A Novel Approach to Identify a Singer in a Video Song using Spectral and Cepstral Features

S. Metilda Florence*[1], S. Mohan[2]
[1]SRM University, Chennai, India
[2]CCIS, Al Yamamah University, Kingdom of Saudi Arabia.
[*]Corresponding author: E-Mail: metilda_florence@yahoo.com

ABSTRACT

The Automatic Annotation of a Singer in a Video Song System enables the user to search for their favorite Singer's video song from video store. The proposed methodology performs the search in a video store by comparing the content of the video rather than the user's textual query and tags associated with the videos. We make an effort for identifying underlying Singer in the video songs by mining their audio features, frequencies and onset values. Various algorithms, filters and classifiers are used to implement this system, namely to mention a few Chebychev infinite-impulse response(IIR) filter , inverse comb filter, Naive Bayes, Sequential Minimal Optimization (SMO) Classification Algorithms etc. This paper focused on the extraction of information about the video contents automatically. The extracted information can serve as an initial step for various data access methods such as surfing, searching, comparison, and classification. It is worth mentioning that annotating music information in a video is an emerging task and was not much covered in past research papers. In the proposed System three Singers namely S.P. Balasubramaniam, Susheela and Swarnalatha are selected for analysis. For each Singer 100 video songs of length 10 seconds duration are taken. From these video songs the vocal track alone are extracted by using IIR digital filter and inverse comb filter. Mathematical functions are applied to calculate the Spectral and Cepstral features from the extracted signal. These features are applied to five classifiers for classification. This System gives a maximum of 95% accuracy in identifying a Singer in a video song using SMO classifier.

KEY WORDS: Content Based Search, Video annotation, Singer Identification, Spectral Features, Mel Frequency Cepstral Coefficients.

1. INTRODUCTION

A recent statistics of YouTube states that it has more than 1 billion users. Every day people watch hundreds of millions of hours of videos on You Tube. In near future watching online videos will increase in huge amount. To reuse the material available in a video store, there is a requirement to annotate the accessible material. In several video production companies, this task is still executed manually. The proposed technique will annotate the video automatically from the audio information. Normally in Video Annotation, the videos are annotated in following categories: Genre Classification (cartoon, commercial, sports, movie, news, music), Objects in the video (car, mountain, sky, road, man, animal, birds etc.) and Semantic Level (desert, indoor, outdoor, sea shore etc.). In these works, music in the video is not concentrated much. Our proposed work is fully concentrated on music and categorise the video based on music as follows: Singer Identification, Instrument Recognition (Piano, Violin, Guitar etc.), Mood Classification (Happy, Sad, Angry etc.) and Genre Classification (Rock, Pop, Classical etc.). Except Singer Identification module, the work on other modules is already published by the same author in Proceedings and Journal (Metilda, 2014; 2015).

In this paper, the implementation and results of Singer Identification module is discussed in detail.The main contribution of this paper is the use of music to annotate video, which is a much less explored problem. For Singer classification, three legends namely Balasubramaniam (SPB), Susheela and Swarnalatha are selected. More contributing video clips of 10 seconds duration are collected from Internet and Video CDs. For each singer 100 video songs are taken. From each video song the vocal track alone extracted. Mathematical functions are applied to calculate the MFCC and Spectral features. These features are directly applied to standard classifiers for classification. Since there is no single classification algorithm which is recognized to perform well for all applications, it is required to carry out a comparative study on the same set of signals to determine the best classifier. The classification accuracy is depends on both the classifier and strength of the features that are extracted.

2. PROPOSED SYSTEM

The proposed System is depicted in Figure.1. More contributing video clips from VCDs and the Internet are collected. Each collected video file is processed by transforming and trimming it to 10 seconds duration. The Audio tracks from video clips are extracted with a sampling rate of 44.1 kHz. From these extracted audio tracks the vocal is isolated and then processed by the feature extraction phase in order to extract the features. By using efficient classifiers the extracted feature set is classified based on Singers. Five different classifiers are used to train and test the dataset.

