Content-Based Music Information Retrieval (CB-MIR) and Its Applications toward the Music Industry: A Review

Y. V. SRINIVASA MURTHY and SHASHIDHAR G. KOOLAGUDI, National Institute of Technology Karnataka (NITK)

A huge increase in the number of digital music tracks has created the necessity to develop an automated tool to extract the useful information from these tracks. As this information has to be extracted from the contents of the music, it is known as content-based music information retrieval (CB-MIR). In the past two decades, several research outcomes have been observed in the area of CB-MIR. There is a need to consolidate and critically analyze these research findings to evolve future research directions. In this survey article, various tasks of CB-MIR and their applications are critically reviewed. In particular, the article focuses on eight MIR-related tasks such as vocal/non-vocal segmentation, artist identification, genre classification, raga identification, query-by-humming, emotion recognition, instrument recognition, and music clip annotation. The fundamental concepts of Indian classical music are detailed to attract future research on this topic. The article elaborates on the signal-processing techniques to extract useful features for performing specific tasks mentioned above and discusses their strengths as well as weaknesses. This article also points to some general research issues in CB-MIR and probable approaches toward their solutions so as to improve the efficiency of the existing CB-MIR systems.

CCS Concepts: • **Information systems → Music retrieval**; **Information retrieval**; **Multimedia and multimodal retrieval**; **Content analysis and feature selection**; **Speech/audio search**;

Additional Key Words and Phrases: Artist identification, indian classical music, instrument identification, music annotation, music genre, music mood estimation, music recommendation system, music related features, open problems in music information retrieval, query-by-humming/singing, segmentation of vocal and non-vocal regions, survey of music information retrieval

ACM Reference format:
Y. V. Srinivasa Murthy and Shashidhar G. Koolagudi. 2018. Content-Based Music Information Retrieval (CB-MIR) and Its Applications toward the Music Industry: A Review. *ACM Comput. Surv.* 51, 3, Article 45 (June 2018), 46 pages.
https://doi.org/10.1145/3177849

1 INTRODUCTION

Advances in technologies such as networks, compact discs (CDs), cloud storage, and so on have created enormous data in various forms leading to the need for tools to simplify or automate the process of information extraction. One such approach is content-based information retrieval

Authors' addresses: Y. V. S. Murthy, National Institute of Technology Karnataka (NITK), Department of CSE, Surathkal, Mangalore, 575025, India; email: urvishnu@gmail.com; S. G. Koolagudi, National Institute of Technology Karnataka (NITK), Mangalore, India; email: koolagudi@nitk.edu.in.

ACM 0360-0300/2018/06-ART45 $15.00
https://doi.org/10.1145/3177849

(CBIR). This approach can be used to handle queries related to multimedia data since the present search engine mechanism, which is based on keywords, may fail to handle queries related to multimedia, as it is difficult to form a textual query for multimedia data such as image, audio, and video. Development of a system based on query by multimedia is the possible solution for such cases. Over the past decades, CBIR has been one of the important research areas, and many useful tools have been developed using this approach [246]. Image compass and query-by-image content (QBIC) are the most popular techniques for extracting information based on similarity measures of the contents. Similarly, music information retrieval (MIR) technology helps to accomplish the task of extracting needed information from an audio signal. Generally, an audio signal[1] is highly complex in nature since it contains a lot of information related to artists, genre, emotions, instruments, raga, repeating patterns, and so on. Considering the varieties of music information, there is a large scope for research on content-based MIR (CB-MIR).

CB-MIR has many applications such as music indexing and cataloging, personalized music collection, music recommendation, music classification, copyright protection, and so on. For developing these applications, it is important to obtain meta-information about an audio signal. Since many of the audio tracks do not have proper meta information, the CB-MIR is expected to address the following issues:

— Identification of the presence of vocal regions in a music clip.
— Recognition of the performing artist, gender, song category (instrumental/solo/duet/trio/chorus), singer tracking, composer information, and so on.
— Classification of the music clip into its relevant genre such as rock, pop, hip-hop, folk, classical, and so on.
— Identification of raga of a song from a variety of Indian classical music (ICM). Raga is treated as a melodic framework in ICM, which is built by the varying pitch of the singers voice.
— Conversion of an audio clip into an equivalent text (annotation) so as to identify the lyrics and facilitate query by text (QBT).
— Categorization of a song based on emotional patterns of both vocal and non-vocal segments.
— Identification of the class of instruments such as percussion, string, keyboard, and others, so that the songs can be categorized depending on their textures (monophonic, polyphonic, or hetero-phonic).
— Listing of the audio clips based on a given query using similarity measures, such as query by humming (QBH) and query by example (QBE).

A general CB-MIR system is expected to process a music clip and extract some or all of the above mentioned meta-information. Research outcomes on CB-MIR have been published since the beginning of the 21st century[2] [171]. However, there has been a rapid growth in this field due to the efforts of the Music Information Retrieval Evaluation eXchange (MIREX), a music research community that evaluates the works related to MIR tasks. This activity is coordinated by International Music Information Retrieval Systems Evaluation Laboratory. MIREX was started in 2005 and, until now, it has received around 1,700 papers on various tasks of MIR,[3] shown in Figure 1.

There are a few survey articles available in the literature that are already published on MIR tasks [23, 62, 116, 196, 197, 212, 233]. The first review article on various MIR systems designed to measure the similarity was published in 2005 [233]. Later, in 2006, another review article focused on music genre and features with respect to genre [196]. However, the work has not focused on the

[1] The words *audio signal* and *music signal* are used interchangeably in this article.
[2] This information is based on the papers that are published in the area of music information retrieval.
[3] http://www.music-ir.org/mirex/wiki/MIREX_HOME.

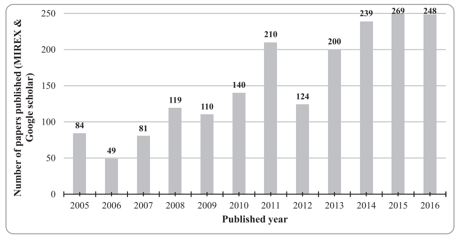

Fig. 1. Increasing trend of research in CB-MIR. Source: Journal and conference articles from http://www.music-ir.org/mirex/wiki/MIREX_HOME and http://scholar.google.com based on MIR keyword.

important aspects of MIR tasks. Consolidation of the combinations of features and classifiers useful for the tasks of music classification is also another focus area of the research community. The use of certain relevant features and classifiers are briefed in another review article, which mainly focused on music annotation [62]. Music annotation tags every portion of a music clip based on gender, mood, artist, style, instruments used, and so on [230]. In addition, these tags can be directly used to optimize the searching process. One other review article has discussed the various approaches for the QBH task [116]. Different variations of dynamic time warping (DTW) are discussed in this article. Song et al. superficially reviewed the works related to general music recommendation tasks [212]. Similarly, another article also briefly explained the works of vocal and non-vocal segmentation [23]. However, this survey does not cover all the issues of vocal and non-vocal segmentation tasks. The very recent article published by Reference [197] studied the various features and classification techniques that are helpful in extracting contextual information. Major contributions are on the works that are proposed in various competitions such as MIREX, MusiClef [172], and the Million Song Dataset (MSD) challenge [18]. This work has mostly focused on the tasks based on music similarity measures. The details of reviewed articles are clearly illustrated in Table 1.

In the present article, the essentials of basic MIR tasks such as vocal/non-vocal segmentation, artist identification, genre classification, raga identification, QBH, music emotion classification, instrument identification, and music annotation are considered. Along with this task-specific information, the article also covers general features and classifiers used in MIR. The applications of each MIR task toward the music industry are discussed in detail, especially for music indexing and recommendation. The existing research works of the CB-MIR field have concentrated less on Indian music. As Indian music contributes to a significant portion of the digital world, it is essential to develop a sophisticated MIR system that automatically retrieves the information needed. The article also focuses on the important problems that are yet to be solved in MIR for Indian music along with general aspects of MIR. The main intention behind this is to motivate upcoming MIR researchers to work on Indian music as well. Moreover, a separate section has been dedicated to raga identification since raga is a base for ICM. The section also discusses the fundamental details and applications of ICM.

The remaining parts of the article are organized as follows: The available datasets for various tasks of MIR are categorized and detailed in Section 2. Section 3 contains the features and various classifiers used for audio information retrieval. The approaches for vocal and non-vocal segmentation are discussed in Section 4. Section 5 addresses the issues in artist identification. The works of genre classification are reviewed in Section 6. The works on ragas of ICM and the related issues are reviewed in Section 7. The similarity measures for comparing music clips and their applications to QBH are reviewed in Section 8. The works on music emotion recognition, instrument identification, and music annotation are subsequently discussed in Sections 9, 10, and 11, respectively. Section 12 lists some important future research directions and Section 13 concludes the review article.

Table 1. Already Available Survey Articles, Their Scope and Limitations

Sl. No.	Reference	Covered Articles/Areas	Scope/Limitations
1	[233]	The authors have reviewed the systems that were implemented for finding the similarity and music indexing. Some of applications that were discussed are Audentify!, C-Brahms, CubyHum, Musipedia, Shazam, Sound Compass, Super-MBox, and Theme-finder.	The focus is only on the applications mentioned in the list and ignored all other systems developed for tasks of music information retrieval.
2	[196]	Music genre is one approach that is universally used for music cataloguing and indexing. A review on the existing works of genre extraction is done in the article.	The article was mainly focused on only the genre, ignored all other aspects of music classification.
3	[246]	The review has provided the solutions and open problems for music classification and the authors have concentrated on music similarity, structure analysis, cognitive psychology, and transcription.	The article mainly focused on the systems developed for similarity and transcription although there are several approaches for music classification.
4	[62]	The review has mainly concentrated on the works done for annotating music clips. In addition to that, some other important MIR tasks are also discussed. It is also possible to find the categories of features and classifiers considered for MIR tasks.	Only few important MIR tasks were selected and an overview was given for them.
5	[116]	The review was focused only on the applications of query by humming.	Note-based matching is the main focus and a very few methods were discussed.
6	[212]	The approaches for developing music recommendation systems are first-time reviewed. The review had given information about the types of methods and user modeling techniques.	However, the techniques available for music recommendation and their definitions have been provided with this article instead of existing research and scope.
7	[23]	Signal-processing approaches for segmenting vocal and non-vocal regions were reviewed in the article.	The survey has concentrated on earlier methods for segmentation. Very few articles are considered for review.
8	[197]	The survey has focused on the recent approaches and possible future directions for music similarity and indexing along with the open problems.	Mainly concentrated on similarity approaches with specific features.
9	**This Article**	This article will give complete information about all the possible CB-MIR techniques and will also have provided the applications of each issue toward the music industry. Critical review has been done for each CB-MIR task and also the scope for improvement is discussed. In addition to that, the category of features and their use for CB-MIR is clearly explained.	

Table 2. List of Some Available Datasets for Researchers to Experiment a Specific Task of MIR

Sl.No.	Task of MIR	Dataset(s) Available	Remarks
1.	Instrument Recognition	200DrumMachines†, Drumpt, ffurhmann†, Good-sounds.org, GSD, Holzapfel:onset, IDMT-SMT-Drums, IRMAS, Medley, NSynth, and RWC†.	Of these, RWC dataset is with 50 instruments and highly used in literature. The other datasets contain clips with less number of instrument categories.
2	Vocal, Non-vocal Segmentation/ Source Separation	Bach10, CCMixer†, DREANSS†, IDMT-MT, iKala, JGDB†, Medley†, MIR-1K†, QUASI, and SMD†.	The tracks that are listed here are based on their regional languages. Moreover, the tracks with complex background have been less considered.
3.	Artist Identification	Artist20†, C224a, C3Ka,C49Ka,C111Ka, CorpusCOFLA, FlaBase, Holzapfel:onset†, MIR-1K†, Musiclef†, SASD†, and 1517-artists†.	Though the number of available artists with C111Ka (110,588 artists) is huge, the clips have not been taken with complex background accompaniment.
4	Genre Classification	Ballroom†, Bodhidharma-MIDI, C224a, C3Ka, Coidach†, Extended-Ballroom†, FlaBase, FMA†, GTZAN†, Homburg†, ISMIR2004Genre†, Medley, MSD†, RWC,† and Uspop2002†.	This task has been highly concentrated in the literature. Hence, some useful datasets are already collected. Of these, GTZAN has been considered as a benchmark for many works.
5.	Mood Estimation	Amg1608, DEAM†, DEAP, EmoMusic, Emotify†, GMD, MoodDetector:Bi-modal, MoodDetector:Multi-modal, and MoodSwings†.	The datasets on music mood estimation are less focused. Emotify is the one which provides nine categories of moods.
6.	Query-by-Humming/ Singing	ACM_MIRUM†, ADC2004, APL, Back10, Ballroom†, DAMP, ENST-Drums, Extended-ballroom†, FlaBase:Fugue, GiantSteps:Tempo, GNMID14, GTZAN†, Hainsworth†, Holzapfel:onset, INRIA, ISMIR2004Tempo†, JordanClassical, JordanJazz, MIREX05Train†, MTG-QBH†, OrchSet, RockCorpus, Sargon, SPAM, UMA-Piano†, UNIQUE†, Uspop2002†, and Zanoni-Giorgi.	Music similarity measurement and QBH are the other important aspects of MIR. The datasets that are available for both the tasks are less focused except ISMIR2004Tempo and MIREX'05.
7.	Onset Detection, Transcription or Raga Identification	Ballroom†, Carnatic Rhythm†, Chopin22, CMMSD, Giant_Steps_Key, GPT†, GTZAN†, Holzapfel:onset, LabROSA-APT†, LackMIDIDataset, MAPS†, MARG-AMT†, McGillBillboard, Mirex06Train†, Modal, MusicNet, SMC:MIREX†, SU-AMT, TONAS, and Zanoni-Giorgi.	Apart from Western music, Indian classical, and Hindustani are also completely dependent on notes that are less focused in literature.
8.	Music Annotation	Beat-box-set1, CAL500†, CAL10K†, Musiclef2012†, OMRAS2†, TagATune†, and Uspop2002.	There are not many useful datasets have been created for the task of music annotation excluding CAL500 and CAL10K.
9.	Others*	AudioSet†, Covers80†, Jamendo, Last.fm†, LFM-1b, MARD, MMTD, PhenixAnechoic, Phonation, PlaylistDataset, QBT-Extended, RWC†, ThisIsMyJAM, TPD, and UrbanSound8k.	The datasets for music recommendation are highly needed. Last.fm is the one which is providing with some limited meta-information.

