Content Based Audio Retrieval using Chord

[1]Priyanka Gaonkar, [2]Dr. Satishkumar Varma, [3]Rupali Nikhare
[1]M.E. Student, [2]Professor, [3]Assist. Professor
[1]Department of Computer Engineering,
[1]Pillai College of Engineering, New Panvel, Navi Mumbai, India

***Abstract*—Music information retrieval (MIR) is a science of retrieving information from music signal. Extracting high-level features of music such as melody, harmony or rhythm from the raw music signal is a critical and challenging process in Music Information Retrieval (MIR) systems. Using one of such feature will be helpful in searching and retrieving relevant musical audio track effectively and efficiently from large collection of musical audio tracks. Content-based concept is based on the content of a given audio document and extracting the necessary information from it. The proposed content based audio retrieval system retrieves relevant songs from a large dataset of music audio tracks based on melody similarity. Relevant song retrieval is done by recognizing and extracting chord progressions (CPs). Chord progressions means transitions between adjacent chords which contains music information related to tonality and harmony which is helpful in effectively distinguishing whether music audio tracks are similar to each other or not. Various combination of features (4, 6 and 9) are used which consist of features like audio spectral centroid, spectral projection, audio spectrum flatness, audio spectrum spread, spectral crest factor, spectral decrease, spectral flux, spectral kurtosis, spectral Mfcc. Then supervised statistical machine learning model such as Support Vector Machine (SVM) and Hidden Markov Model (HMM) is used for recognizing and extracting CPs from music signals. Input to the system is music audio file and output is relevant list of ordered similar audio file from the database along with their emotions like joyful, angry, depression, content or normal. Music file here belongs to any one of 5 different emotion categories. This database consists of stored predefined audio file emotions and their features vector values. There are total 300 audio files in the database out of which 70% of audio files are used for pre-storage and 30% are used for testing. Finally, based on similarity between CPs of input audio track and CPs of all songs in the database, relevant songs are returned to the user in a ranked list as the retrieval results.**

***IndexTerms*—Audio retrieval, Support Vector Machine (SVM), Hidden Markov Model (HMM), Audio Features**

I. INTRODUCTION

Content Based Audio Retrieval (CBAR) is one of the most popular research areas on social web sites but it is critical research topic. On internet, there are many music soundtracks, with the same or similar melody but sung and recorded by different people. It is very complex to find out the particular song from the multiple copies of music files having same name. For that we are developing mechanism to quickly and reliably retrieve relevant songs from a large dataset of music audio tracks based on melody similarity. A melody is a linear succession of music tones. CBAR, in terms of melody similarity, has several applications such as near duplicate audio detection, relevant song retrieval and recommendation, etc. In typical scenarios, a user can find audio tracks similar to his favorite melody using an audio example, or music companies can recommend to user's new music albums with similar melodies according to listening records. These applications need large-scale CBMIR techniques.

In existing work, low-level features such as pitch, short-time Fourier transform (STFT), Chroma and Discrete Fourier transform (DFT), are used to describe music signal. But use of these low-level features for music audio content analysis and summarizations, are inefficient and inflexible for music information retrieval (MIR) task. In comparison, mid-level features (chord, rhythm, instrumentation) represented as musical attributes are able to better extract music structures from complex audio signals and retain semantic similarity. Chord is one of the important mid-level features of music. Chord is group of three or more musical notes. A chord progression is a sequence of musical chords. A chord sequence contains rich music information related to tonality and harmony, which is helpful for effectively distinguishing whether music audio tracks are similar to each other or not. Therefore, they are able to effectively and efficiently search and retrieve relevant musical audio track through a large collection of musical audio tracks. An audio may have emotions in the audio track. Song can sound like emotional song. Emotions in an audio depend on the chord, chord sequence, and chord progression. There are multiple emotions felt by human beings which are expressed in audio form through vocal sound when spoken from mouth of the person. An audio may symbolize that the audio is a happy song, normal song, angry song, depressed song, sad song.

II. Theoretical Background

A. Audio Features

Different Audio features are described in this section

1) Audio Spectral Centroid (ASC), which is the midpoint of the spectral energy distribution of that sound. It acts as the "balance point" of the spectrum.
2) Spectral projection is obtained by multiplying the NASE matrix with a set of extracted basis functions. This spectrum projection is used to represent features of a spectrum after projection onto a reduced rank basis.
3) Audio Spectrum flatness (ASF) is a measure, which is used in digital signal processing, to characterize an audio spectrum. This is measure in decibel.It is measure of noisiness of the spectrum; hence high flatness values indicate noisiness. Low flatness values generally indicate the presence of harmonic components.
4) Audio Spectrum spread (ASS) is describes concentration of the spectrum around the centroid Lower spread values would mean that the spectrum is highly concentrated near the centroid and higher values mean that it is distributed across a wider range at both sides of the centroid.
5) Spectral Crest factor is used to measure of a waveform such as sound. It is also called as the ratio of peak values to the effective value.
6) Spectral Decrease represents the amount of decreasing of the spectral amplitude.
7) Spectral Flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions that correspond to successive signal frames.
8) Spectral Kurtosis is pointedness of a spectrum can be used to indicate "pitchiness".
9) Mel-frequency cepstrum centroid (MFCC) 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.Mfccs are used as an excellent feature vector for representing musical signal.