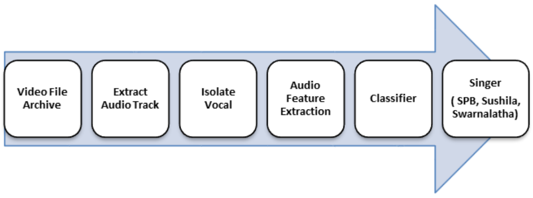

Figure.1. Block diagram of proposed System

Video File Archive and Audio Track Extraction: In the initial stage of this work, video files are collected from the internet and Video CDs. For each singer 100 songs are collected. Duration of all songs was equally set to 10 seconds. During this short period, we can cut only the singing part of the song by avoiding intro, chorus, outro etc. This will help us to extract only the singer voice with back ground music. Using Matlab code audio track is extracted without opening each video files. This audio track contains both vocal and instrumental music in it.

Vocal Isolation: In this phase, vocal is isolated from the extracted audio track. The most of the energy in the singing voice falls between 200Hz and 2000 Hz (Cook, 1990). Since frequency range of singing is concerned, a direct method would be used to detect energy within the frequencies bounded by the range of vocal energy. A simple method is to filter the audio signal with a band-pass filter which permits the vocal range to pass through while weakening other frequency areas. To achieve this Chebychev infinite-impulse response (IIR) digital filter of order 12 is used. This filter has the musical effect of suppress other instruments that fall outside of this frequency region. But even in popular music, the voice is not the only instrument creating energy in this region. Drum, for example, scatter energy over an extensive collection of frequencies, a significant amount of which falls in our range of interest. So another measure is needed to separate the voice from these other sources.

Singing voice is highly harmonic (Cook, 1990) and other high energy sounds in this region, particularly Drum are not as harmonic and distribute their energy more widely in frequency. To exploit this variation, an inverse comb filter bank is used to detect high amounts of harmonic energy. By passing the filtered signal (F) through a set of inverse comb filters with varying delays, we can find the fundamental frequency which the signal is most weakened. By taking the ratio of the total signal energy to the maximally harmonic attenuated signal, Harmonicity is measured.

$$Harmonicity = \frac{F_{original}}{MIN_i(F_{filtered,\, i})} \qquad (1)$$

By thresholding the Harmonicity against a fixed value, we have a detector for harmonic sounds. Most of these signals relate to region of singing.

Feature Extraction: Music signal is described using various numerical values extracted from the signal. These are called as features of the signal. A large amount of different feature sets, mainly originating from the area of speech recognition, have been proposed to characterize audio signals (Tzanetakis, 2002). The features used to signify timbre texture are based on typical features proposed for music-speech separation. From the available features the most relevant features are selected for this System. They are Spectral and Ceptral features. Spectral features used are Spectral Centroid, Spectral Rolloff and Spectral Flux. Cepstral feature used is MFCC.

The purpose of feature extraction is to preserve useful information, eliminate noise and other unwanted information. Aforesaid features are extracted by the steps given in the following sections. Finally all these features are combined together for creating the dataset for our System.

Spectral Centroid: The Spectral Centroid is a measure used in Digital Signal Processing to characterize a spectrum. It indicates where the "center of mass" of the spectrum is. Perceptually, it has a robust connection with the impression of "brightness" of a sound. It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fast Fourier Transform, with their magnitudes as the weights:

$$\mathrm{c} = \frac{\sum_{n=0}^{N-1} S(n)d(n)}{\sum_{n=0}^{N-1} d(n)} \qquad (2)$$

Where *d(n)* represents the weighted frequency value, or magnitude of bin number *n*, and *S(n)* represents the center frequency of that bin.

