Music Genre Classification and Retrieval using MFCC

Aarti Singh[1] Rutvij Gadhiya[2] Aditya Sawant[3]
[1,2,3]Department of Electronics & Telecommunication Engineering
[1,2,3]Thakur College of Engineering & Technology, Mumbai, India

Abstract— With the advancement of technology in today's era there is a need of a efficient Music Information Retrieval (MIR) system which can organize and search through the large number of music files available on the internet. It addresses the problem of querying and retrieving certain music files from the large music data set. Classification is the most fundamental and essential component of the MIR system. The other essential part is selection of features and classifier for developing an accurate and efficient content based retrieval system. The retrieval can be done on the basis of text, genre, artist etc. In this paper genre based retrieval is considered. There are various features in a music signal and depending upon the features there are various feature extraction methods, also there are various classifiers. In this paper various features along with different feature extraction method and different classifiers are mentioned.

Key words: Classification, Content Based Retrieval, Feature Extraction

I. INTRODUCTION

Music types permit the classification of music into general gatherings of related melodies for use in applications, for example, tune recommender frameworks dependent on type. Many existing melodies are unlabelled[1], and new tunes are continually being discharged. While hand-naming them all eventual excessively monotonous, arranging them by kind can fit them into classifications for use in suggestion. Likewise, regardless of whether the craftsman or maker names a tune as a particular kind, finding the melodies that are most comparative as far as recurrence structure instead of emotional marking may give proposal data that the client would appreciate more.

Since the former age, people used to easily categorize, organize and search through the music archives based on certain distinguishing levels; as music clips belonging to the same level would show similar characteristics. There are various distributions of levels that can be used for exhibiting retrievals task. Melody or Instrument level find the music clips containing specific melody fragment or that sounds like a reference music excerpt or that has the same instruments played for composing the music clip. The next is Artist level or composer or Performer level which find music clips by a specific artist or a performer or a specific composer, Similarly there is Emotion level which find music clip according to the emotion of the person, and the last is the Genre[2] and Title[3] level which finds the music clip having same title and genre category. Out of this level specified the most commonly used is the genre level which is used to organize the music store and libraries. Since music clips belonging to the same genre would mean that they have certain common characteristics as they are framed by using similar types of instruments or have identical rhythmic patterns, the Retrieval Task can be performed on generic level efficiently.

II. RELATED WORK

There is an need to closely examine the theoretical analysis of various parameter and technique used in the field of music Retrieval The chapter gives a detailed study of performance evaluation carried out by various researchers so as to gain a generalized idea about the importance and scope of the project And also this chapter will help us decide the Method by which the implementation will occur.

Trisiladevi c nagavI[4] proposed a paper on content based Audio retrieval using MFCC(Mel frequency cepstral coefficients) in which a technique to build a retrieval system was specified using Acoustic similarity and also used short-merge technique. In this paper clustering was also used along with short merge for better Retrieval. The feature coefficients using MFCC were obtained and mean of this feature coefficients were computed for all audio files. The database was created by singing songs from professional and non-professional singers. The feature obtained from MFCC were clusted according to the similarity between them. The retrieval accuracy for MFCC and Clustering is 70% with a response time of about 1*2 minutes and MFCC with clustering and short merge is 85% with a response time of 2-3 minutes.

Nilesh.M.PatiL[5] proposed a paper on Content Based audio Classification and Retrieval: A Novel approach, were he discussed the Research on MIR (Music Retrieval).Various Feature Technique and Classification is discussed .various application on content based audio is stated .A novel Approach was proposed in this paper to get a better retrieval accuracy . The approach uses EMFCC as a feature extraction method and PNN as a classifier.

J. Su, T. Hong and Y. chen [6]proposed a paper on fast Music Retrieval with Advanced Acoustic Features. To get a high performance search the paper proposes a algorithm that fuses Depth-Search Strategy and advanced MFCC. The advanced MFCC is more effective as it represent acoustic feature. As the proposed system consist of both depth-search and low-level features it is more effective the system which proposes only low level features. The set was taken from the web which consisted of 30 genre music signals with 15 music per genre. The system was executed with the help of C++.