Note: All the databases mentioned in the table may not be available publicly. Others include the datasets for cover songs, lyrics transcription, query-by-text, and event identification. The symbol indicates that the dataset is publicly available.

2 DATASETS AND THEIR USE IN THE WORKS OF MIR

The difficulty in arranging the tracks into different categories is also increasing, with a daily increase in the number of digital tracks. The task-relevant tracks are essential while developing the MIR system [28]. For instance, different instrument clips are preferable for developing an application for instrument identification instead of clips with audio and polyphonic sounds. In this regard, it is useful if the benchmark datasets are known concerning the subtask of MIR. Identifying task-specific datasets helps the researchers in comparing their works with the state-of-art systems. Moreover, this creates better research facilities for MIR researchers by providing a proper dataset [70].

Many times, the copyright issue of commercial audio clips is a paramount cause that leads to the use of existing datasets. The datasets are publicly available and contain clips with copyright exculpated information. Table 2 shows the list of datasets that were created during the past two decades being used for various MIR tasks. The contents of the list have been prepared based on

Table 3. Highly Utilized and Publicly Available Datasets with Their Detailed Information

Sl. No.	Year	Datasets	Ref.	#Clips	Purpose‡	Sl.No.	Year	Datasets	Ref.	#Clips	Purpose‡
1.	2001	RWC	[70]	465	IR	12.	2008	1517-Artists	[200]	3,180	AI
2.	2002	GTZAN	[217]	1,000	GC	13.	2009	MIR-1K	[85]	1,000	VOD, AI
3.	2003	USPoP	[135]	8,752	QBH	14.	2009	OMRAS2	[58]	1,52,410	MA
4.	2004	BallRoom	[229]	698	GC	15.	2009	TagATune	[74]	25,863	MA
5.	2004	ISMIR2004	[27]	1,458	GC	16.	2010	CAL10K	[223]	10,271	MA
6.	2005	103-Artists	[198]	2,445	AI	17.	2010	UNIQUE	[200]	3,115	QBH
7.	2005	Homburg	[84]	1,886	GC	18.	2011	MSD	[18]	10,00,000	GC
8.	2006	Codaich	[154]	26,420	GC	19.	2012	MusiClef	[172]	1,355	AI, MA
9.	2007	LMD	[208]	3,227	GC	20.	2016	Ext.BallRoom	[148]	4,180	GC
10.	2007	Artist20	[50]	1,000	AI	21.	2017	AudioSet	[66]	20,84,320	AEI
11.	2007	CAL500	[230]	500	MA	22.	2017	FMA	[14]	1,06,574	GC

‡AEI—Audio Event Identification, AI—Artist Identification, GC—Genre Classification, IR— Instrument Recognition, MA—Music Annotation, QBH—Query by Humming, and VOD—Vocal Onset Detection.

different sources, notably wiki, ISMIR, Colinraffel.com, and audio content analysis websites. All the datasets mentioned in the table do not contain precise information. Many of them are not available publicly. Moreover, some clips only contain a limited number of clips. The prominent datasets that are publicly available and highly used in the past two decades have been identified from the list and mentioned in Table 3 with some necessary information.

An effective MIR system can be built if the task-specific benchmark datasets are available. It is observed from the literature that the datasets available are less complex, and recorded with limited scope. The datasets with incomplete and monotonic information are not suitable for many real-time applications. Considering the literature, the genres of eastern countries are less focused, especially Indian categories. As they contribute to a major portion of the digital music world, it is essential to develop a sophisticated MIR systems for them as well.

3 FEATURES AND CLASSIFIERS FOR AUDIO CLASSIFICATION

Audio songs are mainly available in the form of high-quality audio CDs recorded with a sampling frequency of around 44.1KHz at offline and online stores. Direct processing of these high-quality audio songs for information retrieval consumes large memory and processing time. Generally, numeric features are extracted that resemble the signal characteristics and compactly represent the original audio songs. There are an enormous number of features that have been introduced with the support of various signal-processing techniques and statistical methods to simplify the tasks of speech processing. The majority of them are used to characterize the music as well. Some additional features are also introduced to model the music signal in a better way. The first hierarchy of audio features is identified by Reference [196] to produce the survey on genre classification. They also introduced three kinds of features, namely timbre, pitch, and rhythm. Later, the taxonomy was revised by Reference [246] who categorized them into short-term, long-term, semantic, and compositional feature sets. Although the features mentioned in the article are mainly based on few concepts of music research such as music similarity, transcription, and cognitive psychology, they cannot be generalized and used for all MIR tasks. Hence, the two taxonomies have been combined and enhanced in Reference [62] to present a generalized hierarchy of audio features. In the present article, a similar kind of hierarchy with the additional features and their importance for all MIR

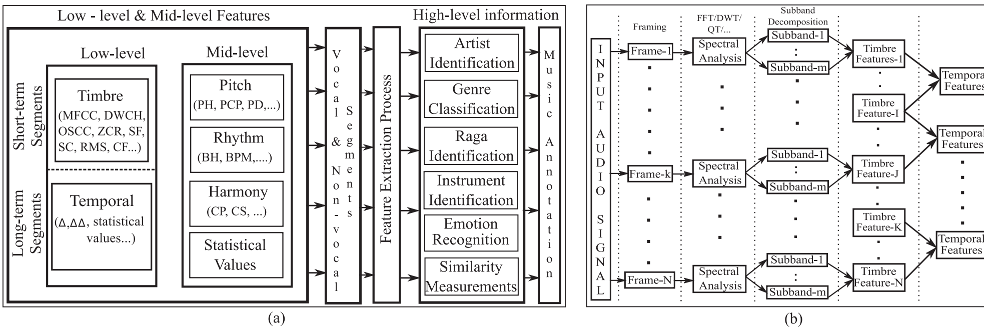

Fig. 2. (a) Audio feature classification as low-level, mid-level, and high-level information. (b) Process of extracting low-level features.

tasks has been given. The features are mainly classified into (i) low level, (ii) mid-level, and (iii) high-level features, shown in Figure 2(a).

In general, the low-level features are extracted from the smaller segments of length 10∼100ms, known as frames. The mid-level features are extracted from a syllable, word, or utterance. Similarly in music, if a feature is extracted at note level or using some set of low-level features, then they will be generally called mid-level features [108]. Low-level features carry abstract characteristics of a frame. They cannot represent the characteristics of an entire signal. Mid-level features provide abstract characteristics of an entire signal or set of segments. They can be computed on longer segments or by applying statistical operations on low-level features [46, 181, 249]. High-level features provide semantic information such as annotated information, which is useful for labeling the clip and helps for easy retrieval. The combination of low- and mid-level features are used to decide the high-level information such as genre, mood, instrument, artist, and so on. The following subsections describe various low- and mid-level features.

3.1 Low-Level Features

These are the very common block-based features that have, in general, shown better characterization for various tasks of music information. They are further classified into timbre and temporal features.

The number of vibrations caused to produce sound waves in a second is known as *pitch* [also called *fundamental frequency (F0)*] of a note. The strength of the signal can be measured by computing the sum of squares of samples called *energy* of the signal. Timbre is the quality of a musical tone which helps to differentiate the voice or instruments even when their pitch and energy are the same. For instance, if guitar and piano are playing the same note at the same scales, then timbre of those instruments helps to classify them. In psycho-acoustics, timbre is defined as the voice quality of a musical note, sound, tone color, or tone quality that distinguishes various kinds of sound sources [52, 184]. Timbre features are generally computed from the short-time segments of length 10–100ms, called *frames*. The main advantage of this approach is the technical simplicity and availability of well-established methods to process stationary signals in terms of effectiveness and complexity. The general process for low-level feature extraction is shown in Figure 2(b). Initially, the input signal is divided into chunks of frames that are converted into frequency domain using various transformations such as Fourier transform, constant Q-transform, wavelet transform, and so on. A sub-band decomposition technique is applied to the frequency domain signal. Each sub-band is analyzed to extract the timbre features of a frame. A combination of timbre features is used to extract temporal features. The low-level features can be extracted from both time and frequency domains. The important features found in the literature that are extracted from time-domain information are root mean square (RMS) energy, zero crossing rate (ZCR) [17, 132, 165, 234], and crest factor (CF)[4] [81]. To analyze the signal in the frequency domain, various transformations such as discrete Fourier transformation [3], discrete wavelet transformation [25], and constant Q-transformations [199] are performed. From the spectrum obtained through transformation, it is possible to extract features like spectral roll-off, spectral centroid (SC), spectral flux (SF), and bandwidth using statistics. Instead of Fourier transformation (FT) on complete signal, the short-time Fourier transformation can be used to extract potential features such as mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs), octave scale cepstral coefficients (OSCCs), Daubechies wavelet coefficient histograms (DWCHs), spectral flatness measure [4, 13], spectral crest factor [32] and amplitude spectrum envelope (ASE) [105, 123]. To obtain MFCCs, the sub-bands during spectrum computation are linearly spaced up to 1000Hz and are logarithmically spaced at higher frequencies. MFCCs can model music patterns better than other spectral features [165]. Moreover, the segments of variable length are used to extract cepstral features based on inter-beat segments, which are more relevant than the traditional equal-sized block-processing approach. This has led to the invention of new kinds of features, i.e., OSCCs [144]. This approach is extended to extract relevant features directly from MP3 files with slight modifications in discrete cosine transformation (DCT); this process is called modified discrete cosine transformation (MDCT). An attempt has also been made to extract the features from both recording channels (left and right) since vocals of both channels are common and the non-vocals vary for most of the times. To identify the spectral distribution in both channels, the stereo panning spectral featureshave been introduced [235, 236]. However, many of the timbre features mentioned above are adopted from the works of speech processing. As there are several variations of speech and music, an effort is yet to be done to extract the distinct timbre effect for efficient analysis.

The temporal variation in the signal helps in several music classification tasks. Temporal features are a kind of low-level feature extracted on top of timbre features. These are helpful to observe the temporal feature transformation of the given signal. Generally, statistical parameters such as mean, variance, co-variance, and kurtosis, which are computed from a large number of local windows, are the features [234]. The means and variances have been computed from a timbral texture to form a feature vector, called *MuVar* [234]. The means of covariance values are computed from a covariance matrix to form a feature vector called *MuCov* [146]. *MuVar* and *MuCov* are also explored in the literature to observe the temporal evolution [132]. The same operations are performed on the frames of larger lengths and named *MuVar*^{2} and *MuCov*.^{2} Normally, the process of computing temporal features considerably increases the computational complexity. Hence, feature integration is done using the other techniques such as amplitude regression, multi-variate auto regression, multi-variate Gaussian regression, and diagonal auto regression (DAR) to reduce the complexity issues [146]. Along with the other available techniques, probabilistic models are also used to extract the temporal features. One such model is hidden Markov models (HMM) [190, 246], which models the time series data using hidden states. In HMM, each frame is treated as a single state that helps to provide the feature set for the current frame based on the output probabilities of the previous frames.

3.2 Mid-Level Features

Human ears can perceive the intrinsic properties of any music with the help of integrated biological mechanism. Low-level features failed to capture much information from a given song clip.

[4]Crest factor is a ratio of amplitude peak and RMS value and is obtained as $CF = \frac{peak(|signal|)}{rms(signal)}$.

Thus, mid-level features are introduced and they are mainly used for the tasks such as QBE, query by singing/humming (QBSH), cover song detection, raga identification, and so on. The three broad categories of mid-level features are: (i) pitch—the fundamental frequency, (ii) rhythm—the recurring pattern of tension, and (iii) harmony—a mixture of notes that are played simultaneously and successively to produce chords and chord progressions [254]. In music processing, pitch plays an important role for different applications such as QBH and raga identification. Other factors such as context, loudness and timbre also influence the pitch. In the musical context, the pitch is not a single F0 since every instrument has its own harmonic frequency series. Multi-pitch estimation is necessary for such cases. Few algorithms [110, 224] were designed especially to estimate the multiple pitch values. These algorithms are helpful in extracting the pitch values at frame level and song level using pitch histograms (PHs). The PHs are used to recognize the genre and mood of a song with the additional support of MFCCs and other perceptual features. Along with the PHs, other features such as pitch class profile (PCP) can also be used. The first note of the *C* major scale is the note *C*. If it is pitched around 261.63Hz, then the low-C and high-C would be around 65.40Hz and 1046.50Hz, respectively. Though there are several variations in pitch frequency, all the variations will be considered as the same pitch class [120]. PCPs and harmonic pitch class profiles are helpful in extracting the chroma (pitch class) features. Chroma features are helpful in analyzing the melody of a song, including *gamakas*.[5]

The occurrence and recurrence of patterns can be discriminated using rhythmic features. These features are mostly helpful in recognizing the repeated pattern in a song clip. The most repeated pattern in any song is known as a beat. The features such as beats per minute (BPM) and tempo are useful to estimate the beat locations. Another way of computing beat features is by taking the envelope of an auto-correlation for a given input signal. The regularity in peaks of the auto-correlation signal helps to compute beat histograms [234]. In the literature, rhythmic features are also used for mood estimation tasks [59, 143]. The results indicate that the mood of a song is highly correlated to the rhythm. It is observed that, normally, each mood is fixed to some value of a scale [251, 252].