B. Hidden Markov Model and Support Vector Machine

In this paper, Hidden Markov Model(HMM) and Support Vector Machines (SVM) method are proposed for content-based audio classification and retrieval. HMM has shown to be one of the powerful classifier in Audio retrieval and classification processing. The features extracted from the test's audio are considered to be a sequence of events and then used as the input for the HMM. It represents a set of states and the probabilities of making a transition from one state to another state.The typical usage in audio classification is to train one HMM for each class.SVM models the boundary between the different classes instead of modeling the probability density of each class. SVM algorithm is a classification algorithm that provides state-of-the-art performance in a wide variety of application domains.

III. Literature Survey

In this section, we refer to the relevant past literature that use the various chord recognition methods which are used for Music information retrieval (MIR) task.

Heng-Tze Cheng, Yi-Hsuan Yang, Yu-Ching Lin, I-Bin Liao and Homer H. Chen [1] proposed in Automatic Chord Recognition for Music classification and retrieval propose a new method to construct chord features for music emotion classification and evaluate its performance on commercial song recordings. They also investigated mid-level music feature construction and its applications to music classification and retrieval. With recognition accuracy as competitive as existing systems and simplicity and time-efficiency advantages, the proposed N-gram-based chord recognition system is particularly attractive for practical applications. The two proposed new chord features, longest common chord subsequence and chord histogram, are useful for music analysis, management, and retrieval. With these two mid-level music features, this system able to achieve good improvement over existing approaches that use only low level features for emotional valence prediction.

Jingzhou Yang, Jia Liu, Wei-Qiang Zhang [2], this paper present Query by humming (QBH) is an efficient way to search the song from a large database. We proposed a note-based system, which consists of noted-based linear scaling (NLS) and noted-based recursive align (NRA), and makes use of both note information and the difference among the distributions of humming. Comparison experiments against several other widely-used algorithms in QBH reveal that our proposed system can get a good balance between computation time and recognition rate.

Laurent Oudre, CédricFévotte [3], this paper describes a probabilistic approach to template-based chord recognition in music signals. The algorithm only takes Chroma gram data and a user-defined dictionary of chord templates as input data. No training or musical information such as key, rhythm, or chord transition models is required. The chord occurrences are treated as probabilistic events, whose probabilities are learned from the song using an expectation–maximization (EM) algorithm. The systems are tested on two evaluation corpuses; the first one is composed of the Beatles catalog (180 pop-rock songs) and the other one is constituted of 20 songs from various artists and music genres. The chord transcription output by our automatic chord transcriber is a sequence of chord labels with their respective start and end times. This output can be used for song playback—which constitutes the main aim of our system—but also in other applications such as song identification, query by similarity or structure analysis.

J. Osmalskyj, J.J. Embrechts, S. Piérard, M. Van Droogenbroeck[4], This Paper presents a feed-forward neural network for chord recognition. The method uses the known feature vector for automatic chord recognition called the Pitch Class Profile (PCP). Although the PCP vector only provides attributes corresponding to 12 semi-tone values, we show that it is adequate for chord recognition. The method uses the known feature vector for automatic chord recognition called the Pitch Class Profile (PCP). Although the PCP vector only provides attributes corresponding to 12 semi-tone values, we show that it is adequate for chord recognition. The use of a simple 12-bin PCP vector based on the Discrete Fourier Transform, we show promising results and fast processing, which would have probably not been achieved with more complex pre-processing steps.

R.Thiruvengatanadhan,P.Dhanalakshmi [5], this paper present music classification using SVM and GMM. Accuracy of system is based on features and classification model. In this paper both Time domain and Frequency domain features are extracted from audio signal.Time domain features are Zero Crossing Rate (ZCR) and Short Time Energy (STE). Frequency domain features are spectral centroid, spectral entropy and spectral roll-off. After extracting feature, classification is carried out, using Support Vector Machine (SVM) and Gaussian Mixture Model (GMM). SVM is trained and tested for different kernel function and performance. Both SVM and GMM are comparing based on performance. But GMM are providing better performance compared to SVM.

S.Suguna, J.BeckyElfreda [6], this paper present cepstral Features for audio retrieval. Audio information retrieval has been performed on GTZAN datasets using weighted Mel-Frequency Cepstral Coefficients (WMFCC) feature which is a kind of cepstral feature. The results obtained for the various stages of feature extraction WMFCC and retrieval performance plot has been presented. The use of Distance-from-Boundary (DFB) and Support vector machine (SVM), audio retrieval and classification task which use Mel cepstral feature had been performed on a database which consists of 409 sounds of 16 classes. The results for the various stages of WMFCC feature extractions have been obtained. With the help of WMFCC feature, audio retrieval task has been performed on the GTZAN database.

ShrutiVaidya,Kamal Shah [7], proposed application for vector quantization for audio retrieval. In this paper, study of audio content analysis for classification is presented, in which an audio signal is classified according to audio type. Different features like, zero-crossing rate (ZCR), low short-time energy ratio (LSTER), spectrum flux (SF), spectrum roll off (SR) are used with K-nearest-neighbor (KNN) and Fast Fourier Transform (FFT). Reducing number of audio features reduces the accuracy levels.