Babu Kaji Baniya, J. Lee, and Z. N. Li[7] in his paper of Music genre classification using Support vector machine in which he adopted four groups of audio features viz, Dynamic, rhythm ,spectral and harmony for the genre classification. Further this features were integrated using mean , standard deviation (low order moment)and skewness, kurtosis(high order moment) & also considered covariance matrix. After performing two stages of low and high order moments the feature dimension were increased so they were controlled by two feature reduction method viz, Minimum Redundancy Maximum Relevance(MRMR) and Principle Component Analysis (PCA). 63% feature's from the

complete feature set were selected for genre classification. The accuracy obtained by using this reduction technique with SVM was 87.9% for MRMR and 78.2% for PCA. Since MRMR performs maximum relevance from complete set of feature it gives good accuracy with SVM.

Changsheng Xu, N. Maddage, and X. Shao [8]proposed a paper on music Genre classification using Support vector Machine. Dataset contained 100 music signals collected from different places and had 4 types of genres viz, Jazz, pop, classical and rock Feature were extracted using Beat spectrum, linear prediction coefficients, zero crossing rate, short time energy and MFCC. It used 3 nonlinear SVM Classifiers to obtain the boundaries between classic/jazz and pop/jazz and classic/jazz. Different SVM classifiers uses different features therefore many feature extraction method were used. It was proposed as a better technique then Euclidean distance. This method was also applied to other classifiers but a significantly higher accuracy with minimum error was obtained using SVM.

III. Music Information Retrieval

With the advances in the field of information technology, there has been an enormous increase in production and storage archives for digital audio, images and videos. Human beings are capable of distinguishing different types of audio. As given an music query human can easily identify type, mood and similarities with other music. This human based classification of music content is limited to size of database, bigger the database lesser is the accuracy. Music information retrieval (MIR) is one of the best solution to existing methods. MIR systems are used to classify music files to some music information and search similar music files from the large database that are available. Two types of MIR methods are described below

A. Text based Music Retrieval:

This method is also known as ‘’Tag based music retrieval’’[9] or ‘’Query-by-tag’’. The functioning of this method is based on textual metadata. There are two types of metadata, one is factual metadata that deals with objectives (viz. artist, year of publication, album, etc.) and the other is cultural metadata that content subjective concepts of music (viz. mood, emotion, genre etc.). Tag based retrieval system searches similar music from the database using text query, where text query are generally assigned to every music clip in the database. Factual information can limit the utility of metadata. Retrieval efficiency based on factual metadata gets hampered on incorrect spelling of keywords. The cultural metadata is also less effective method for retrieval tasks as the information provided by the tags is inconsistent.

B. Content based music retrieval:

Content based music retrieval[10] method is the promising solution for the problems of tag based music retrieval method. It allows the user to easily find the music clips. Content based music retrieval approach works on entire music clip. Given the instructions the aim of this method is to automatically retrieval all the similar music clip from the database. Hence this method is also referred as ‘’Query-by-example’’. In this method initially , various relevant feature that can represent the raw music waveform are extracted from each clip from the database and are stored in the feature database section in the form of feature vectors. The feature vector of music query is compared with each of the feature vector from the feature database. The feature vector which exhibits a closet match with the feature vector of the query music is selected and its corresponding music file is retrieved.

IV. Database

The database available with sufficient music clips is the database created by G.Tzanekatis and P.cook known as GTZAN database[11]. The database consist of 1000 music clips of each half a minute. This music clips are distributed in 10 genres and each genre consists of 100 music clips. The ten genre are rock, classical, jazz, Blue, pop, metal, disco, reggae, hip-hop, country. The tracks have a sampling rate of 22050 Hz.