The third important feature is harmony, that can be recognized through several factors. Of these, one is chord sequence (CS). Harmony is quite different from melody since melody obtains the horizontal information and harmony obtains the vertical information of a song. Melody is the linear succession of musical notes and is a combination of rhythm and pitch. Harmony is the combination of simultaneous notes or *chords*. The CS can be extracted by some chord-detection algorithms found in the literature [69, 97, 230]. These sequences are also helpful in detecting the multiple fundamental frequency values present in the chord since a chord is the combination of more than one note played together. The harmony features are used in the literature for the cover detection of a song [11] and song similarity [51]. Although mid-level features can capture the intrinsic properties of a music clip such as pitch, BPM, melody, harmony, rhythm, and the like, they alone are sometimes not sufficient enough to achieve good results. The combination of both low-level and mid-level features would give better results in music processing tasks [108]. In pattern recognition applications, it has been difficult to establish a strong correlation between specific tasks and features. In such cases, a set of features is used initially, and, later, feature selection techniques such as elimination, correlation, and so on are applied to reach the optimum feature set [8, 61, 131, 170, 204, 205].

The process of selecting a suitable classification model is the next important step while developing MIR. There are three categories of classification models, namely (i) unsupervised, (ii) semi-supervised, and (iii) supervised. Since the audio data is highly non-linear and a majority of them are classification problems, unsupervised classification models may not handle them effectively.

[5] A *gamaka* is an ornament which gives soothing effect for the *raga* of ICM.

Several supervised classification models such as artificial neural networks (ANNs), Gaussian mixture models (GMMs), support vector machines (SVM), AdaBoost (AB), generalized linear models (GLMs), k-nearest neighbor (KNN), sparse restricted Boltzmann machine (SRBM), and so on have been considered for a variety of MIR tasks. Since the classifier selection is completely dependent on the feature vector constructed, it is highly difficult to suggest a single classifier for the specific task. The performance of the system with different classification models is given in respective sections. For instance, if the data falls under normal distribution, then GMM is the better classfier in such cases.

4 VOCAL/NON-VOCAL SEGMENTATION

An audio signal is a combination of pure vocals, instrumental region, silent regions (SIL), and vocals with background instruments. Since a majority of the users are interested in listening to popular songs, identifying the popular song structure is an interesting research assignment. In this subsection, two important issues are observed while processing audio clips, which are discussed along with their possible solutions. Identification of SILs is the foremost pre-processing step in any speech and audio-processing task. Generally, the length of the silence portion in a song is negligible, because more than 99% of the audio song is occupied by either a singing voice or an instrumental sound. The second issue is segmenting the vocal and non-vocal regions as the music signal is the complex cohesion of these two components.

As vocals are usually accompanied with background music, segmentation becomes a challenging task. Segmentation is a prerequisite for singer identification, emotion recognition, instrument classification, lyrics transcription, and so on. One of the interesting commercial applications of vocal and non-vocal segmentation is the *karaoke* system. *Karaoke* is a Japanese word, which means only music track without vocals. This is helpful for music enthusiasts to learn singing for many existing compositions or to use the tracks in concerts for simulating reality. Presently, the extraction of karaoke tracks is being done manually during recording, which needs a lot of manual effort and time. Segmentation of vocal and non-vocal regions is an essential step in designing an automated *karaoke* system.

For automating segmentation, several approaches have been reported in the literature. Initial attempts were made to analyze the signal in time domain by using simple features such as energy, ZCR, and so on, as these values get a sudden jump when vocal region appears [255]. However, it is not always true since the drum sound comprises high energy components when compared to vocals when vocals are accompanied by background music. In addition, it is understood that analysis of a music signal in its time domain is not sufficient for accurate segmentation. Spectral analysis is also employed.

Different kinds of transformations, including FT, are available to represent time domain signal in frequency domain. It should be noted that the majority of energy of vocals formant falls in the frequency range of 200Hz to 2000Hz. Therefore, suppressing other frequency values helps to locate the singing voice segments. This can be done by using any of the available infinite impulse response filters such as Butterworth, Chebyshev, and so on [106]. This approach can be used to separate the background accompaniment, which helps in locating the vocal segments easily. However, it is very difficult to observe the frequency range of vocals and music in a mixed clip [240]. The change in the shape of a spectrum of vocal and non-vocal regions has created much research interest. By analyzing the spectrum, formants and MFCCs are computed and used to understand the characteristics of vocal and non-vocal segments [191].

In some works, it is found that *fluctogram* gives much information when compared to a spectrogram. The sub-semitone and pitch-continuous fluctuations can be viewed using a simple cross correlation followed by shifting operation. The resultant of this operation forms a new visual representation named *fluctogram*. Features based on the *fluctogram* have been extracted; however, very little effort has been made in that viewpoint [47, 124]. Prominent human formant values are mainly observed in the range of 2–3kHz. Basic cepstral features such as MFCCs also carry important music/vocal information [26, 133, 161]. While computing MFCCs, the length of the frames is always fixed. Frames of variable lengths are introduced [144] based on inter-beat-times to improve the performance of a system, known as OSCCs. The results convince us that the OSCCs are more suitable for music modeling than the MFCCs. Frequency analysis along with temporal behavior is considered for vocal characterization by using Δ (velocity) and $\Delta\Delta$ (acceleration) features of MFCCs. Similarly, there are other features found in literature such as Δlog energy, modulation energy, harmonic coefficients, and delta MFCCs for locating the singing voice [35]. Vibrato in the singing voice is also useful in locating the vocal segments efficiently [151, 166]. The trending deep neural networks (DNNs) are also utilized in some works to separate the source information [209]. Table 4[6] summarizes the research contributions to locate the singing voice segments along with their limitations and scope of improvement.

In a majority of the cases, the task of locating singing voice segments has been considered as a sub-task for singer identification. It is true that the small portion of the singing voice is enough for such tasks. A few works have concentrated on segmenting the complete music that may be helpful for the applications like *karaoke* [166, 206]. Since it takes high computational time for proper segmentation, feature dimensionality reduction is also a necessary task. Some works have concentrated on optimizing the features using feature-selection algorithms that help in selecting the suitable features for locating vocal segments [188]. Nevertheless, room is still open for an accurate system that segments the vocal and non-vocal regions in a given song clip of any kind.

5 ARTIST IDENTIFICATION

Artist identification is one important attribute available with a music clip. Singer identification, recognition of composer, and artist identification of a concert are the variations in artist identification. In a majority of the time, it is possible to observe the unique styles (singing/performing or writing/composing) of an artist while performing. Through the implicit learning capability of humans, they can discern the differences by listening to a sample audio clip [62]. If a person is familiar with a specific singer's tone, it is possible to recognize the singer by a small piece of the audio clip. At present, music stores are utilizing the efforts and expertise of music professionals to label the artist information for the unknown songs of their music databases. However, it is practically difficult to label millions of tracks available in the digital market manually, and, sometimes, it becomes unreliable. The complex audio signals do not give proper artist-specific information by simply looking at them [106]. The difficulty in identifying the singer through signal-level analysis has created an opportunity to develop an automated system for artist identification. The applications of automation of artist identification include music recommendation, cataloging, and indexing. It can also be used in issuing copyrights for tracks to avoid music plagiarism.

Artist identification is a $one - in - n$ class classification problem since it deals with identifying a singer among n possible singers. The difficulty is to handle a large music database. In this scenario, "singer similarity" based approaches are more useful and suitable. In the mutual phase, similar singers may be grouped together by using clustering algorithms. One important observation is that the artists will maintain similar voice patterns and common characteristics while rendering songs, although the occasion is different. This would help in artist identification.

Traditional speech-processing techniques for speaker identification [9, 187] may not be suitable for the task of singer identification. In spontaneous speech, the pitch of a speaker involuntarily

[6]Expansions for the acronyms are given in the Appendix.

Table 4. Summary of Works on Vocal and Non-vocal Segmentation

Sl. No.	Title of the Article	Composition of Database	Feature(s)	Accuracy %	Remarks	Future Scope	Limitations
1	Artist detection in music with Minnowmatch [248].	82 clips (male and female)	FFT values and MFCCs	85.10	Two classifiers, namely SVM and ANN, are used for segmenting singing voice segments.	As FFT gives good discrimination for vocal and non-vocal portions, statistical operations on FFT values may improve the accuracy.	Accuracy of the system comes down with the increase of database.
2	Singer identification in popular music recordings using voice coding features [106].	20 full-length songs	Chebyshev-IIR and Harmonicity	55.40	Chebyshev-IIR filter is applied to enhance vocal regions and attenuate other frequency regions. Later, harmonicity is applied to detect singing voice segments.	Frequency analysis of singer and non-vocal regions may improve the accuracy of detecting singing voice locations.	It is assumed that formant energy always falls below 4KHz. Due to the advancements in technology and music rendering, distinguishable/useful formants may be extracted up to 12KHz.
3	Automatic singer identification [255].	English and Chinese clips	Energy, ZCR, and SF	70.00	A sudden increase in the value can be observed for specified features when singing voice starts.	Identifying similar kind of time-domain features may reduce the complexity issues	The sudden change in the specified values can be found in case of pure vocals. As the background accompanies vocals in a majority of vocals, the approach could not be practical.
4	Singer identification based on vocal and instrumental models [144].	110 tracks (English and Chinese)	OSCCs	83.58	Inter-beat frames are considered instead of fixed-size frames to compute cepstral coefficients and named as OSCCs. Better performance is observed with OSCCs when compared to traditional MFCCs.	There is a need to develop a system that can divide the signal into variable length frames instead of shorter and fixed length frames. It may be helpful in reducing complexity issues.	The proposed system is not suitable to identify all the vocal and non-vocal regions. The OSCCs may confuse as to segment using inter-beat segmentation due to vocals involvement.
5	Singing voice separation from monaural recordings [133].	Popular English songs	Intonation and Viterbi algorithm	89.44	Inverse comb filtering is applied to reduce the background accompaniment and, later, vocal frames are identified by observing high energy levels when vocal region starts.	A thorough analysis of filtering techniques may help in reducing the background accompaniment, which further helps in properly detecting the vocal onset detection.	The dataset contains very few songs and may not be sufficient to generalize the results.

(Continued)

Table 4. Continued

Sl. No.	Title of the Article	Composition of Database	Feature(s)	Accuracy %	Remarks	Future Scope	Limitations
6	Automatic singer identification based on auditory features [26].	140 clips (English)	MFCCs	92.10	At first step, low-pass filter is applied to suppress background accompaniment. Sparse representation classifier (SRC) is used to locate the vocal segments. MFCCs are used as features	Reduction of background score and enhancement of the singing voice may increase the performance.	The detailed explanation is not found using the SRC classifier.
7	Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations [166].	300 clips (small and longer) clips	MFCCs, stat{pitch} and vibrato	87.05	Baseline MFCCs, statistical values of pitch, and vibrato features were used to observe the variations in vocal non-vocal regions.	Signal-level analysis on popular songs may give some repeated patterns that may be helpful in locating singing segments.	It is observed that the computational complexity increases if the clip length is longer.
8	A low-latency, real-time-capable singing voice detection method with Long Short-Term Memory (LSTM) recurrent neural networks [124].	149 clips	Fluctrogram Analysis and spectral features	89.06	Fluctrogram is introduced to compute the pitch and available spectral features such as MFCCs, Spectral contraction and flatness are added to improve the performance.	Proper analysis on fluctogram may give suitable temporal features that help in improving the accuracy of vocal onset detection.	For experimentation, only a single genre is considered, which is not sufficient to rely on the approach.

Note: Only some relevant and widely cited articles are listed.

changes with factors such as emotion, loudness, and so on, whereas in singing, controlled pitch modulation is necessary for melody. Singers are trained to vary pitch while rendering music and have control on vocal parameters such as respiratory system, laryngeal muscle activity, articulation, and so on. In simple terms, singers are trained to vary the vocal parameters systematically; this gives an evident reason to recognize the singers through the analysis of voice parameters [22, 204].

In the literature, several techniques have been proposed for singer identification [106, 138, 144, 156, 178, 204, 226, 228, 255, 256] and artist identification [15, 107, 248]. Some of the important approaches for singer and artist identification given in the literature are presented below.

In many works, MFCCs are used as base-line features for singer modeling as they are already well-established features for speaker identification [141, 192]. Compared to speech, the music contains more high-frequency components (many instrumentals) in the frequency range of 200 to 15,000Hz. To have the expected soothing effect, the music signal is maintained at a very high sampling frequency (above 40KHz). A slight modification in MFCC extraction process produces tweaked MFCCs, which are used for artist identification by using complete frequency bandwidth (up to 22,000Hz) [248]. Cepstral mean subtracted MFCCs (CMSMFCCs) have been proposed to improve the classification accuracy as they can capture the variations among singers [178]. These features are computed by subtracting the cepstral mean from each vector of MFCCs. Moreover, the temporal behavior of MFCCs is considered to observe the singing pattern variations among singers through Δ and $\Delta\Delta$ MFCCs [15]. OSCCs have also been proposed for singer identification, where the cepstral features are computed on frames of variable lengths [106, 255], which helps to characterize the harmonic structure of a singer. To compute OSCCs, framing is done based on inter-beat duration rather than traditional fixed-length frames.

In general, specific vibrato and pitch profiles are followed by the singer while performing [219]. Therefore, features that resemble human perception have a high role in many music-processing applications. One such approach is warped linear prediction, where all coefficients are extracted at warped scale [79, 216]. A warped scale is closely related to the logarithmic one and highly resembles the functioning of a human ear. Warped linear prediction coefficients (WLPCs) are used [106] to recognize the singer. The results of the above works convey that the WLPCs exhibit better singer characterization as compared to the conventional Linear Predictive Coding values (LPCs). In general, the same kind of instruments will be used to provide the background for the singer while they are performing in concerts. Hence, the performance of the singer identification may be improved with the combination of non-vocals instead of vocals alone. In some works, the LPCs are utilized to dense the cepstral coefficients for the task of singer identification [256]. From the literature, it may be observed that warped LPC-based cepstral coefficients may be explored further for singer identification.