IV. CONTENT BASES AUDIO RETRIEVAL SYSTEM (CBARS)

A. System Architecture

The basic operation performed by this system is as follows. First, the feature vectors are estimated for each and every signal from the database. Then, feature vectors are estimated for each input signal from the user. The feature vectors of input signal from the user are compared to each feature vectors of audio signal form the database. This database consists of predefined audio file emotions and their features vectors are calculated and store in matrix. If majority of the features of input test audio file matches with the stored database then that audio file is retrieved. Here we are extracting top 10 similar audio files which are available in the database. When we search something from any search engine then we get top 10 results on first search. These top 10 retrieved webpages are ranked according to various factors like number of users visited or like number of time spent on particular website or particular page of that website. This system is designed so that we obtain top 10 similar audio files based on the audio features and emotion of the audio. Then supervised learning model is used for recognizing CPs from each Feature vector from input audio track and all audio tracks in the database. System uses Support Vector Machine (SVM) and Hidden Markov Model (HMM) supervised machine learning algorithm. Finally, based on similarity between CPs of input audio track and CPs of all songs in the database, relevant songs are returned to the user in a ranked list as the retrieval results. The system architecture of this system is shown in figure 1, which represents how the Content based audio retrieval system proceeds.

The system consists of following phases:

- Feature Extraction
- Classification
- Histogram computation

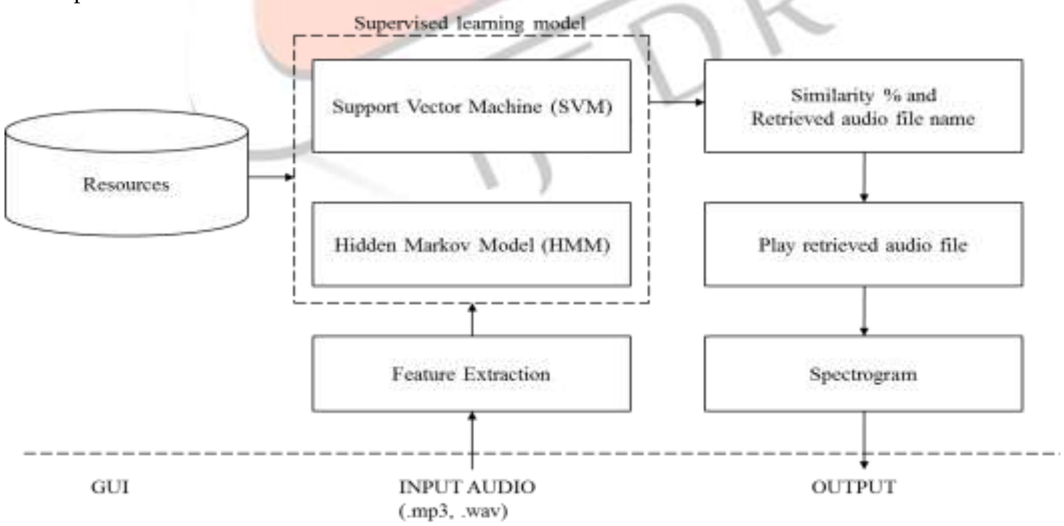


Figure1 System Architecture for CBARS

B. Working of CBAR System

Following figure shows, working of the CBAR System

User Interface

The user interface is shown in figure 2 where user can select an audio file by clicking on the buttons displayed on the user interface.