V. Features and Feature Extraction Method

A. FEATURES:

1) Timbre:

In music, timbre, otherwise called tone colour or tone quality is the apparent sound nature of a melodic note, sound or tone. Timbre recognizes distinctive kinds of sound creation, for example, choir voices and melodic instruments, for example, string instruments, wind instruments, and percussion instruments. It likewise empowers audience members to recognize distinctive instruments in a similar class (for example an oboe and a clarinet, both woodwind instruments). This type of feature can be extracted using LPC(linear prediction Coefficients),MFCC(Mel frequency Cepstral Coefficients) etc.

2) Pitch:

Pitch is a perceptual property of sounds that permits their requesting on a recurrence related scale, or all the more generally, pitch is the quality that makes it conceivable to pass judgment on sounds as "higher" and "lower" in the sense related with musical melodies. Pitch can be resolved just in sounds that have a recurrence that is clear and stable enough to recognize from noise. Pitch is a noteworthy sound-related trait of melodic tones, alongside span, tumult, and timbre. Pitch might be measured as a recurrence, yet pitch is definitely not a simply objective physical property; it is an emotional psycho acoustical characteristic of sound.

3) Rhythm:

Rhythm generally means a "development set apart by the managed progression of solid and frail components, or of inverse or diverse conditions". This general importance of customary repeat or example in time can apply to a wide assortment of repetitive regular marvels having a periodicity or recurrence of anything from microseconds to a few seconds (likewise with the riff in a stone music melody); to a few minutes or hours, or, at the most outrageous, even over numerous years.

4) Tempo:

In musical terminology, tempo is the speed or pace of a given piece. In classical music, rhythm is regularly demonstrated with a guidance toward the beginning of a piece (frequently utilizing ordinary Italian terms) and is normally estimated in beats per minute (or bpm). In present day traditional

structures, a "metronome mark" in beats every moment may enhance or supplant the ordinary rhythm stamping, while in current classes like electronic move music, beat will commonly basically be expressed in bpm.

5) Dynamics:

In music, the dynamics of a piece is the variety in loudness between notes or expressions. Elements are demonstrated by explicit melodic documentation, frequently in some detail. Nonetheless, elements markings still require understanding by the entertainer relying upon the melodic setting: for example a piano stamping in one a player in a piece may have very extraordinary target din in another piece, or even an alternate area of a similar piece. The execution of elements additionally stretches out past tumult to incorporate changes in timbre and now and again rhythm rubato.

B. FEATURE EXTRACTION METHODS:

1) MFCC(Mel frequency cepstral coefficient):

In sound processing, the mel-frequency cepstrum (MFC)[12] is a portrayal of the transient power range of a sound, in view of a straight cosine change of a log control range on a nonlinear mel size of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that all things considered make up a MFC. They are gotten from a kind of cepstral portrayal of the sound clasp (a nonlinear "range of-a-range"). The distinction between the cepstrum and the mel-frequency cepstrum is that in the MFC, the frequency groups are similarly divided on the mel scale, which approximates the human sound-related framework's reaction more intently than the directly dispersed recurrence groups utilized in the ordinary cepstrum. This recurrence distorting can take into consideration better portrayal of sound, for instance, in sound compression. MFCCs are ordinarily utilized as highlights in discourse acknowledgment frameworks, for example, the frameworks which can consequently perceive numbers spoken into a phone.

MFCCs are additionally progressively discovering utilizations in music data recovery applications, for example, sort order, sound comparability measures, etc.