Primarily, the following points are to be considered while developing an application for singer recognition. Commercially available audio files are always accessible in compressed formants (e.g., MP3), whereas a majority of the works in the literature are experimented on raw files (e.g., *wav*). MPEG Audio Layer-3 (MP3) is one of the techniques used to compress the audio file. Identifying and extracting the features from MP3 clips helps in designing a real-time system for music processing. A few works are only reported in which features are directly extracted from MP3 clips [138]. New features and approaches are essentially required to extract the singer relevant information from MP3 clips. Another important issue in singer identification is locating multiple singers and identifying the overlapped regions. Existing systems are helpful in characterizing and recognizing a single singer. The quantity of duets and trios is much more than the solo songs. Thus, there is a need for the approaches to recognize multiple singers, track the location of singers, and so on. This approach is helpful to those who are learning to sing songs on empty (vocals absent) tracks [64, 106, 226, 227]. A summary of the literature with their limitations and scope in artist identification is depicted in Table 5.

Singing voice mostly occupies a place between the dominant musical instrument and speech [159, 160]. The spectrogram of a singing voice reflects vowels with a harmonic structure. Hence, the harmonicity helps in recognizing the singer from a given clip. At the same time, the features based on articulatory techniques are also helpful in determining the singer as they outperform in speaker identification tasks. The above statements hint at combining music and speech related features to improve the singer recognition accuracy. Considering the fewer efforts, singer identification has to be explored with wider dimensions at least in the context of Indian music. Singing quality of an artist has a direct correlation with one's timbre. Hence, estimating the timbre would benefit the task of singer identification. Moreover, the vocal tract and excitation-level features along with rhythmic features of a performer may further be useful in detecting the singer more accurately.

6 GENRE CLASSIFICATION

Music genre is a concept that categorizes the music clips based on tradition. It can be identified by musical style and musical form [168]. Music genre is another important factor to categorize and index the music clips. At present, a majority of the audio clips of online music stores are organized using their genre information. Genre categorization is usually done based on the intrinsic music patterns and instrumentals. All songs with similar patterns can be grouped into a single genre class. They are relevant because of the musical differences in culture, artist, and composers. The analysis of these properties is essential for music classification. However, identification of the genre is not an easy task for naive listeners, whereas music professionals can do the same by their proficiency and experience [182, 196]. Inconsistencies while categorizing the music based on genre is an important issue.

Processing of speech would be little simpler since it is limited to few clearly characterizable notions such as emotion, language, gender, and so on. In the case of music classification, identifying generally acceptable taxonomy itself is a big issue. The set of principles such as objectivity, independence, similarity, and consistency are helpful in creating a hierarchical taxonomy of genre [174]. Moreover, an album containing different songs may not belong to the same genre. In such situations, each song has to be labeled with a different genre. From the statistics taken from three online music stores, *AllMusic* (http://www.allmusic.com), *Amazon* (http://www.amazon.com), and *MP3* (http://www.mp3.com), it has been observed that a huge number of genre classes are available with no certainty regarding their genre. For Indian music, the musical websites www.raaga.com and www.gaana.com categorize the music clips based on Carnatic, Hindustani, and classical. The Indian music needs to be analyzed for genre taxonomy. The clip that has two natures can be labeled as two genres. For instance, the intrinsic properties of *rockabilly* and *punk* can be merged to form a new genre called *psychobilly*. It leads to the increase in complexity of taxonomy.

This survey article contains some of the following approaches for genre classification [5, 6, 72, 98, 131, 134, 158, 177, 213, 234, 237]. In the initial stage, traditional short-term features are extracted from fixed-length frames of size 10~50ms. In this case, at least a second of a music clip is essential to recognize the genre. There is always a chance of ignoring the information available with longer segments [157, 175]. Thus, feature-integration tasks based on *late information fusion*[7] can be considered on top of short-term features using auto regressive, multi-variate auto regressive [203], and diagonal auto-regressive models [158]. Moreover, temporal behavior in music progression helps to distinguish the patterns of various genres. To observe the temporal behavior, modulation spectral contrast, and modulation spectral valleys are computed and used for genre classification

[7] An integration is based on the outputs of a classifier (e.g. a majority voting).

Table 5. Excerpts of the Articles Published on the Issue of Artist Identification

Sl. No.	Title of the Article	Composition of Database	Feature(s)	Accuracy %	Remarks	Future Scope	Limitations
1	Artist detection in music with Minnowmatch [248].	82 clips (male and female)	FFT values and MFCC	85.10	Artist classification is done with two classifiers. SVM gives good performance when compared with NN for more artists.	Artist's timbre can be detected using the statistical operations on FFT.	Database with fewer artists gives good accuracy.
2	Singer identification in popular music recordings using voice coding features [106].	NECI Minnowmatch testbed	LPC and WLPCs	45.30	Warped scale is introduced and combined with linear scale to extract LPCs.	Features that are extracted using variable-length frames and perceptual scales may be helpful in developing real-time systems.	A little bit of improvement is found when compared to traditional LPCs. However, the mentioned performance may not be sufficient to standardize the system.
3	A singer identification technique for content-based classification of MP3 music objects [138].	200 clips (male and female)	PMCV and FMCV	66.00	DCT is applied on frames of MP3 clips and named as MDCT. Phone- and frame-level features are extracted for experimentation.	Feature extraction on MP3 files (compressed) may be useful for the tasks of MIR, which is to be thoroughly explored.	The database with few clips has been considered for experimentation and less accuracy is observed.
4	Automatic singer identification [255].	45 (English and Chinese) clips	LPC and MFCCs	80.00	Singing voice locations are identified automatically. Further, GMM classifier is used to classify the singers.	Increase in database size and understanding the voice qualities of singers may be helpful for singer identification.	Database size is very small and it may difficult to model all modulations of singers using it.
5	Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals [228].	260 (solo and duet tracks)	MFCCs	82.80 (solo)	Solo and duet clips are considered to extract multiple singers' information.	The system may be extended to locate the singer information and track the singer.	Performance of locating target singer and tracking target singer is not per expectation.
6	Automatic singer identification based on auditory features[26].	140 clips	MFCCs, LPCCs, and GTCCs	90.00	Combination of three cepstral features are used to improve the performance.	As cepstral features are highly correlated to human perception, they can be used to characterize the singer.	It is observed that the performance gets degraded with the increase in the number of singers.
7	Combining evidence from mel-cepstral features and cepstral mean subtracted features for singer identification [178].	500 (14M and 6F)	MFCCs CMSMFCCs	84.50	Cepstral mean is subtracted to observe the temporal variation of singers' information.	Temporal fluctuation estimation may be helpful to identify singer	Vocal locations are manually marked, which may not meet the real-time applications.

Note: Only some relevant and widely cited articles are listed.

[123]. In addition, spectral similarity and fluctuation patterns (FPs) are used to characterize different genres as they capture the temporal behavior of the song [173]. Generally, the music composers and the singers share some common attributes such as rhythmic structure, pitch information on instrumentals, and so on. These common features also help to label the genres. The features like rhythmic content, beat histogram, pitch content, timbral texture, and so on are important features used in this category [234].

In addition to the short-time features, which are not sufficient to capture the global information [132, 234], histogram and wavelet analysis are performed to identify distinct properties of various genre classes [43, 145, 150]. As wavelets are similar to the human perceptual scales, they help to categorize the music as humans normally do [128]. DWCHs are a kind of wavelet feature that can classify the music clips based on the genre with reasonably good accuracy. Performance improvement is observed with these coefficients when compared with short-time cepstral coefficients [129, 130]. As genre is one of the important attributes in many tasks of music classification, efforts are required to fulfill the issues of existing systems. The issues of genre classification from audio clips are summarized in Table 6 for easy understanding.

7 RAGA IDENTIFICATION

7.1 Swaras (Notes) and Their Functions

Raga plays a vital role in ICM, which is an important attribute to classify classical music. In the Indian subcontinent, music is mainly categorized into two types: (i) Carnatic music and (ii) Hindustani music. Region-wise, Carnatic music is believed to be evolved from south India, and Hindustani is from North India, Bangladesh, and Pakistan. These two are heterophonic[8] in nature. Raga and tala are the basic melodic and rhythmic structures for both the categories. The variations in raga and tala are voluminous [241]. The Sanskrit meaning of a raga is *color* or *hue*. Technically, raga is a sequence of melodic atoms (notes) in which the pitch values are modulated with respect to time [119, 195]. All note frequencies of a raga always depend on the base note.

A notation that represents the fundamental frequency of a sound is called a *note*. In Indian music, a *note* is called a *swaram*.[9] Each note is related to one frequency based on the unison (also known as *tonic frequency* or *shadja*). The range of tonic frequency for male singers is from 100Hz to 180Hz, and for female singers it ranges from 160Hz to 280Hz. For popular musical instruments, the range is around 140–200Hz [12]. In general, there are seven swaras in ICM: Sa, Ri, Ga, Ma, Pa, Dha, and Ni. These swaras are nearly similar to Do, Re, Mi, Fa, So, La, and Ti of the Western solfege. Each note is labeled with a symbol, a separate name, and is related to the specific *chakra* and *God*. Table 7 describes the expansion and meaning of each note along with the related animal, position of that note in the human body, and the God to which it is related [91]. A fixed frequency ratio with base note *shadja* is used to differentiate the swaras. Each swara has two to three variations except the *shadja* and *panchama*. Table 8 explains the variations of each swara, the scale, and the ratio to unison by assuming the tonic frequency as 220Hz and their nomenclature in Carnatic and Hindustani music. In general, the seven swaras, i.e., Sa, Ri, Ga, Ma, Pa, Dha, and Ni, are called as *pure* or *shuddha* swaras. The variations of these swaras are known as *teevra* or *vikruta swaras*. For instance, *rishabha* is flattened to obtain the *teevra rishabha* and Ma is sharpened to obtain *teevra madhyama*. The swaras in '*shudhdha* form are Sa and Pa only. The Western C is labeled as *shadja* in ICM. Although there are 16 swaras—including variations—in ICM, only 12 *swara sthanas* are

[8]"Hetero" means another and "phone" means sound; single variation for the single melody by at least two performers simultaneously.

[9]Swara and note, unison and tonic are used interchangeably in this article.

Table 6. Excerpts of the Articles Published on the Issue of Genre Classification

Sl. No.	Title of the Article	Composition of Database	Feature(s)	Classifier(s)	No. of GC's	Accuracy %	Remarks	Future Scope	Limitations
1	Musical genre classification of audio signals [234].	100 clips	Timbre, rhythm, and pitch	GMM and KNN	10	60.00	Rhythmic-related features are extracted to detect the genre class.	Exploration of pitch features probably leads to better results.	Thirty seconds of clips are considered for test which may take longer time.
2	A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques [72].	200 clips for training and testing	Time frequency and DWPT	KNN and EM	5	83.50	Wavelet transformation is applied to extract both time and frequency information.	Ensemble classifiers may improve the classification accuracy.	There is an issue with over fitting in the case of large databases.
3	Features for audio and music classification [155].	180 clips	AFTE	GMM	6	70.00	Tempered envelopes are used to analyze the signal characteristics.	Auditory features may give better performance if they are critically analyzed.	High feature dimensions are used that may lead to complexity issues.
4	Evaluation of feature extractors and psycho-acoustic models for music genre classification [134].	GTZAN and ISMIR database	AP, SSD, and RH	SVM	8	77.00	Rhythm histograms are analyzed for different genres to detect the (dis)similarity.	Detailed study is to be done in psycho-acoustic transformation w.r.t. genre.	Performance is very low and average accuracy of 72% is achieved .
5	Aggregate features and AdaBoost for music classification [17].	1500 clips	RCEPS, MFCC, SC, Rolloff, and LPC	AB.Tree, AB.stamp, ANN, and SVM	10	82.34	Variations of AdaBoost classifier are considered to check the performance.	A music clip may contain multi-genre information. Room is yet open in this aspect.	Feature dimension is high it may not reliable for real time systems.
6	Automatic classification of musical genres using inter-genre similarity [6].	1000 clips (30 sec each)	MFCC, ΔMFCC, and ΔΔMFCC	GMM, IGM, and IIGM	10	92.50	Inter genre similarity (IGM) classifier is introduced to identify genre similarities in an iterative manner.	Use of mid-term features may reduce the complexity issues to detect genre.	Longer clips of length 30 seconds are used to recognize genre information which some times may not be required.

(Continued)

Table 6. Continued

Sl. No.	Title of the Article	Composition of Database	Feature(s)	Classifier(s)	No. of GC's	Accuracy %	Remarks	Future Scope	Limitations
7	Temporal feature integration for music genre classification [158].	1210 clips	DAR and KNN	GLM	11	50.00	Classification done by professionals is compared with the performance of developed system.	Analysis with the support of experts may give an idea to design a proper system.	Humans can classify genres and they succeed up to 50-60%.
8	Content-based information fusion for semi-supervised genre classification [213].	1458 Clips	MFCC-earth mover's distance (EMD), FP and SH	K-Means and EM	6	77.00	Empirical mode decomposition (EMD) is applied to get the similarity score among two MFCCs.	Analysis using probabilistic methods will give idea while choosing the features.	Very little improvement is observed with this approach.
9	Music genre classification based on local feature selection using a self-adaptive harmony search algorithm [89].	GTZAN Database	Intensity, Pitch, Timbre, Tonality and rhythm	SVM	10	97.20	Self-adaptive harmony search algorithm is applied to select acoustic features. Up to 55% dimensions are reduced.	Experiments are needed to be done to select the relevant and useful features using evolutionary techniques.	GTZAN database is standardized one. However, it became old to accommodate new music trends.

Note: Only some relevant and widely cited articles are given.