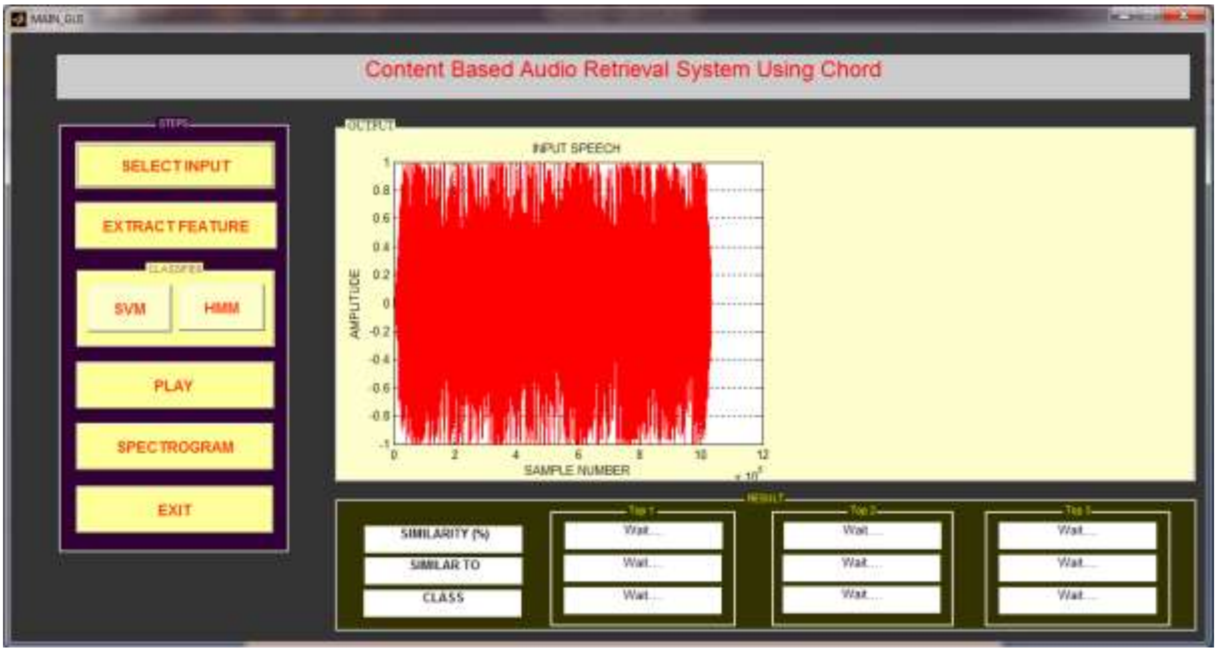


Figure 2 Audio input displayed in graph

Audio Feature Extraction

After selecting audio file, then next step isfeature extraction is shown in figure 3 where all features for selected audio is calculated and displayed to the user. There are 9 features calculated for 1 audio input.

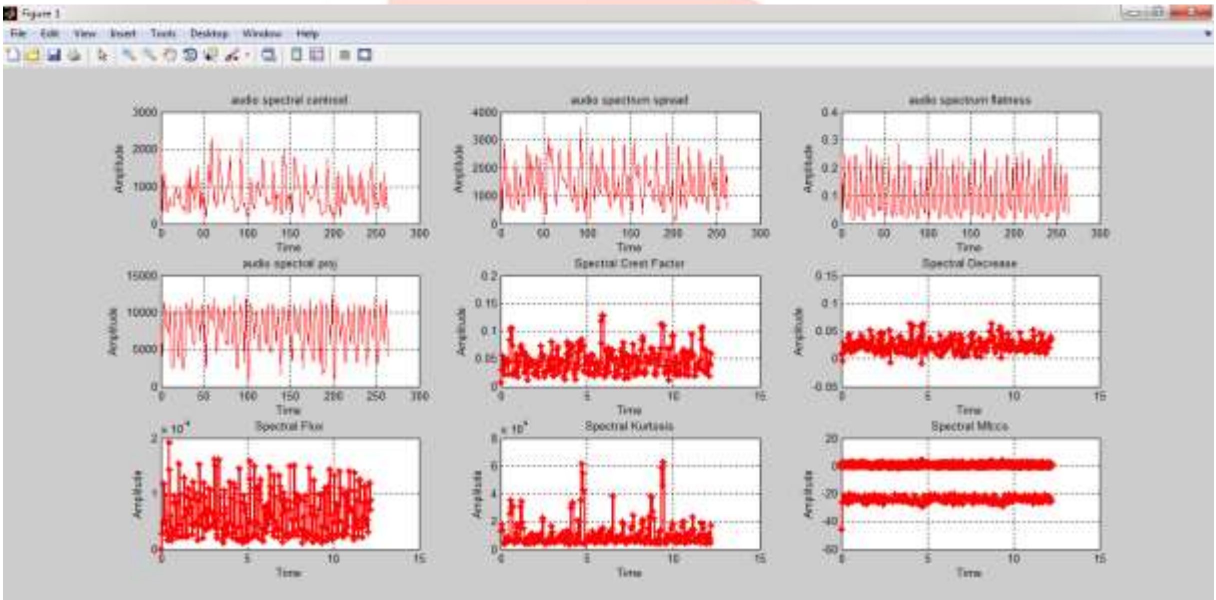


Figure 3 Audio Feature Extraction

Calculating similarity % using SVM Classifier

Similarity % for input audio is calculated using SVM classifier is shown in figure 4. Features for input audio is displayed using red bar graph on the left and features for retrieved audio is shown in green bar graph on the right. For a selected audio input similarity % is calculated for top1 audio file as 85.4168%. It means that the audio input is 85.4168 % similar to the stored audio file named "77.mp3". Similarly, for top2 and top3 retrieved audio file similarity % is calculated as 82.6573% and 78.5647% respectively. If the similarity % is equal to 100% then the audio input matches with the stored audio. This percentage is obtained by using SVM classifier.The user can play the audio extracted from the database using SVM classifier. When user clicks on the SVM button and then the play button, then audio retrieved by SVM classifier is played.