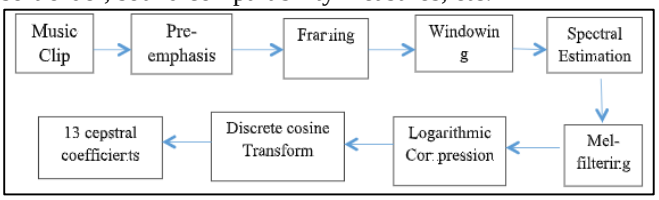

Fig 1: Block Diagram of Mel Frequency cepstral Coefficients

2) PLP(Perceptual linear prediction):

The Perceptual Linear Prediction PLP model developed by Hermansky. PLP models the human speech based on the concept of psychophysics of hearing. PLP discards irrelevant information of the speech and thus improves speech recognition rate. PLP is identical to LPC except that its spectral characteristics have been transformed to match characteristics of human auditory system. PLP approximates three main perceptual aspects namely: the critical-band resolution curves, the equal-loudness curve, and the intensity-loudness power-law relation, which are known as the cubic-root.

The PLP speech analysis method is more adapted to human hearing, in comparison to the classic Linear Prediction Coding (LPC). The main difference between PLP and LPC analysis techniques is that the LP model assumes the all-pole transfer function of the vocal tract with a specified number of resonances within the analysis band. The LP all-pole model approximates power distribution equally well at all frequencies of the analysis band. This assumption is inconsistent with human hearing, because beyond 800 Hz, the spectral resolution of hearing decreases with frequency and hearing is also more sensitive in the middle frequency range of the audible spectrum

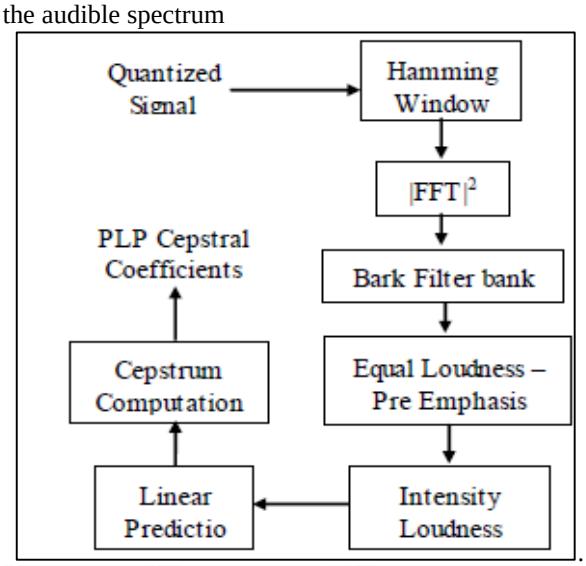

Fig 2: Block diagram of plp

VI. MUSIC CLASSIFIERS

A. SVM (Support Vector Machine):

In machine learning, support-vector machines (SVM) are administered learning models with related learning calculations that dissect information utilized for characterization and relapse examination. Given a lot of training examples, each set apart as having a place with either of two classifications, a SVM[14][15] training algorithm assembles a model that allots new guides to one classification or the other, making it a non-probabilistic parallel straight classifier (despite the fact that strategies, for example, Platt scaling exist to utilize SVM in a probabilistic order setting). A SVM demonstrate is a portrayal of the precedents as focuses in space, mapped with the goal that the instances of the different classifications are partitioned by an unmistakable hole that is as wide as would be prudent. New precedents are then mapped into that equivalent space and anticipated to have a place with a class dependent on which side of the hole they fall.

Notwithstanding performing direct grouping, SVMs can proficiently play out a non-straight characterization utilizing what is known as the part trap, certainly mapping their contributions to high-dimensional element spaces.

At the point when information is unlabelled, administered learning is unimaginable, and an unsupervised learning approach is required, which endeavours to discover characteristic bunching of the information to gatherings, and after that map new information to these framed gatherings.