Spectral Rolloff: The Spectral Rolloff is the frequency R_t under which 95% of the power distribution is concentrated.

$$\sum_{n=1}^{Rt} S_t[n] = 0.85 * \sum_{n=1}^{N} S_t[n] \qquad (3)$$

Where *n* is time index ranging from *0 <= n <= N-1, N* is duration of file and *t* is current time frame.

Spectral Flux: The Spectral flux is defined as the squared difference between the normalized magnitudes of successive spectral distributions.

$$SF = \sum (F_t[n] - F_{t-1}[n])^2 \qquad (4)$$

Where $F_t[n]$ and $F_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current time frame *t*, and the previous time frame *t*-1, respectively.

Mel-Frequency Cepstral Coefficients (MFCC): The MFCC is a very common and efficient technique for signal processing. It describes the spectral shape of the signal. Its computation involves five main steps, including the conversion of signal frame into a Mel scale representation in order to emphasize the middle frequency bands. The MFCC transformation has been proved useful for computing music similarity.

To extract MFCC features, the audio signal is divided into a number of overlapped frames. To minimize the ringing effect (Chandwadkar, 2012), multiply each frame by a Hamming window hwd (h) is given in (5).

$$hwd(h) = 0.54 - 0.46\cos\left(\frac{2\Pi h}{N\text{-}1}\right), 0 \le h \le N - 1 \qquad (5)$$

Where N is the length of Hamming window. FFT is then applied on each pre-emphasized, Hamming windowed frame to obtain the corresponding spectrum. The audio samples are sampled at 44.1 KHz. To extract features, music samples are segmented into 23 milliseconds (ms) frames to get accurate FFT. When compared with other sample rates and segment size combinations, 44.1 KHz and 23 ms gives the best performance. For windowing Hamming window is used because combination of Mel frequency and Hamming window gives good results. For each window, thirteen MFCC coefficients are calculated. Feature vector of MFCC for each window of 23 ms is obtained. As the size of this feature vector is very large, instead of using these features directly for classification, their mean and standard deviations are obtained. Totally it produces (13 standard deviation values, 13 mean values) 26 features.

Classifiers: It is necessary to use more than one classifier to get the average accuracy. In the proposed System, five efficient classifiers are used to train and test the dataset. They are given as follows: Naive Bayes, Sequential Minimal Optimization (SMO), Multiclass Classifier, J48 and Random Tree.

3. EXPERIMENTAL RESULTS AND DISCUSSION

Dataset: From the internet and VCDs 300 video songs are collected. By using video cutter tool the video songs are trimmed to 10 seconds duration. In Matlab, all the video files are read one by one to extract the audio track and to isolate the vocal. From the isolated vocal track Spectral and Cepstral features are extracted.

Classification: For classification , we used WEKA tool (Anguita, 2012). In this tool , 10 cross validation technique is used for classification. Cross-validation is a model validation technique for assessing how the results of statistical analysis will generalize to an independent dataset. The confuson matrix for each classifiers are given in Table.1.

Table.1. Confusion matrices for different classifiers

Singers	SPB	Sushila	Swarna
SPB	95	1	4
Sushila	1	94	5
Swarna	8	5	87

Table 1(a). Naïve Bayes Classifier

Singers	SPB	Sushila	Swarna
SPB	97	0	3
Sushila	1	96	3
Swarna	2	5	93

Table 1(b). Sequential Minimal Optimization

Singers	SPB	Sushila	Swarna
SPB	96	0	4
Sushila	5	92	3
Swarna	4	7	89

Table 1(c). Multiclass classifier

Singers	SPB	Sushila	Swarna
SPB	96	1	3
Sushila	1	94	5
Swarna	4	8	88

Table 1(d). Random Forest classifier

Singers	SPB	Sushila	Swarna
SPB	81	2	17
Sushila	2	89	9
Swarna	8	9	83

Table 1(e). J48 classifier

Table.2. Analysis Report

Classifier	Accuracy %			Overall %
	SPB	Sushila	Swarnalatha	
Naïve Bayes	95	94	87	92
SMO	97	96	93	95
Multiclass	96	92	89	92
Random Forest	96	94	88	93
J48 classifier	81	89	83	84