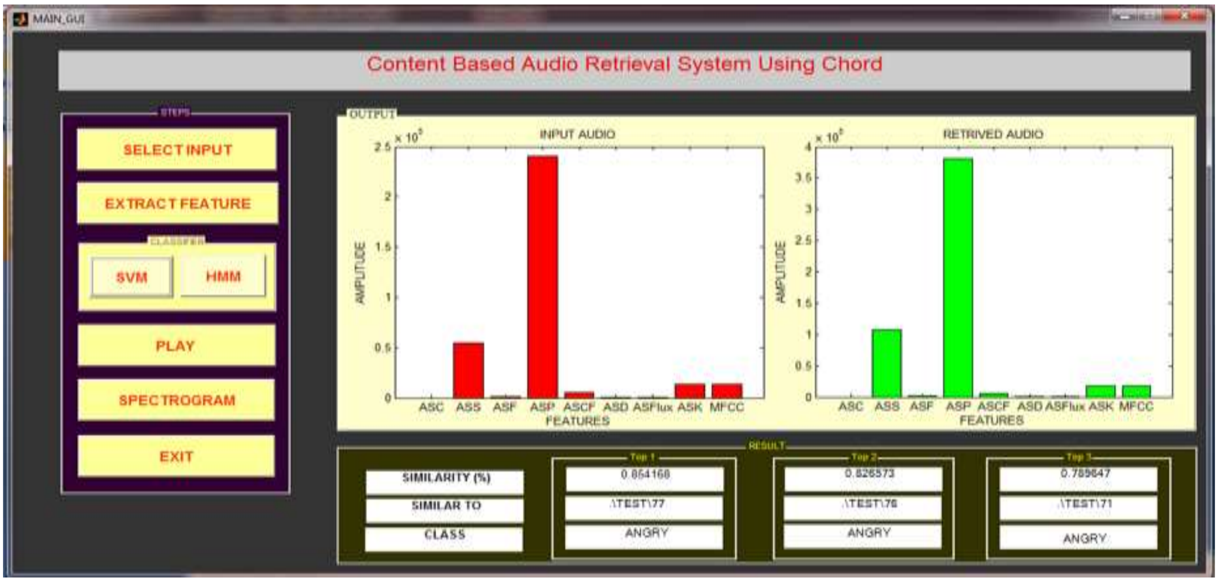
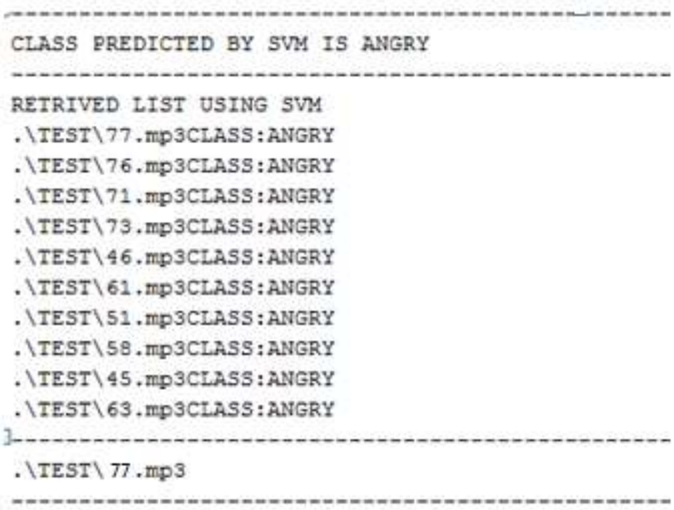


Figure 4 Calculation of similarity % using SVM Classifier

List of top 10 audio files names retrieved are displayed to the user in the background which is shown in figure 5.



```
---------------------------------------------------
CLASS PREDICTED BY SVM IS ANGRY
---------------------------------------------------
RETRIVED LIST USING SVM
.\TEST\77.mp3CLASS:ANGRY
.\TEST\76.mp3CLASS:ANGRY
.\TEST\71.mp3CLASS:ANGRY
.\TEST\73.mp3CLASS:ANGRY
.\TEST\46.mp3CLASS:ANGRY
.\TEST\61.mp3CLASS:ANGRY
.\TEST\51.mp3CLASS:ANGRY
.\TEST\58.mp3CLASS:ANGRY
.\TEST\45.mp3CLASS:ANGRY
.\TEST\63.mp3CLASS:ANGRY
---------------------------------------------------
.\TEST\ 77.mp3
---------------------------------------------------
```

Figure 5 Top 10 audio files retrieved using SVM

Calculating similarity % using HMM Classifier

Similarity % for input audio is calculated using HMM classifier is shown in fig 4.6. Features for input audio is displayed using red bar graph on the left and features for retrieved audio is shown in green bar graph on the right. For a selected audio input similarity % is calculated as 69.8858 %. It means that the audio input is 69.8858 % similar to the stored audio file named "113.mp3". This value is less than the value obtained by SVM classifier. The user can play the audio extracted from the database using HMM classifier. When user clicks on the HMM button and then the play button, then audio retrieved by HMM classifier is played.