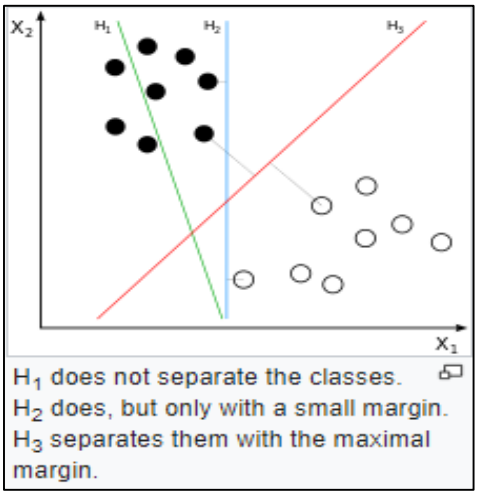

Fig 3.Hyper-plane

B. kNN(k nearest neighbours):

In pattern recognition, the k-nearest neighbors calculation (k-NN) is a non-parametric technique utilized for arrangement and regression. In the two cases, the information comprises of the k nearest preparing precedents in the component space. The yield relies upon whether k-NN is utilized for grouping or relapse:

In k-NN characterization, the yield is a class participation. An article is grouped by a majority vote of its neighbors, with the item being doled out to the class most basic among its k closest neighbors (k is a positive number, normally little). On the off chance that k = 1, at that point the item is essentially allocated to the class of that solitary closest neighbor. In k-NN relapse, the yield is the property estimation for the item. This esteem is the normal of the estimations of its k closest neighbors.

k-NN is a kind of occurrence based learning, or languid realizing, where the capacity is just approximated locally and all calculation is conceded until order. The k-NN calculation is among the least complex of all machine learning calculations.

Both for characterization and relapse, a valuable strategy can be utilized to dole out load to the commitments of the neighbours, so that the closer neighbours contribute more to the normal than the more far off ones. For instance, a typical weighting plan comprises in giving each neighbour a load of 1/d, where d is the separation to the neighbour.

The neighbours are taken from a lot of items for which the class (for k-NN grouping) or the article property estimation (for k-NN relapse) is known. This can be thought of as the preparation set for the calculation, however no unequivocal preparing step is required.

A characteristic of the k-NN calculation is that it is touchy to the nearby structure of the information.

C. HMM(Hidden Markov Model):

Hidden Markov Model (HMM) is a statistical Markov demonstrate in which the framework being displayed is thought to be a Markov procedure with imperceptibly (for example concealed) states.

The Hidden Markov model can be spoken to as the least complex powerful Bayesian system.

In less difficult Markov models (like a Markov chain), the state is specifically obvious to the eyewitness, and along these lines the state progress probabilities are the main parameters, while in the concealed Markov demonstrate, the state isn't straightforwardly unmistakable, however the yield (as information or "token" in the accompanying), reliant on the state, is noticeable. Each state has a likelihood appropriation over the conceivable yield tokens. Hence, the succession of tokens created by a HMM gives some data about the grouping of states; this is otherwise called example hypothesis, a subject of language structure acceptance.

A Hidden Markov model can be viewed as a speculation of a blend display where the shrouded factors (or inactive factors), which control the blend part to be chosen for every perception, are connected through a Markov procedure as opposed to autonomous of one another. As of late, concealed Markov models have been summed up to pairwise Markov models and triplet Markov models which permit thought of progressively complex information structures and the demonstrating of nonstationary information

VII. PROPOSED SYSTEM:

The proposed system can be shown above, the whole system will be divided in three parts. The input to the system is given from the GTZAN [11]database, the music signal is divided into frames and then each frame give to extract the feature, the feature method is selected from the above mentioned in the 4.2 section, the feature vector is given to classifier which generates a classified output which is treated as query song, and using this query song the Retrieval process is initiated.

