepstrum