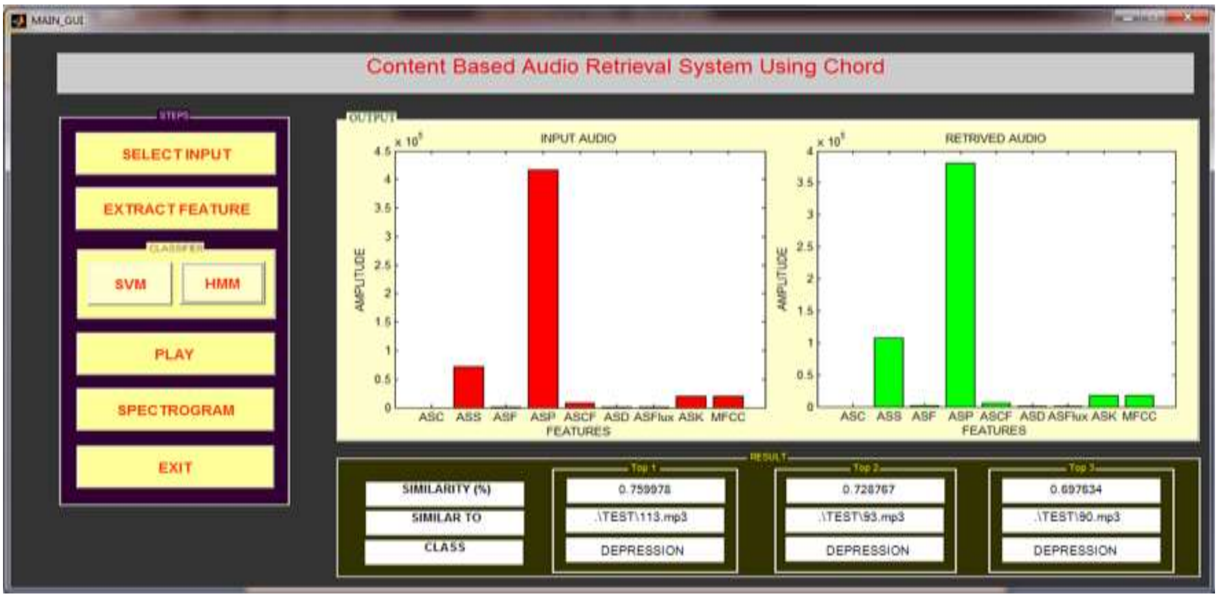


Figure 6 Calculation of similarity % using SVM Classifier

List of top 10 audio files names retrieved are displayed to the user in the background which is shown in figure 7.

```
------------------------------------------------
CLASS PREDICTED BY HMM IS DEPRESSION
------------------------------------------------
RETRIVED LIST USING HMM
.\TEST\113.mp3CLASS:DEPRESSION
.\TEST\93.mp3CLASS:DEPRESSION
.\TEST\90.mp3CLASS:DEPRESSION
.\TEST\86.mp3CLASS:DEPRESSION
.\TEST\105.mp3CLASS:DEPRESSION
.\TEST\83.mp3CLASS:DEPRESSION
.\TEST\96.mp3CLASS:DEPRESSION
.\TEST\98.mp3CLASS:DEPRESSION
.\TEST\82.mp3CLASS:DEPRESSION
.\TEST\92.mp3CLASS:DEPRESSION
------------------------------------------------
------------------------------------------------
.\TEST\113.mp3
------------------------------------------------
```

Figure 7 Top 10 audio files retrieved using HMM

Audio Spectrogram

Input audio spectrogram is shown in figure 8 where x-axis is having normalized frequency and y-axis is having time factor. Exit button from the GUI is shown is figure 8 where user can come out of the GUI by terminating the GUI process.

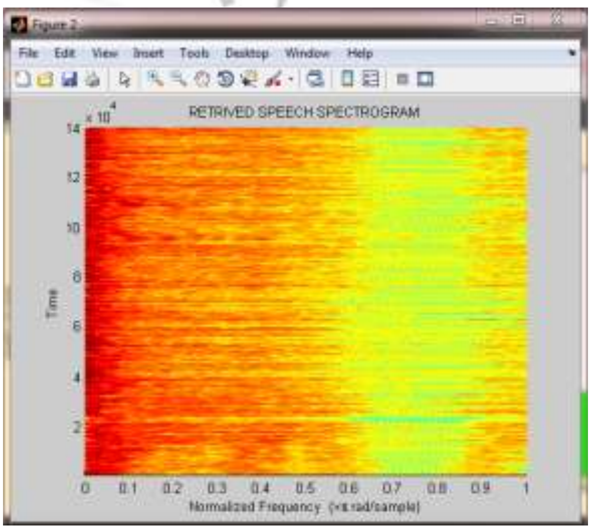


Figure 8 Input audio spectrogram

Feature Value

Feature value matrix is shown in Figure 9 for the trained files. Here rows specify feature values and column specify features for particular trained file.