Table 7. Swaras (Notes) in ICM and Their Associations

Swara	Expansion	Meaning	Animal	Chakra	God	Western Equivalent
Sa	Shadja	Ocean	Peacock	Base of spine	Agni	Do
Ri	Rishabha	Unbeaten	Skylark	Genitals	Brahma	Re
Ga	Gandhara	Sky	Goat	Solar plexus	Shiva	Mi
Ma	Madhyama	Middle	Dove	Heart	Vishnu	Fa
Pa	Panchama	Fifth	Cuckoo	Throat	Naarada	So
Dha	Dhaivata	Earth	Horse	Third eye	Ganapathi	La
Ni	Nishada	Hunter	Elephant	Crown of the head	Sun	Ti

Table 8. Swaras, Their Scales and Ratios

Symbol	Solfa	Scale	Ratio	Natural	Carnatic/Hindustani Word
Sa	Do	C	1:1	220.0	Shadja
Ri1		C#	16:15	234.7	Shuddha/Komal Rishabha
Ri2	Re	D	9:8	247.5	Chatushruti/Teevra Rishabha
Ga1	Re	D	9:8	247.5	Shuddha Gandhara
Ri3		Eb	6:5	264.0	Shatshruti Rishabha
Ga2		Eb	6:5	264.0	Sadharana/Komal Gandhara
Ga3	Mi	E	5:4	275.0	Antara/Teevra Gandhara
Ma1	Fa	F	4:3	293.3	Shuddha/Komal Madhyama
Ma2		F#	45:32	309.8	Prati/Teevra Madhyama
Pa	So	G	3:2	330.0	Panchama
Dha1		G#	8:5	352.0	Shuddha/Komal Dhaivata
Dha2	La	A	5:3	366.7	Chatushruti/Teevra Dhaivata
Ni1		A	5:3	366.7	Shuddha Nishada
Da3		Bb	9:5	396.0	Shatshruti Dhaivata
Ni2		Bb	9:5	396.0	Kaisiki/Komal Nishada
Ni3	Ti	B	15:8	412.5	Kakali Nishada
Sa’	Do’	C’	2	440.0	Shadja’

Note: *Sa* is starting note and *Sa’* is ending note.

considered due to the common ratio shared by some swaras. For instance, Ri2 & Ga1, Ri3 & Ga2, Dha2 & Ni1, and Dha3 & Ni2 share the same ratios.

Obviously, raga is neither a single scale nor a single tune. The different scales used in raga rendering, various kinds of *ornaments* and *pakads* are useful in identifying a raga [185]. Similar to the pieces of a chess game, the notes of a raga can be used in different ways. The significance of certain notes is very high compared to the rest and can be used to express the mood of a raga. Such notes are called *jeeva swaras*. *Graha swaras* are the notes that appear at the starting point of the melodic phrase. The notes that occur at the closing point of the melodic phrase are termed as *nyasa swaras*. The extended one is called *deerga swara*, frequently occurring ones are *amsa swaras*, and the less used swaras are *alpa swaras* [41]. The swap in locations of swaras does not make any difference to the raga. Involving *Sa* results in another raga. In ICM, the hierarchic structure of note, labeled as *That*, is created by Vishnu Narayan Bharatkhande for Hindustani music [19, 20] and Raamamaatya for Carnatic music Melakarta system. The combination of 12 swarastanas (swara positions) is considered to build 72 various ragas in the Melakarta system. Such 72 ragas are

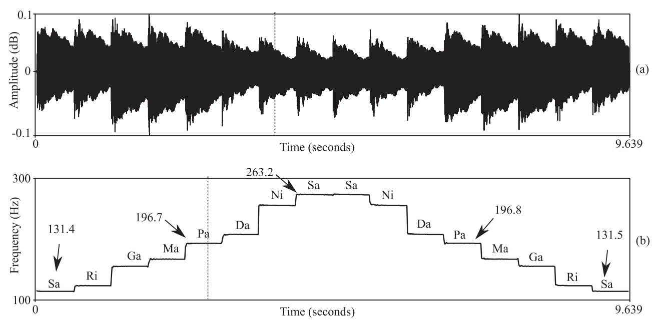

Fig. 3. A music signal of around 10 seconds showing the arohan-avarohan pattern of a *mayamalavagowlai raga*: (a) music signal recorded from a keyboard and (b) pitch contour of a music signal.

known as *janaka* or *parent ragas* [215]. The ragas evolved from parent ragas are known as *janya* or *child ragas*. However, a limited version of ragas is presented in *That* system, and each contains seven notes, in parent ragas. The others are derived from them.

7.2 Arohan, Avarohan, and Pakad

Arohan (ascending) and *avarohan* (descending) progressions are rendered using the notes of a raga. The usage of notes in framing the melodic phrase is determined by the order of notes in the progressions. The melodic phrase, which is ascending in nature, uses the notes that have ascending progression of the pitch. For instance, *bhairavi raga* has its arohan → *Sa Ga2 Ri2 Ga2 Ma1 Pa Dha2 Ni2 Sa'* and avarohan → *Sa' Ni2 Dha1 Pa Ma1 Ga2 Ri2 Sa*. Note that both arohan and avarohan patterns are different [118]. Simple representation of an arohan and avarohan pattern for raga *mayamalawagowlai* is shown in Figure 3. The keyboard is used to play this raga. Some properties of raga get repeated to establish signature of raga. One of them is a sequence of notes known as *pakad*, which is often visited by an artist throughout the performance.

7.3 Gamakas

Unlike the note rendering in Western classical music, a fixed notation is not available for ICM. The rapid oscillatory movement about the note is one of the several forms of improvisations in music rendering, which are together called *gamakas*. The sliding movement from one note to another is one of the forms of *gamaka*. There are several ways to move between the notes. In general, 15 types of gamakas are popular and universally accepted [92]. Gamakas refer to an ornamentation (*alankaras*) that beautifies the note patterns and creates special feeling while listening to the raga [220]. Some constraints need to be remembered by the artist while performing a raga using gamakas. For example, in *mohana raga*, there is no gamaka for Ri1 as Ga1 is very near, and it is not possible to render a gamaka after Ri1 without touching Ga2. However, Sa and Pa do not have any gamakas, so they are called *achala swaras* or *immovable notes*.

7.4 Vadi, Samvadi, and Jati

The first and essential note of a raga is called *vadi*. The next prominent one is called *samvadi*. The artist emphasizes the vadi and samvadi notes in a performance and renders them for significant durations. Others are called *anuvadi*, and those that are completely absent are *vivadi*. The number of notes in a raga plays a key role in classifying them. *Sampoorna raga* contains all seven notes, *shadhav* contains six notes (swaras), audhav contains five swaras and *surtar* contains only four swaras. A music rendering *shadhav-shadhav* contains six swaras in arohan and six in avarohan progression.

Several approaches have been explored in the literature to recognize the ragas from a given music clip [10, 33, 111, 121, 201, 214, 218]. Fundamental frequency plays an important role in identifying the notes and then ragas, as all the notes are tuned to specific ratios of the tonic frequency. A majority of the raga identification works have used F0 as a baseline feature. Along with pitch, some other related features such as pitch class distributions and pitch class dyad distributions are also computed to characterize the raga [10, 34, 102]. First-order pitch distributions and template-matching techniques are also popularly used for the identification of patterns of the raga [112]. Each singer is comfortable with the specific fundamental frequency (pitch) while performing in a concert. Thus, a singer can be characterized by the unique fundamental frequency. Moreover, it helps to recognize the raga as the identified pitch can be considered as tonic [214]. Probability density function (PDF) of pitch contours is also helpful to recognize the individual notes accurately. Hence, the same technique has been considered to improve the raga classification performance [218]. However, very few works on identifying raga and tala of a song have been reported in the literature. Table 9 summarizes some important raga-identification works reported in the literature. Room is still open to design reliable systems for raga identification in real concerts, raga transcription, tala recognition, and so on.

Raga identification is an important task of MIR, having several useful applications. The performance of a singer can be objectively judged during a live stage performance. An automatic music tutor machine can be designed for beginners. With this motivation, different approaches have been proposed in the literature to identify and transcribe a raga from the given music clip. The process of estimating the raga in monophonic clips could be simpler. The same task is highly challenging in the case of polyphonic clips due to background accompaniment. Besides, the pattern differences of various ragas and the involvement of *gamakas* are also important aspects that complicate the raga processing. The task of raga identification or note transcription can be efficiently done using the variations of a class of pitch features. As of now, a few works are reported on identifying the raga from the clips of live concerts. The results are not encouraging [112]. One more important observation is that the process of raga identification can also be helpful in recognizing the singer since the tonic frequency of a singer seems to be highly subjective. Instrument identification is also reported by determining the tonic frequency [111]. Properly identified ragas can also be helpful in finding the similarity in music, QBH, and genre classification. In this regard, certain statistical analysis on pitch class profile may be helpful to obtain relevant features that further enhance the task of raga identification.

8 QUERY BY HUMMING (QBH)

Image and audio are the most popular categories of multimedia information. At present, keyword/text-based search engines are most widely used and are available directly to the users. However, mapping the query is an important and a useful task when the search requirement is on multi-media data. This involves complexity in describing the information as a query statement. In general, it may be convincing for many of us to recollect the content of picture or music; however, it is not easy to form the query for the recollected information. Drawing the shape of an image and humming the tune of a music piece can be used to efficiently search for them. If an automated system is designed to search the database based on the provided image or humming a clip, then most of the human requirements will be met with respect to search engines. This issue can be resolved by using CBIR for both images and music clips [211, 239, 253]. To extract the image content, some tools are already available in the commercial market. The popular ones are

Table 9. Summary of Literature for the Task of Raga Identification

Sl. No.	Title of the Article	Composition of Database	Category	Feature(s)	Accuracy %	Remarks	Future Scope	Limitations
1	Raga identification by using swara intonation [10].	Four ragas	Hindustani	Pitch class distributions	87.00	Swara intonation is analyzed using pitch distributions and KL-dist measures.	Partition boundary detection and consonance pair examination may be useful in improving the accuracy.	Twelve partitions are assumed to have equal temperament and may not give accurate mean value.
2	Raga mining of Indian music by extracting arohana-avarohana pattern [207].	90 clips 50 ragas	Carnatic	Pitch values	95.00	Note transcription and raga detection are done by using distinct notes and their combinations.	It can be the base system for developing a complete note transcription system.	The system may not give good recognition for polyphonic clips due to rhythm effect.
3	Carnatic music analysis: Shadja, swara identification, and raga verification in alapana using stochastic models [189].	*Sampurna* ragas	Carnatic	PDF	91.50	PDF is used to estimate the variance of /sa/ and /pa/, which is helpful to recognize *shadja.*	This approach has to be tested on all kinds of *janaka* and *janya ragas.*	The experiments are limited to *sampoorna ragas* in which all notes were present.
4	A knowledge based signal processing approach to tonic identification in Indian classical music [12].	722 Clips	Carnatic	Group delay histograms	95.28	Tonic frequency has been estimated using template matching and segmented histograms.	A system is to be implemented to discriminate /sa/ and /pa/ where they affect performance.	The system is good enough to detect tonic of Carnatic and may not be effective for Hindustan music.
5	Raga recognition based on pitch distribution methods [112].	1415 recordings	Carnatic and Hindustani	First-order PDs and template matching	92.00	Raga recognition is done using template matching and first order pitch distributions.	Analysis on melodic phrases helps to detect the raga even when there is an accompaniment of ornamentation.	Inclusion of *gamakas* may not give proper results.
6	A multi-pitch approach to tonic identification in Indian classical music [194].	364 Excerpts	Hindustani and Carnatic	Pitch	93.00	Sinusoidal extraction and salience functions are used to detect tonic frequency.	Estimating the tonic frequency for both male and female is still unsolved.	The proposed system may not detect tonic in case if different gender information is provided.

(Continued)

Table 9. Continued

Sl. No.	Title of the Article	Composition of Database	Category	Feature(s)	Accuracy %	Remarks	Future Scope	Limitations
7	Identifying ragas in Indian Music [121].	*CompMusic* Dataset	Carnatic	SVM	83.39	Melody extraction algorithm is applied to detect *gamaka* and raga.	Temporal information may be helpful in recognizing raga even there is a *gamaka* effect.	Not much improvement is found with the melody extraction algorithm.
8	Raga Classification for Carnatic Music [218].	162 clips and 12 ragas'	Carnatic	Pitch and PDF	89.50	36-dimensional feature set is extracted from the PDF based on pitch contour.	An attempt to classify ragas' in hetero-phonic music clips may be used to develop a reliable system.	Database with 12 ragas is considered for experimentation. Only monophonic and polyphonic clips are considered.
9	Identifying Gamakas in Carnatic Music [242].	158 clips	Carnatic	Pitch and ASD	75.85	*Gamakas* are identified using the sequence of Attack, Sustain and Decay.	ASD modeling with the support of other techniques may help to identify the *gamaka*.	To detect *gamaka* in a music clip, the analysis on ASD sequence alone may not suffice.

Note: Only some relevant and widely cited articles are listed.

CBIC and Image Compass [60, 122]. It is not easy to develop a search engine for hummed clips as compared to drawn images. This is because a rough sketch can be easily drawn even without proficiency. However, QBH should be in a position to process and accept a wrong humming as an input due to lack of awareness in pitch modulation, voice parameters, and so on [42, 67]. This section addresses some important issues and existing research in the area of QBE/QBH particularly, passing the hummed tunes as a query to search the database.

Most of the time, one can hum the tune of one's interested/favorite song, however, it is difficult to recollect the name of the album, lyrics of the song, artist, composer, language, and so on. This can be resolved using QBH and the process is mainly categorized into three parts: (i) collecting the music clips from the digital world and creating the database after preprocessing them, (ii) formulating the query based on the tune hummed by the user, and (iii) selecting the relevant clip, which is more proximate to the hummed one. In practice, several complications will arise while implementing the system for QBH; some of them are discussed here.

One important issue related to QBH is the length of the hummed clip and non-matching humming positions. The humming length may vary from time to time and people to people. A singer may not hum the same portion of the song. Hence, the starting point and length may always have to be dynamically changed to match the hummed tune with the existing audio database. Sliding window based approach may be one of the reasonable solutions to this problem [57]. Another complication is wrong humming due to a lack of proficiency. All listeners may not hum in exactly the same way as the audio data. The pace of the tune hummed may be faster or slower compared to the original song. In this case, the duration may not match. The issue of fast and slow humming tunes can be resolved using time-normalization methods [186]. Moreover, the matching pattern should be extracted in the shortest time with high efficiency using suitably sophisticated algorithms. A robust QBH system is expected to handle all these issues, including poor humming, wrong pitch rendering, wrong note duration, wrong keys, and the like. The system should also be in a position to handle the noise and distortions.