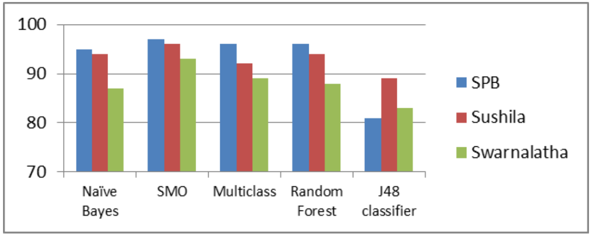

Figure.2. Comparison of different classifiers performance

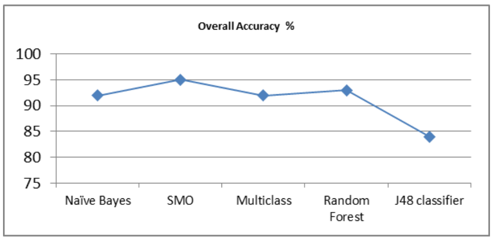

Figure.3. Overall Classification Accuracy for different classifiers

Table.1 depicts confusion matrices for all the five classifiers. In Table 2, the overall accuracy of all these classifiers is listed. These tables show that, SMO classifier has given the highest accuracy of 95 % and J48 classifier given the least accuracy of 84%. To conclude, Singer identification module gives a maximum of 95% accuracy in identifying a Singer in a video song using SMO classifier. FigureS.2 and 3 gives the pictorial representations of the accuracy percentages.

4. CONCLUSION

A novel and efficient approach for Identifying a Singer in Video Song is presented. This will enable the music lovers to locate their favorite Singer's Video Song. Current search engines will search the video by their tags not by content. Our proposed system will identify the song based on the Voice track in the video. We achieved maximum of 95 % accuracy. In the proposed System three Singers namely SPB, Susheela and Swarnalatha are selected for analysis. For each Singer 100 video songs of length 10 seconds duration are taken. From these video songs the vocal track alone are extracted by using IIR digital filter and inverse comb filter. Mathematical functions are applied to calculate the Spectral and Cepstral features from the extracted signal. These features are applied to five classifiers for classification. This System gives a maximum of 95% accuracy in identifying a Singer in a video song using SMO classifier This System restricts the identification for 10 seconds period for experimental purpose. This can be extended for the entire song. In future, this work can be extended to cover more Singers. Size of the dataset can also be increased by including more contributing features from the audio track. This may increase the accuracy percentage.

REFERENCES

Chandwadkar DM, Sutaone MS, Role of Features and Classifiers on Accuracy of Identification of Musical Instruments, in Proceedings of CISP, 2012.

Cook P.R, Identification of Control Parameters in an Articulatory Vocal Tract Model, with Applications to the Synthesis of Singing, Ph.D. Thesis, Stanford University, Stanford, CA, 1990.

Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto and Sandro Ridella, The 'K' in K-fold Cross Validation, in ESANN, proceedings, 2012.

George Tzanetakis, Perry Cook, Musical Genre Classification of Audio Signals, IEEE Transactions On Speech And Audio Processing, 10 (5), 2002.

Metilda Florence S, Mohan S, A Novel Search Engine for Identifying Musical Instruments in a Video File, in International Journal of Applied Engineering Research,10 (14), 2015, 34144-34148.

Metilda Florence S, Mohan S, Automatic Video Annotation for Music Genre Based on Spectral and Cepstral Features, in ELSEVIER Proc. Int. Conf. on Applied Information and Communications Technology, ICAICT, 2014, 27 – 32.

Metilda Florence S, Mohan S, Automatic Video Annotation for Music Mood using PCA with Rhythm and Cepstral Features, in ELSEVIER Proc. Int. Conf.on Emerging Research in Computing, Information, Communication and Applications, ERCICA, Bangalore, 2014, 355 - 360.