TRN

205x70 double

	1	2	3	4	5	6	7	8	9	10
1	339.0199	3.4938e+03	425.2117	579.2155	375.1833	252.1922	233.0160	598.4950	109.7961	142.6
2	448.5521	3.5243e+03	904.6684	570.9135	616.7513	136.5466	186.6181	524.5080	213.4970	135.6
3	457.2739	3.5040e+03	750.2884	294.1766	524.4355	206.2751	251.1191	795.0941	181.4830	213.5
4	543.6657	3.5291e+03	1.6037e+03	229.2222	561.6797	324.2616	316.8470	398.9753	202.8896	198.6
5	313.7451	3.5344e+03	664.4620	242.6962	1.6699e+03	1.4360e+03	211.5329	191.5566	178.7306	212.8
6	353.4344	3.5171e+03	707.2774	396.4226	682.4818	506.8659	155.4518	212.1398	394.2472	316.2
7	326.4610	3.4959e+03	583.4694	841.0646	885.3674	464.0837	128.9869	606.6532	280.4483	234.6
8	282.7651	3.5212e+03	790.0909	608.6407	1.0171e+03	460.8927	153.5394	580.4343	218.6159	403.5
9	387.2431	3.5438e+03	323.4515	346.9631	646.7299	368.1157	274.7119	676.3087	268.7328	676.6
10	274.9605	3.5056e+03	238.7485	533.7945	334.9900	592.7455	293.2028	881.6204	222.9235	705.0
11	230.1278	3.5183e+03	418.5252	289.5991	629.2170	1.0279e+03	342.0270	675.0336	309.5054	406.9
12	335.1221	3.5069e+03	591.7994	719.8007	2.3580e+03	313.0218	1.1503e+03	376.4686	188.3992	195.7
13	507.7975	3.4960e+03	0	327.8688	499.7051	311.5800	595.2003	1.2393e+03	280.5807	579.3
14	700.5028	3.4814e+03	0	256.9877	547.3432	738.9450	543.2270	500.9237	300.9326	389.7
15	553.3733	3.5456e+03	0	435.2540	611.6850	1.0688e+03	556.5498	691.3668	315.3661	257.8
16	604.5703	3.5140e+03	0	553.3028	480.6272	897.2233	401.2135	536.8665	370.4498	331.7
17	284.6769	3.5342e+03	4.6918e+03	298.7426	457.6448	538.4936	462.5010	403.3585	469.0585	430.0

Figure 9 Feature Value

V. RESULT ANALYSIS

A result obtained from the various runs of the experiments conducted is shown in table 1. There are 4 sets of data files taken. First set is 75 audio files, second is 150 audio files, and third is 225 audio files and finally 300 audio files. All these sets are manually divided for training and testing process into 70% by 30 % ratio. Correct file classification means total number of files correctly classified as similar by the system by using different features. Incorrect file classification means total number of files incorrectly classified as non-similar by the system. Numbers of features taken for experimentations are 4, 6, and 9 respectively for each data set. Features for training audio files are calculated and stored for testing later on. Set of 4 features considered are: Audio Spectral centroid (ASC), Spectral projection, Audio Spectrum Flatness (ASF) and Audio Spectrum Spread (ASS). Set of 6 features considered are: Audio Spectral centroid (ASC), Spectral projection, Audio Spectrum Flatness (ASF), Audio Spectrum Spread (ASS), Spectrum Crest Factor and Spectral Decrease. Set of 9 features considered are: Audio Spectral centroid (ASC), Spectral projection, Audio Spectrum Flatness (ASF), Audio Spectrum Spread (ASS), Spectrum Crest Factor, Spectral Decrease, Spectral Flux, Spectral Kurtosis and Spectral Mfcc.

In table 1, similarity % using SVM and HMM classifier with 9, 6 and 4 features is shown. Total numbers of audio files used are 300 for experimentation. Set of 75 and total 4 sets of data is used. Out of which 70% is used for predefining the data, also called as training and 30 % is used for testing. Here we will use the word defined or train simultaneously. In serial number 1, total numbers of audio files selected are 75. Out of which manually selected 56 audio files are used for predefining or labeling, also called as training and 19 audio files are used for testing. 19 files are tested under combination of features (9, 6 and 4). Similarly, the data in training and testing set are incremented.

Table 1Results obtained from the various runs of the experiments

SN	Total No. of audio files	Similarity % for SVM and HMM					
		9 features		6 features		4 features	
		SVM	HMM	SVM	HMM	SVM	HMM
1	75, Defined: 56, Test: 19	53.26	27.53	45.70	25.23	37.39	20.20
2	150, Defined: 112,Test: 38	59.66	38.73	51.88	33.15	42.64	26.72
3	225, Defined: 168, Test: 57	65.44	46.29	56.84	39.07	47.44	31.56
4	300, Defined: 210, Test: 90	68.94	50.86	59.96	42.94	50.59	35.08

Average similarity % for SVM and HMM classifier with 9, 6 and 4 features for 210 trained and 90 tested files is shown in figure 4.10. Where x-axis specifies number of trained files and y-axis specifies Average Similarity in %.