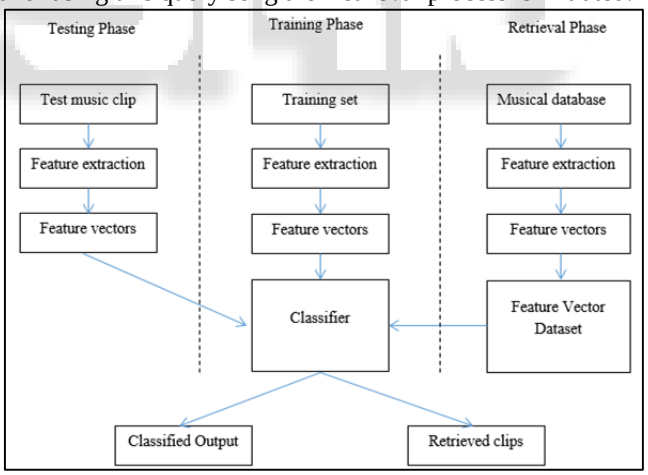

Fig 4: Block Diagram of the proposed method

VIII. RESULT

We used MFCC as Feature extraction technique and SVM as the classifiers and got the Accuracy as 56%. The evaluation is showed below.

We analysed the classification on different parameters of SVM, such as kernel function, value of gamma, value of C etc. The kernel function such as Linear are considered, also the value of C and gamma are considered different for each kernel function. Total 200 test samples are considered, 20 from each genre. The Fig 8-3 shows the confusion matrix of testing data with linear kernel.

It can be noted from the Fig5 that out of 20 songs given in each genre more than half off of songs are classified correctly which resembles to 56-60% accuracy in each of the genre levels with a 95% accuracy for classical genre. Also from the table it can be noted that the average classification for rock, country and disco is much less compared to all the other genre level in the GTZAN database. The perfect Value of C and Gamma is obtained by tuning the SVM model.

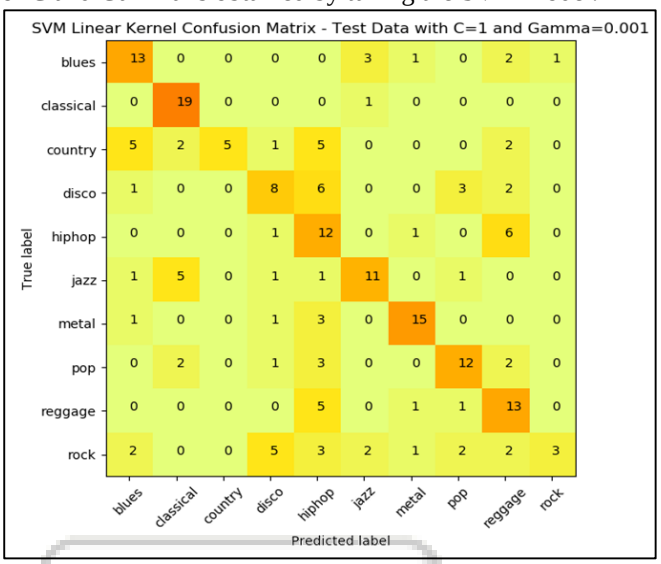

Fig 5: SVM linear kernel confusion matrix

The Fig 6 shows the Accuracy and Classification report for the linear kernel, It can be seen that precession of country genre level is the highest and for hip-hop is the lowest among all the other genre levels, also the recall and f1-score is highest for the classical and it is lowest for the rock genre.

```
accuracy:
55.50000000000001
Classification Report:
              precision    recall  f1-score   support

       blues       0.57      0.65      0.60        20
   classical       0.68      0.95      0.79        20
     country       1.00      0.25      0.40        20
       disco       0.44      0.40      0.42        20
      hiphop       0.32      0.60      0.41        20
        jazz       0.65      0.55      0.59        20
       metal       0.79      0.75      0.77        20
         pop       0.63      0.60      0.62        20
      reggae       0.45      0.65      0.53        20
        rock       0.75      0.15      0.25        20
```

Fig 6: Classification report of linear kernel.