The techniques introduced in the literature to identify a song based on a hummed tune are computationally expensive. Note-based QBH system may give a better solution since the comparison was based on notes instead of signals [251]. Two algorithms, namely note-based linear scaling (NLS) and note-based recursive align (NRA), are reported in this category. These two are the enhancements of linear-scaling (LS) and recursive algorithm (RA), which are used to recognize the melody in QBH systems [95, 250]. The issue in both LS and RA techniques is the expensive time needed to recognize the song, as they are frame-based approaches. The use of pitch contour is also one of the known techniques in implementing a QBH system. Pitch tracking based on notes helps for the design of QBH [67]. In this approach, normally, a given note is categorized into similar, greater, and smaller. Based on this, the sequence of hummed tunes can be compared with the existing song to identify the required one.

DTW is one of the most popular dynamic programming (DP) techniques for comparing two non-time aligned sequences and measuring the distance. It is popularly used in the applications of query by singing/humming (QBSH) [39, 42, 44, 94, 96, 104, 116, 126, 152, 164, 257]. Comparing pitch contour of two sequences is sufficient, rather than comparing the entire signal [93]. In some works, the normalized pitch values are considered as features to avoid the variance [1]. Tempo variations are also useful in identifying the matching patterns effectively [42, 152, 164]. The main issue with DTW is the drastically increasing computational complexity when the comparison is extended to the whole song. However, better accuracy can be achieved with DTW when compared to other methods. Since time is an important factor, especially for real-time systems, a modernized version of the sub-sequence matching method is used in practice for QBH systems [78]. This can be implemented by segmenting the given sequence into parts, and then DP-based DTW is applied for matching. A sliding window technique is used to check the matched pattern in all segments. In addition to DTW, earth movers' distance (EMD) and proportional transportation distance (PTD) are also available to measure the distance of two given one-dimensional signals [193]. EMD and PTD have better applications in the case of image comparison; however, they are also used for music signal comparison [232].

In fact, DTW itself has several variations to reduce the computational time, namely continuous DTW (cDTW), edit distance on real sequence (EDR) [29], edit distance with real penalty (ERP) [30], and so on. The difference between cDTW and EDR/ERP techniques is that the cDTW ignores the unmatched sequence by indexing the matched patterns, whereas EDR and ERP are used to measure the similarity based on the distance. They are developed to extract the perfectly matched sequences, and may not be useful for the QBH systems. This is due to the user always inputting the query with poor humming and wrong pitch. Of these, ERP can be used since it can handle noise environments as well. Recently, these sub-sequence matching techniques are being widely used for implementing QBH systems. Some research findings have reported that the longest common sub-sequence method gives improvement along with DTW when compared to plain DTW [40, 78]. Another intelligent approach for QBH systems would be to keep the popular music patterns in the database by collecting them manually along with the given meta information. It reduces the burden of comparing the music clip with the entire database. The given query is compared with the selected music pieces if their metadata matches. However, accuracy of the system is highly dependent on the reliability of available music clips [39, 86, 94, 115, 125, 126, 179, 238, 257].

Singing a song is completely different from humming it. Singing or humming discrimination method is useful in distinguishing the singing from humming. On this basis, an effort has been made to categorize the music clips based on singing or humming [243]. Locality-sensitive hashing (LSH) is initially proposed to distinguish singing from humming [73]. Based on this, note-based locality sensitive hashing (NLSH) and pitch-based locality sensitive hashing (PLSH) are proposed to reduce the complexity issues of LSH [244]. LS and RAs are also widely used techniques for QBH systems. The advancements over LS and RA algorithms such as boundary alignment linear scaling (BALS) and key transposition recursive alignment (KTRA), which gives better accuracy when comparing the input hummed tune with the songs in the database are also used [73]. However, the performance of these systems is not appreciably high in terms of time and computation. Literature shows that the KTRA gives better results compared to the other matching algorithms [101]. The literature findings on QBH system are summarized in Table 10, which contains different existing approaches and their limitations.

The evaluation process of the QBH systems is based on *top-k* ranking system [90]. The process of evaluating the performance using top-k ranking for $count_q$ number queries is given in Equation (1):

$$Top - KAccuracy = (Count_k / Count_q) * 100, \tag{1}$$

where $Count_k$ is the count of identified similar clips in the list of k and $Count_q$ is the total number of queries.

For instance, if 47 clips are correctly indexed at position one for a given 100 queries, then the top-1 ranking accuracy is equal to 47%. Figure 4 shows that the combination of LS, BALR, and KTRA gives better performance when compared separately. Two different datasets have been considered and the average performance of both the datasets have been shown. One is Think IT Corpus, which has 106 MIDI files and 355 queries. The other one is Jang's collection,[10] with 48 MIDI files and 4,431 queries. The performance of DTW is better, but it is affected by high time

[10] http://www.music-ir.org/mirex/wiki/2010:Query-by-Singing/Humming_Results.

Table 10. Excerpts of the Articles Published on the Issue of Query by Singing/Humming (QBSH)

Sl. No.	Title of the Article	Composition of Database	Feature(s)	Distance Measures	Accuracy %	Remarks	Future Scope	Limitations
1	Melody matching directly from audio [152].	90 MIDI clips	Pitch	Time warping	80.00	Enhanced auto correlation and time warping algorithms are used to extract the matching clips.	There is a scope to improve the system using temporal variations of prosodic and spectral features.	Pitch alone may not be sufficient to extract the matching clips.
2	Warping indexes with envelope transforms for query by humming [257].	100 MIDI clips	Envelops	DTW and LDTW	80.00	Envelops of signal have been compared, using DTW to find the similarity.	Normalized and smoothed envelops may be helpful to achieve better accuracy.	The envelops may vary for hummed clips and original clips because of background accompaniment.
3	An improved query by singing/humming system using melody and lyrics information [243].	2154 Mandarin songs	Pitch and Lyrics	Similarity matrix	72.00	Lyrics are extracted along with pitch to improve the accuracy.	Design of acoustic models may be helpful to get the proper phonetic information.	As background music is always accompanied with songs, simple similarity matrix may not be sufficient.
4	A fast query by humming system based on notes [251].	1000 MIDI clips	Notes	NLS and NRA	96.10	Combination of note-based linear scaling (NLS) and note-based recursive align (NRA) are used to measure the distance between the query and database clips.	Experimentation on voiced humming may be useful for real-time applications instead of MIDI clips.	Absolute pitch values are considered and only MIDI files are taken for experimentation.
5	A sub-sequence matching with gaps-range-tolerances framework: a query-by-humming application [115].	5643 MIDI clips	Subsequence	SMBGT and HMM	45.00	SMBGT is proposed to find matching clips. There are performance variations with variable tolerances.	Design of systems based on gaps and tolerances may improve the QBH system.	The sub-sequence information alone may not be sufficient for this task.
6	A query-by-humming system based on locality sensitive hashing indexes [73].	9940 MIDI & song clips	Pitch boundary & KTA	NLSH, PLSH, BALS, and KTRA	89.00	Several matching algorithms such as NLSH, PLSH, BALS, and KTRA are introduced in addition to LS.	There is a scope to improve the system for handling wrong input.	MIDI clips are considered for testing and may not be suitable for real time.
7	A two-stage query-by-singing/humming system on GPU [101].	8431 MIDI clips	Key transposition algo. (KTA)	LS and DTW	78.00	LS and DTW are combined and labeled as LS-DTW. This is done to utilize the features of both approaches.	Developing a system with combinatorial techniques is always better to overcome the disadvantages of individual systems.	As there is a standard metric for QBH evaluation, it is always better to use that approach.

Note: Only some relevant and widely cited articles are listed.

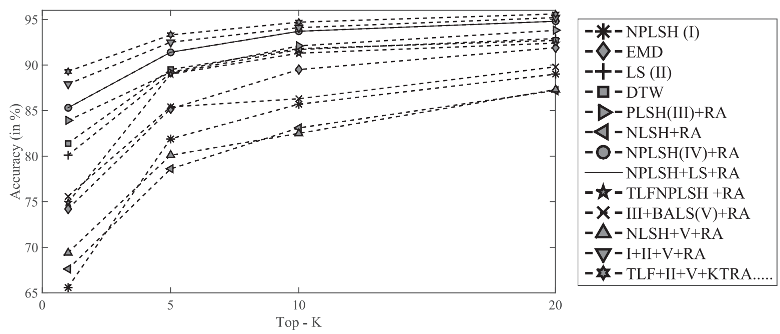

Fig. 4. Performance of various sequence-matching techniques.

complexity. However, a perfect matching algorithm is yet to be designed, and the relevant feature vectors need to be identified for better performance in less time.

9 MUSIC EMOTION RECOGNITION

Emotion classification based on music patterns is another important aspect of CB-MIR, which helps to recommend or fill the playlist based on users' emotional needs. The aim is to categorize the songs based on emotional patterns such as happy, angry, sad, and so on. Emotions are difficult to process because of inherent complications. It is impractical to compare the performance of the systems due to lack of a benchmark dataset. Recently, MIREX has created a standard dataset that is used to check the reliability and effectiveness of the works received for their competition. However, the dataset is not generalized to cover all important genre categories. The effort to create a benchmark dataset in the context of Indian popular music is almost nil. Hence, there is a need to create a standard dataset and develop an approach that can classify music based on emotions. Emotion classification is highly ambiguous due to many psychological aspects related to the emotions of a song. In the literature, some approaches such as Thayer's model [222], Hevner's model [82], and TWC model [221] have tried to address the issue of emotions. All these models are designed by expert psychologists and used by various MIR scientists; however, these models lack the support of listeners. In these cases, a listener's opinion is collected through majority voting, which creates an open ballot to collect the options from a variety of users, and then a majority can be labeled as the emotion for the song [48, 87].

The features used in literature for the task of emotion recognition are almost similar to those used for genre classification. From the analysis of the literature, it is clear that the low-level spectral features, such as MFCCs, LPCCs, Δ features, and so on, are helpful in categorizing the emotion of an audio clip [114, 129, 131, 142]. Some experiments have also been conducted using rhythmic features for categorizing the emotions in music [59]. These features are combined with other low-level features to improve the performance [36]. It is perceptual observation that the smoothness in placid emotions such as happy or sad is high compared to strong emotions like anger [162]. To estimate the smoothness among the changing multiple sounds, articulation-based features are more useful [36]. Some experiments are conducted on tempo-based features, and the tempo of angry clips is faster than that of placid emotions. This analysis supports the necessity in the use of articulation and rhythmic features for mood estimation.

The process of low-level feature extraction demands more time, as these features are extracted at every 20~40ms. To resolve this problem, octave-based spectral contrast (OSC) is explored [98]. These features are extracted at every spectral sub-band instead of fixed small-length segments known as frames and used for music mood classification [98, 142, 252]. The results of this approach convince us that the OSCs are better than traditional MFCCs for music mood estimation. Moreover, the emotion classification in music is a multi-label learning (MLL) problem because a song may contain more than one emotion in it [132]. To address this issue, sophisticated algorithms are introduced with the support of a KNN classifier [225]. The literature also contains the efforts to identify the mood of a song based on the instrumental region. In fact, the emotion of a song clip can be recognized by focusing on vocal as well as non-vocal regions. Table 11 gives some overview of existing literature with possible future directions. Based on the literature, the development of a reliable emotion-recognition system from songs based on the analysis of both vocal and non-vocal regions may give better performance.

10 INSTRUMENT IDENTIFICATION

Instrumental identification is very crucial for indexing music clips, genre classification, emotion recognition, and so on. The complexity involved in finding the timbre of an instrument is very high. Hence, the classification of instruments of an audio clip has created an impact on researchers to delve deep into it. Generally, music can be either monophonic (single instrument) or polyphonic (multiple instruments). The process of instrument identification is a sequence labeling problem in which each and every segment is to be labeled with the appropriate instrument(s). Perceiving the timbre information of instruments in music clips can solve various issues in medical and social fields [16, 37, 103, 153, 183]. In music healing therapy, music is used as an important mean to address many psychological issues related to anxiety and stress. The instruments with soothing timbre and genre are affective in healing different levels of ailments. Though it is highly subjective, knowing timbre influences treatment on positive side [146]. It is also reported from the ancient *Sangeeta Shastra* (logic of music) of Indian scriptures that the music can also interact with nature [167]. For instance, *Thansen*, the court musician of the musical king *Akbar*, used to bring down the rain automatically by playing a raga called *MeghMalhar* and he used to lighten the lamps by rendering the raga called *Deepak*. This is one of the crazy motivations for processing instruments as sounds.

Earlier research and findings have extracted the information about instruments from monophonic clips [2, 24, 49, 53–56, 65, 100, 117, 139, 140, 149, 180]. However, the applications of processing monophonic clips for instrumentals are very limited, except for few occasions. Polyphonic clips are generally available in the digital world with the combination of multiple instruments. Processing polyphonic clips being interested in instrumentals may be observed [56, 63, 75, 80, 109, 127, 137]. However, it is an open problem to tag all instruments present in a given clip of polyphonic music. To achieve this, the hierarchical clustering method is used to recognize multiple instruments from the clips of *jazz* category [56]. The combination of timbre and temporal features are used to cluster the instruments of each segment [137]. A particular segment is assigned to one of the existing clusters, based on their minimum distance. Each cluster represents specific instrument. This task also helps to know the presence/absence of a particular instrument for a given clip. Another challenging issue in polyphonic music processing is that instruments overlap, which needs multiple tagging of a single segment. From the music recommendation point of view, processing of instruments from music clips is a crucial point that needs to be addressed with special care [80, 109].