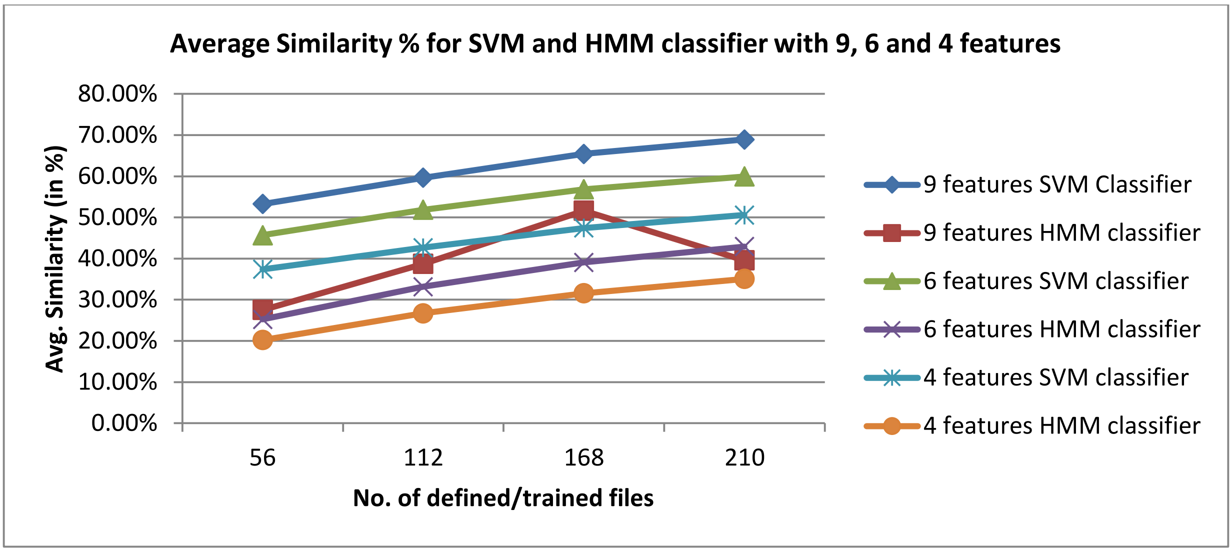


Figure 9 Average Similarity % for SVMand HMM classifierwith 4, 6 and 9features

VI. Conclusion

The use of content-based approaches has always been a common strategy in the music information retrieval area. Content Based Audio Retrieval (CBAR) is one of the most popular research areas on social web sites. The algorithm consists of two key points: recognizing chord progressions from a music audio track based on a supervised learning model related to musical knowledge and computing a summary of the audio track from the recognized chord progressions by locality sensitive hashing (LSH). Statistical approaches for extraction of chord progressions using SVM and HMM based framework is proposed. This system takes predefined/pre-stored, also called labeled data or resources. System predicts and displays the emotion of the input test audio file, calculates and displays Similarity% and name of retrieved audio file (which is most similar) is displayed on the screen. A list of 10 audio files is displayed to the user in the background of this system. As user is not interested in only one audio file which is most similar to the input test audio file but instead of list of top 10 or more audio file names similar to any search engine output results. Then users also have option to hear the retrieved output file by clicking on play button on the user interface. The user can see the spectrogram and finally can close the interface by clicking on exit button. This terminates the working of the system. Else user can continue using the system and start the process again by selecting an audio input test file. And the process repeats. Maximum average similarity % is obtained by 9 features using SVM classifier as compared to HMM classifier model.

VII. Acknowledgment

I am using this opportunity to express my gratitude to thank all the people who contributed in some way to the work described in this paper. My deepest thanks to my project guide for giving timely inputs and giving me intellectual freedom of work. I express my thanks to the head of computer and IT department, M.E. coordinator and to the principal of Pillai College of Engineering, New Panvel for extending their support.

References

[1] Heng-Tze Cheng, I-Bin Liao and Homer H. Chen, "Automatic Chord recognition for music classification and retrieval", ICME 2008, pp. 1505-1508.

[2] Jingzhou Yang, Jia Liu, Wei-Qiang Zhang "A Fast Query by Humming System Based on Notes", In Proc. ISCA, September 2010, pp. 2898-2901.

[3] Laurent Oudre,CedricFevotte, "Probabilistic Template-Based Chord Recognition", IEEENovember 2011, pp.2249 – 2259.

[4] J. Osmalskyj, J.J. Embrechts, S. Pierard, M. Van Droogenbroeck, "NEURAL NETWORKS FOR MUSICAL CHORDS RECOGNITION", Actes des Journées d'Informatique Musicale (JIM 2012), Mons, Belgique, 9-11, May 2012.

[5] R.Thiruvengatanadhan,P.Dhanalakshmi[7],"Speech/Music Classification using SVM and GMM", IJARCSSE, September 2013, pp. 793-801.

[6] S.Suguna,J.BeckyElfreda, "Audio Retrieval based on Cepstral Feature",In Proceeding of International Journal of Computer Applications (IJCA), December 2014, pp. 28-33.

[7] ShrutiVaidya, KamalShah,"Application of Vector Quantization for Audio Retrieval",Volume 88, February 2014, pp. 0975 – 8887.

[8] J. T. Foote, "Content-Based Retrieval of Music and Audio." In Proc. SPIE, vol.3229, 1997, pp. 138-147.

[9] Wang. A "Automatic Identification of Musical Versions Using Harmonic Pitch Class Profiles", PhDthesis,September 2011.

[10] MaksimKhadkevich, "Music signal processing for automatic extraction of harmonic and rhythmic information", PhD thesis, December 2011.

[11] T. E. Ahonen, "Combing chroma features for cover version identification," In Proc. ISMIR, 2010, pp. 165–170.

[12] Y. Yu,M. Crucianu, V. Oria, and E. Damiani, "Combing multi-probing histogram and order-statistics based LSH for scalable audio content retrieval," In Proc. ACM MM, 2010, pp. 381–390.

[13] Support Vector Machine Topic, https://en.wikipedia.org/wiki/Support_vector_machine

[14] Hidden Markov Model Topic, http://www.shokhirev.com/nikolai/abc/alg/hmm/hmm.html