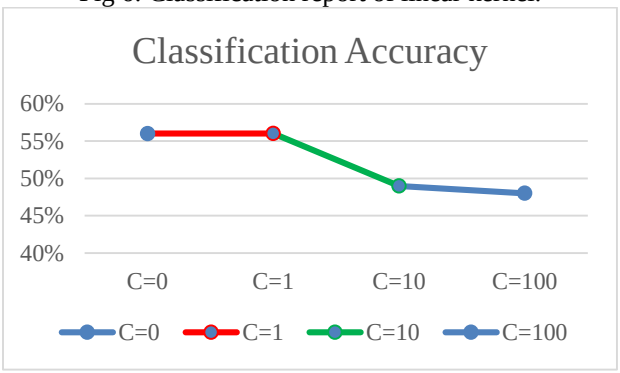

Fig 7: Comparative analysis for value of 'C' for Linear Kernel

IX. Conclusion and Future Scope

The survey gave an idea about Different features and the method through which they can be extracted. It also gave idea about the dimension reduction technique and variety of classifiers. To select the right database is very essential as it will decide our response time for the Music Retrieval. Also the length of the feature vector should not be large as larger the length of the feature vector the greater is the effect on the retrieval efficiency .The Classifier should also be selected that it effective when combined with the feature extraction Technique.

As this work is concentrated only on GTZAN dataset with 1000 songs and 10 genres, it can be extended to various other databases having large number of songs.Also music genre classification based on mixed genre level can be implemented.he database used in this work included only western music clips, the music retrieval task for ethnic music can also be implemented. Apart from genre level, music retrieval based on other musical sense like emotions or artists can be implemented.

Reference

[1] J. G. A. Barbedo and A. Lopes, "Automatic genre classification of musical signals," *EURASIP J. Adv. Signal Process.*, vol. 2007, 2007.

[2] C. Xu, N. Maddage, and X. Shao, "Musical genre classification using support vector machines," *Acoust. Speech, Signal Process. 2003. Proceedings. (ICASSP '03)*, vol. 5, no. March, p. V-429-32 vol.5, 2003.

[3] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimed.*, vol. 13, no. 2, pp. 303–319, 2011.

[4] TRISILADEVI and NAGAVI, "Content based audio retrieval with mfcc feature extraction , clustering and sort-merge techniques," *4th ICCCNT*, 2013.

[5] D. M. U. N. NIlesh M.Patil, "Content-Based Audio Classification and Retrieval : A Novel Approach," no. 2016 International Conference on Global Trends IN SIgnal Processing,Information and Communication, 2016.

[6] J.-H. Su, T.-P. Hong, and Y.-T. Chen, "Fast music retrieval with advanced acoustic features," in *2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW)*, 2017, pp. 357–358.

[7] B. K. Baniya, J. Lee, and Z. N. Li, "Audio feature reduction and analysis for automatic music genre classification," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 2014–Janua, no. January, pp. 457–462, 2014.

[8] C. Xu, N. Maddage, and X. Shao, "Musical genre classification using support vector machines,*AcoustSpeech, Signal Process. 2003. Proceedings. (ICASSP '03)*, vol. 5, no. May, p. V-429-32 vol.5, 2003.

[9] N. Hariri, B. Mobasher, and R. Burke, "Personalized Text-Based Music Retrieval," 2013.

[10] Y. V. S. Murthy and S. G. Koolagudi, "Content-Based Music Information Retrieval (CB-MIR) and Its Applications toward the Music Industry," *ACM Comput. Surv.*, vol. 51, no. 3, pp. 1–46, Jun. 2018.

[11] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," no. 11, pp. 1–29, 2013.
[12] P. P. Singh, P. Rani, A. Professor, and R. Scholar, "An Approach to Extract Feature using MFCC," *IOSR J. Eng. www.iosrjen.org ISSN*, vol. 04, no. 08, pp. 2250–3021, 2014.
[13] K. Xu, W. Hu, and Y. Wang, "An improved singer's formant extraction method based on LPC algorithm," in *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2017, pp. 1–5.
[14] E. Osuna and J. Platt, "Support Vector Machines," *Featur. Sel. Ensemble Methods Bioinforma.*, pp. 68–116.
[15] V. Elaiyaraja and P. Meenakshi, "Audio Classification Using Support Vector Machines and Independent Component Analysis," vol. V, no. 1, pp. 34–38, 2012.