Most of the literature has considered clips with non-overlapping instruments. It is highly impossible to use them for real-time applications. A very few works are found with raw clips that include overlapping information. However, the number of instruments considered for experimentation and size of the database matters for considering them in real-time. Table 12 summarizes the

Table 11. Excerpts of the Articles Published on the Issue of Emotion Recognition from Music

Sl. No.	Title of the Article	Composition of Database	Emotional Classes	Feature(s)	Accuracy %	Approach	Future Scope	Limitations
1	The 2007 MIREX audio mood classification task: Lessons learned [48].	600 clips	Five emotions	Temporal, tonal, and loudness	52.65	Human-based classification is done and later compared with the system performance.	Analysis of the clips for all categories of emotions with mean opinion score may give better reliability.	The performance of system is not up to the mark with specified features.
2	Music mood and theme classification—a hybrid approach [21].	Allmusic.com and Last.fm	Four moods and four themes	Audio features	62.50	The hybrid approach is proposed to group the songs based on the emotions in them.	The theme and mood hierarchy is not standardized. Generalized hierarchy may be helpful to classify songs.	The process of feature extraction and feature selection is not explained properly.
3	SMERS: Music emotion recognition using support vector regression [76].	165 clips	Eleven emotions	Scale, energy, rhythm and harmonics	94.55	Eleven emotions have been classified with the support of support vector regression (SVR).	Proper selection of perceptual features may be useful to detect the emotions in a better way.	SVR is showing better performance compared to other non-linear classifiers. However, task related features are to be selected.
4	Lyric-based song emotion detection with affective lexicon and fuzzy clustering method [88].	981 Chinese clips	Valence and arousal	NLP and fuzzy clustering	60.38	NLP is applied to recognize the words and distribution among valence, and arousal is done using fuzzy clustering.	Extracting lyrics may be helpful along with the support of signal-processing approaches for mood estimation.	Fuzzy clustering alone is considered, which may be the reason for less accuracy.
5	Music emotion classification and context-based music recommendation [77].	120 clips	Eleven emotions	Low-level features and COMUS ontology	61.80	COMUS is used—an ontology to estimate the users' present emotional state based on past behavior and low-level features that are applied for song mood estimation.	Multi-mood estimation may be possible with low-level features as a song contains more than one emotion.	The system gives less accuracy due to improper estimation of users' moods. Moreover, low-level features alone may not be sufficient.
6	An approach of genetic programming for music emotion classification [7].	488 Western clips	Five emotions	Timbre, tonality, and chord	74.4	Two-level classification is applied to identify class of emotion and actual emotion.	A light is to be thrown on evolutionary approaches to reduce the complexity issues and increase the performance.	The performance of system gets degrading when the number of classes is increasing.
7	Audio songs classification based on music patterns [202].	300 clips	Seven emotions	MFCC, stat {pitch} and vibrato	82.00	Modulated features are used to detect emotions and mean opinion score is collected.	Consideration of vocal and non-vocal regions may improve system accuracy.	Increase in database may reduce the accuracy because the features may not be sufficient to discriminate emotions.

Note: Only some relevant and widely cited articles are considered.

Table 12. Excerpts of the Articles Published on the Issue of Instrument Identification

Sl. No.	Title of the Article	Composition of Database	Approach	Instrument Count	Accuracy %	Remarks	Future Scope	Limitations
1	Musical instrument identification in continuous recordings [140].	108 Solo performances	Temporal, harmonic, and perceptual	Seven instruments	85.24	Gradual Descriptor Elimination (GDE) is applied to ignore the irrelevant features.	Evolutionary approaches for feature reduction may give better results.	The features selected using GDE do not give good accuracy when compared with complete feature set.
2	Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps [109].	RWC-MDB-C-2001	Harmonics and LDA	Polyphonic clips	84.10	Frequency components of harmonic structure are analyzed to detect the instruments information when there is an overlapping scenario.	Proper analysis of timbre of instruments in polyphonic environment is yet to be addressed in detail.	When multiple instruments are involved, performance gets degraded.
3	Semi-supervised learning for musical instrument recognition [45].	RWC-MDB	MFCCs	Nine instruments	77.00	Semi supervised learning (SSL) approach is proposed to automatically label the training data instead of manual labeling.	Consideration of multiple scenarios such as multiple instruments, noise, reverberation and so on for instrument identification may be helpful in real-time applications.	Experimentation is done using MFCCs alone and they may not be suitable for multiple instrument recognition as they cannot carry information about high-frequency components.
4	Musical instrument recognition in polyphonic audio using missing feature approach [68].	MUMS Database	LSF and MET	Ten instruments	64.10	Mask estimation technique is used to detect multiple instruments in a given clip.	Combination of temporal features and MET may improve the instrument recognition performance.	The performance may be degraded because of improper feature selection.
5	Automatic ontology generation for musical instruments based on audio analysis [113].	RWC-MDB	LSF and MFCCs	Fifteen Instruments	79.87	Ontology web logic (OWL) is applied to arrange the identified instruments hierarchically.	The system for detecting and classifying instruments may be useful in music indexing applications.	The use of basic MFCCs and LSF may not be able to capture the information of high-frequency components.
6	Identification of Indian musical instruments by feature analysis with different classifiers [99].	150 mono clips	Statistical MFCCs LPC	Three instruments	99.37	Statistical features along with cepstral coefficients are used to identify the class of instruments.	Analysis on temporal variation of instruments may be helpful to annotate them properly.	Monophonic clips are considered, which have less impact in real-time applications.

Note: Only some relevant and widely cited articles are considered.

instrument identification tasks and scope of improvement to recognize the possible instruments from an audio clip.[11, 12]

11 MUSIC ANNOTATION

The most difficult and important task in music processing applications is tagging each portion of the music clip with labels such as genre, artist, style, emotion, and so on. This task is also known as music annotation. Annotation helps in converting the given music clip into textual information containing useful metadata. If multimedia information can be represented in textual form, then the majority of the complexities of search engines are conquered. This motivated researchers to have a look into it [210]. Music annotation is a complicated work and needs the support of other MIR tasks. Initially, a dataset was created by Reference [230] for music annotation, named CAL500,[13] containing 500 songs from different albums. The textual labeling was done manually, and was made available through public portals to train and test the newly developed systems for the task of music annotation. Furthermore, this database (http://slam.iis.sinica.edu.tw/demo/CAL500exp/) is enhanced and used for more sophisticated music annotation [245]. One more issue of music annotation is identifying the performance metrics.

Two useful metrics are suggested to measure the correctness of the system [230]. One is average Area Under ROC curve (AUC) and average precision for ranking n-labels *(precision-at-n)* of a clip. In the initial works, CAL 500 dataset is used with MFCCs and Δ-MFCC features. The support of hierarchical GMMs for the ΔMFCC features is taken to tag the music clip [83, 231]. The standard MFCCs are enhanced by computing the *MuVar* for them. It is reported that better AUC values are obtained with the enhanced features compared to standard MFCCs [170]. This indicates that the variations in cepstral features are helpful in labeling the music clips. It opens up room for working on the features such as *MuCov* (http://marsyas.info), FP [176], rhythmic pattern [136], rhythmic coefficients [247], and other sophisticated features evolved from the modulation spectrum analysis for music annotation.

Since tagging is to be done for every segment of music clip, music annotation is an MLL problem. However, labels of each segment are useful in deciding the high-level tag of the entire clip, such as emotion and genre. It can be done by providing weight values to each tag. This process is known as semantic multi-nominal, which helps in identifying the pertinence of a tag to a music clip [147]. In some cases, content-based tag co-occurrences may not give appropriate results [83, 231]. In such cases, music annotation is useful for tagging the high-level attributes accurately. This kind of experimentation is done using a generative context model with the combination of existing auto-tagger methods [71]. From literature, it is observed that there are several approaches left behind for music annotation. One important aspect of them is language identification of a given song clip [88]. The task of language identification is highly helpful in providing the lyrics information. Moreover, the tasks of MIR can be narrowed down if the information about language is known. The literature on music annotation is summarized in Table 13.

12 SOME FUTURE RESEARCH DIRECTIONS

Based on the critical review of available research outcomes in the area of CB-MIR, the following issues are worth giving immediate research attention to improve the performance of existing CB-MIR systems:

[11] http://www.staff.aist.go.jp/m.goto/RWC-MDB.
[12] http://www.worldcat.org/title/mcgill-university-master-samples-collection-on-dvd/oclc/244566561.
[13] http://labrosa.ee.columbia.edu/millionsong/sites/default/files/cal500HDFS.tar.gz.

Table 13. Excerpts of the Articles Published on the Issue of Music Annotation

Sl. No.	Title of the Article	Database	Feature(s)	Classifier(s)	AUC	Remarks	Future Scope	Limitations
1	On the use of anti-word models for audio music annotation and retrieval [31].	CAL500	MFCCs	GMM	0.56	MFCCs are used to annotate the clips.	Use of tempo and rhythm features instead of simple MFCCs may improve the accuracy.	Better accuracy may not be achieved for music annotation with MFCCs alone.
2	Exploring automatic music annotation with acoustically-objective tags [223].	Swat10K	ENT and ENS	GMM, SVM, and BDS	0.78	Echo Nest Timbre and song features are introduced to annotate song clips.	Independent Component Analysis (ICA) may improve the performance of music annotation.	Auto-tagging can be effective as every frame is mentioned clearly, which may not be possible with specified features.
3	Time series models for semantic music annotation [38].	CAL500	Temporal and Rhythm	HEM-DTM, AGG-DTM, and EM-DTM	0.68	Using HEM-DTM, temporal information is extracted from song clips based on dynamic texture mixture (DTM).	Instead of training with huge data, searching for distributed approaches may reduce the complexity issues.	The subset of CAL500 database is considered for experimentation, which is not very large.
4	A generative context model for semantic music annotation and retrieval [163].	CAL500 CAL10K	MFCCs and ΔMFCCs	SVM, cBOOST, GMM and DMM	0.73	At initial stage, music annotation is done on instruments of a song clip using cepstral features. Later, context model is applied on results to provide high-level contextual information.	Instead of using cepstral features, the use of temporal and perceptual features is helpful in improving the tagging quality.	Simple base-line features are considered to tag the information, which may not be sufficient many times.
5	Learning sparse feature representations for music annotation and retrieval [169].	CAL500	MFCCs	SRBM	0.74	Five different tags are considered to label CAL500 dataset using MFCCs and sparse restricted boltzman machie (SRBM).	Design of multi-modal approach to detect each individual component of a music clip is essential.	The system introduces SRBM for music annotation. It is better to consider additional tags to utilize the system for other standard databases.

Note: Only some relevant and widely cited articles are considered.

— Lack of benchmark datasets in eastern countries, especially for the music of the Indian subcontinent, is a major concern for the music research community. The song categories of the eastern world contribute to a major portion of the digital audio domain. The provision of standard datasets to cover different aspects of MIR highly motivates the researchers to work on this area.
— The features that are computed for various speech and audio processing techniques have been directly used for a majority of MIR tasks without thorough analysis. Some standard correlation analysis may help in deciding the feature set. In this process, it is also possible to reduce the dimensionality, which further minimizes the complexity issues. The task of identifying task-related features for different tasks of MIR systems is still a major problem that needs immediate attention.
— At present, the task of vocal and non-vocal segmentation has been considered just as a subtask of a singer identification for which extracting a small portion would be sufficient. However, there are many other applications for a complete vocal and non-vocal segmentation task. The process of separating source information may simplify the task of locating vocal onset and offset points. This separation is also helpful in developing an efficient *karaoke* system without foreground voice. Hence, a special focus is essential on vocal and non-vocal segmentation.
— The majority of singer identification systems that are accessible at present consider audio clips with a solo singer and minimal monophonic musical accompaniment. This could be the major hindrance for considering them for real-time applications. Hence, there is an immense need to develop singer-identification systems that can handle the clips with multiple singers, overlapping singers, and variety of background instruments. Singer tracking in duets is also an important step toward a complete solution. The process of detecting gender of a singer could be one possible solution that simplifies the task of singer identification.
— The taxonomy of genres is not well-defined in the music industry. Many times it is found that the same track falls in the category of more than one genre. There is a huge number of genre classes that are highly unorganized with many overlaps. The taxonomy creation surely motivates the researchers to design a proper genre classification system.
— The task of raga identification can be improved only with the support of tonic identification. Though a significant amount of work has already been reported on tonic frequency estimation, the approaches reported are not adequately developed for live concerts. As tonic frequency is an essential component in estimating the singer of a song, it is essential to develop a complete tonic identification system. Moreover, all the 72 Melakarta ragas have not been considered in existing works. The task of raga identification and note transcription can help in designing an automated tutor for those who are interested in learning music. A system can also be developed to judge the singers in live performance through objective evaluation.
— The task of instrument identification is highly focused on monophonic clips, and the overlapping issues are less addressed. The approach of independent component analysis may help in estimating the instruments though they are recorded in a polyphonic environments. The timbre of an instrument is also highly helpful for the task of instrument identification. The histogram analysis for different features of various instruments could be another possible solution to distinguish them. Multi-pitch detection is helpful to recognize the multiple instruments in a selected clip. More fine-tuned systems are essential to detect the instruments in polyphonic environment.
— The area of speech emotion recognition has been highly addressed, and many approaches have been proposed. The difficulty in deciding the mood of a song clip is the main reason

for not having standard models and benchmark datasets for music mood estimation. An effort has to be made for a general dataset and standard model of mood estimation in the context of songs. Moreover, the existing works have concentrated only in the song portions, especially of instrumentals. It is also true that the vocals carry much information related to moods. This hints toward extensive efforts on mood estimation from a song based on vocals in it.

— QBH is another important task of MIR that has been implemented mostly on MIDI files. It may not be suitable for real-time scenarios. Recently, some systems have been proposed on query by singing that are also useful to extract or search the clips based on either lyrics or human voice. The accuracy top-1 rank is also not up to the appreciable mark for generalization of performance. From this background, it is said that there is a huge research scope in QBH and QBS.
— The process of annotating each portion of the song is the ultimate solution for MIR that gives complete information. The task is certainly dependent on the other tasks of MIR. At present, the works have focused on labeling the songs based on the lyrics, instruments, and solo singer. There are many other important tasks of MIR that are to be concentrated on to provide complete annotation such as gender, multiple singers, raga, and so on. Moreover, a single portion of a song can be labeled with more than one tag based on the information present. The light is yet to be thrown in this area.
— The present research trend is to use the technology of DNNs in many fields. The DNNs are self-capable of extracting features from the raw input. The only condition to use the DNN is that the data should be large enough. Moreover, it has to be properly labeled. A very few works have used the DNNs in the area of MIR. Since the music data is large in size with high stochastic nature, the use of DNNs may help in extracting features from the complex signal. Further, comparing the works with the existing feature-dependent systems may give a hint at upcoming research.

13 CONCLUSION

Recent works in CB-MIR, along with its important tasks and needs, are presented in this survey article. An up-to-date review is provided on the features and classifiers that are used in CB-MIR. Moreover, the sub-tasks such as vocal/non-vocal segmentation, artist identification, genre classification, raga identification, music emotion recognition, query by humming, instrument identification, and music annotation are discussed from the point of view of available literature. As these are sufficient to extract the meta-information from music clips, their importance for various music-related applications and general possibilities to enhance the existing works are highlighted. At the end of the survey, it is identified that there is a huge scope for further advancements and enhancements in CB-MIR, as most of the issues are yet to be addressed to achieve higher efficiency. The article lists some important gray areas where research initiatives in the area of CB-MIR are essential.

REFERENCES

[1] Norman H. Adams, Mark A. Bartsch, Jonah B. Shifrin, and Gregory H. Wakefield. 2004. Time series alignment for music information retrieval. In *Proceedings of the 5th Annual International Symposium on Music Information Retrieval (ISMIR'04)*. 056–303.

[2] Giulio Agostini, Maurizio Longari, and Emanuele Pollastri. 2003. Musical instrument timbres classification with spectral features. *EURASIP J. Appl. Signal Process.* 2003 (2003), 5–14.

[3] Nasir Ahmed, T. Natarajan, and Kamisetty R. Rao. 1974. Discrete cosine transform. *IEEE Trans. Comput.* 100, 1 (1974), 90–93.

[4] Eric Allamanche, Jürgen Herre, Oliver Hellmuth, Bernhard Fröba, Throsten Kastner, and Markus Cremer. 2001. Content-based identification of audio material using MPEG-7 low level description. In *Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR'01).*

[5] Jean-Julien Aucouturier and Francois Pachet. 2003. Representing musical genre: A state of the art. *J. New Music Res.* 32, 1 (2003), 83–93.

[6] Ulas Bagci and Engin Erzin. 2007. Automatic classification of musical genres using inter-genre similarity. *IEEE Signal Process. Lett.* 14, 8 (2007), 521–524.

[7] Sung-Woo Bang, Jaekwang Kim, and Jee-Hyong Lee. 2013. An approach of genetic programming for music emotion classification. *Int.J. Control, Autom. Syst.* 11, 6 (2013), 1290–1299.

[8] Luke Barrington, Mehrdad Yazdani, Douglas Turnbull, and Gert R. G. Lanckriet. 2008. Combining feature kernels for semantic music retrieval. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR'08).* 614–619.

[9] Claudio Becchetti and Klucio Prina Ricotti. 2008. *Speech Recognition: Theory and C++ Implementation.* John Wiley & Sons.

[10] Shreyas Belle, Rushikesh Joshi, and Preeti Rao. 2009. Raga identification by using swara intonation. *J. ITC Sangeet Res. Acad.* 23 (2009), 1–7.

[11] Juan Pablo Bello. 2007. Audio-based cover song retrieval using approximate chord sequences: Testing shifts, gaps, swaps and beats. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR'07)*, Vol. 7. 239–244.

[12] Ashwin Bellur, Vignesh Ishwar, Xavier Serra, and Hema A. Murthy. 2012. A knowledge based signal processing approach to tonic identification in Indian classical music. In *Proceedings of the 2nd CompMusic Workshop.* 113–118.

[13] Emmanouil Benetos, Margarita Kotti, and Constantine Kotropoulos. 2006. Musical instrument classification using non-negative matrix factorization algorithms and subset feature selection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06).* IEEE, Los Alamitos, CA, 221–224.

[14] Kirell Benzi, Michaël Defferrard, Pierre Vandergheynst, and Xavier Bresson. 2016. FMA: A dataset for music analysis. arXiv:1612.01840.

[15] Adam L. Berenzweig, Daniel P. W. Ellis, and Steve Lawrence. 2002. Using voice segments to improve artist classification of music. In *Proceedings of the 22nd International Conference on Virtual, Synthetic, and Entertainment Audio.*

[16] Kenneth W. Berger. 1964. Some factors in the recognition of timbre. *J. Acoust. Soc. Am.* 36, 10 (1964), 1888–1891.

[17] James Bergstra, Norman Casagrande, Dumitru Erhan, Douglas Eck, and Balázs Kégl. 2006. Aggregate features and AdaBoost for music classification. *J. Mach. Learn.* 65, 3 (2006), 473–484.

[18] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11).* 591–596.

[19] Vishnu Narayan Bhatkhande. 1934. *A Short Historical Survey of the Music of Upper India.* B.S. Sukthankar.

[20] Vishnu Narayan Bhatkhande. 1990. *Hindustani Sangeet Paddhati.* Vol. 1. Luminous Books, Varanasi.

[21] Kerstin Bischoff, Claudiu S. Firan, Raluca Paiu, Wolfgang Nejdl, Cyril Laurier, and Mohamed Sordo. 2009. Music mood and theme classification—a hybrid approach. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR'09).* 657–662.

[22] Eva Björkner. 2006. *Why So Different? Aspects of Voice Characteristics in Operation and Musical Theater Singing.* Ph.D. Dissertation. School of Computer Science and Communication, Speech, Music and Hearing, TMH.

[23] A. Bonjyotsna and M. Bhuyan. 2013. Signal processing for segmentation of vocal and non-vocal regions in songs: A review. In *Proceedings of the International Conference on Signal Processing Image Processing and Pattern Recognition (ICSIPR'13).* IEEE, Los Alamitos, CA, 87–91.

[24] Judith C. Brown. 1999. Computer identification of musical instruments using pattern recognition with cepstral coefficients as features. *J. Acoust. Soc. Am.* 105, 3 (1999), 1933–1941.

[25] Lori Mann Bruce, Cliff H. Koger, and Jiang Li. 2002. Dimensionality reduction of hyper-spectral data using discrete wavelet transform feature extraction. *IEEE Trans. Geosci. Remote Sens.* 40, 10 (2002), 2331–2338.

[26] Wei Cai, Qiang Li, and Xin Guan. 2011. Automatic singer identification based on auditory features. In *Proceedings of the 7th International Conference on Natural Computation (ICNC'11)*, Vol. 3. IEEE, Los Alamitos, CA, 1624–1628.

[27] Pedro Cano, Emilia Gómez, Fabien Gouyon, Perfecto Herrera, Markus Koppenberger, Beesuan Ong, Xavier Serra, Sebastian Streich, and Nicolas Wack. 2006. ISMIR 2004 audio description contest. Technical Report. Music Technology Group of the Universitat Pompeu Fabra.

[28] Michael A. Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. 2008. Content-based music information retrieval: Current directions and future challenges. *Proc. IEEE* 96, 4 (2008), 668–696.

[29] Lei Chen and Raymond Ng. 2004. On the marriage of LP-norms and edit distance. In *Proceedings of the 30th International Conference on Very Large Databases*, Vol. 30. 792–803.

[30] Lei Chen, M. Tamer Özsu, and Vincent Oria. 2005. Robust and fast similarity search for moving object trajectories. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD'05)*. ACM, New York, NY, 491–502.
[31] Zhi-Sheng Chen and Jyh-Shing Roger Jang. 2009. On the use of anti-word models for audio music annotation and retrieval. *IEEE Trans. Audio, Speech, Lang. Process.* 17, 8 (2009), 1547–1556.
[32] Heng-Tze Cheng, Yi-Hsuan Yang, Yu-Ching Lin, I-Bin Liao, and Homer H. Chen. 2008. Automatic chord recognition for music classification and retrieval. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, Los Alamitos, CA, 1505–1508.
[33] Parag Chordia. 2006. Automatic raag classification of pitch tracked performances using pitch-class and pitch-class dyad distributions. In *Proceedings of the International Computer Music Conference (ICMC'06)*. 431–436.
[34] Parag Chordia and Alex Rae. 2007. Raag recognition using pitch-class and pitch-class dyad distributions. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR'07)*. 431–436.
[35] Wu Chou and Liang Gu. 2001. Robust singing detection in speech/music discriminator design. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01)*, Vol. 2. IEEE, Los Alamitos, CA, 865–868.
[36] Bee Yong Chua. 2008. *Automatic Extraction of Perceptual Features and Categorization of Music Emotional Expressions from Polyphonic Music Audio Signals*. Ph.D. Dissertation. Monash University.
[37] Melville Clark Jr., Paul T. Robertson, and David Luce. 1964. A preliminary experiment on the perceptual basis for musical instrument families. *J. Audio Eng. Soc.* 12, 3 (1964), 199–203.
[38] Emanuele Coviello, Antoni B. Chan, and Gert Lanckriet. 2011. Time series models for semantic music annotation. *IEEE Trans. Audio, Speech, Lang. Process.* 19, 5 (2011), 1343–1359.
[39] Tim Crawford, Costas S. Ilipoulos, and Rajeev Raman. 1998. String-matching techniques for musical similarity and melodic recognition. *Comput. Music: Dir. Res.* 11 (1998), 73–100.
[40] Maxime Crochemore, Costas S. Iliopoulos, Christos Makris, Wojciech Rytter, Athanasios K. Tsakalidis, and T. Tsichlas. 2002. Approximate string matching with gaps. *Nordic J. Comput.* 9, 1 (2002), 54–65.
[41] David Cummings. 1997. *Random House Encyclopedic Dictionary of Classical Music*. Random House.
[42] Roger B. Dannenberg and Ning Hu. 2004. Understanding search performance in query-by-humming systems. In *Proceedings of the 5th Annual International Symposium on Music Information Retrieval (ISMIR'04)*. 43–48.
[43] Ingrid Daubechies et al. 1992. *Ten Lectures on Wavelets*. Vol. 61. Society for Industrial and Applied Mathematics (SIAM).
[44] John R. Deller, John G Proakis, and John H. L. Hansen. 2000. *Discrete-Time Processing of Speech Signals*. IEEE, Los Alamitos, CA.
[45] Aleksandr Diment, Toni Heittola, and Tuomas Virtanen. 2013. Semi-supervised learning for musical instrument recognition. In *Proceedings of the 21st European Signal Processing Conference (EUSIPCO'13)*. IEEE, Los Alamitos, CA, 1–5.
[46] Christian Dittmar, Christoph Bastuck, and Matthias Gruhne. 2007. Novel mid-level audio features for music similarity. In *Proceedings of the International Conference on Music Communication Science*. 38–41.
[47] Christian Dittmar, Bernhard Lehner, and Thomas Prätzlich. 2015. Cross-version singing voice detection in classical opera recordings. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR'15)*. 618–624.
[48] XHJS Downie, Cyril Laurier, and MBAF Ehmann. 2008. The 2007 MIREX audio mood classification task: Lessons learned. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR'08)*. 462–467.
[49] Shlomo Dubnov and Xavier Rodet. 1998. Timbre recognition with combined stationary and temporal features. In *Proceedings of the International Computer Music Conference (ICMC'98)*. 102–108.
[50] Daniel P. W. Ellis. 2007. Classifying music audio with timbral and chroma features. In *Proceedings of the 8th Annual International Symposium on Music Information Retrieval (ISMIR'07)*. 339–340.
[51] Daniel P. W. Ellis and Graham E. Poliner. 2007. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, Vol. 4. IEEE, Los Alamitos, CA.
[52] Robert Erickson. 1975. *Sound Structure in Music*. University of California Press.
[53] Antti Eronen. 2003. Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMs. In *Proceedings of the 7th International Symposium on Signal Processing and Its Applications*, Vol. 2. IEEE, Los Alamitos, CA, 133–136.
[54] Antti Eronen. 2001. *Automatic Musical Instrument Recognition*. Master's Thesis. Department of Information Technology, Tampere University of Technology.
[55] Slim Essid, Gaël Richard, and Bertrand David. 2004. *Efficient Musical Instrument Recognition on Solo Performance Music Using Basic Features*. Audio Engineering Society.

[56] Slim Essid, Gaël Richard, and Bertrand David. 2006. Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Trans. Audio, Speech, Lang. Process.* 14, 1 (2006), 68–80.
[57] Christos Faloutsos, Mudumbai Ranganathan, and Yannis Manolopoulos. 1994. Fast subsequence matching in time-series databases. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD'94)*, Vol. 23. ACM, New York, NY, 419–429.
[58] György Fazekas, Yves Raimond, Kurt Jacobson, and Mark Sandler. 2010. An overview of Semantic Web activities in the OMRAS2 project. *J. New Music Res.* 39, 4 (2010), 295–311.
[59] Yazhong Feng, Yueting Zhuang, and Yunhe Pan. 2003. Music information retrieval by detecting mood via computational media aesthetics. In *Proceedings of the IEEE International Conference on Web Intelligence (WIC'03).* IEEE, Los Alamitos, CA, 235–241.
[60] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, et al. 1995. Query by image and video content: The QBIC system. *IEEE Comp. Soc.* 28, 9 (1995), 23–32.
[61] Zhouyu Fu, Guojun Lu, Kai-Ming Ting, and Dengsheng Zhang. 2010. On feature combination for music classification. In *Proceedings of the Conference on Structural, Syntactic, and Statistical Pattern Recognition.* 453–462.
[62] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. 2011. A survey of audio-based music classification and annotation. *IEEE Trans. Multimed.* 13, 2 (2011), 303–319.
[63] Ferdinand Fuhrmann, Martín Haro, and Perfecto Herrera. 2009. Scalability, generality and temporal aspects in automatic recognition of predominant musical instruments in polyphonic music. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR'09).* 321–326.
[64] Hiromasa Fujihara, Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. 2005. Singer identification based on accompaniment sound reduction and reliable frame selection. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR'05).* 329–336.
[65] Ichiro Fujinaga and Karl MacMillan. 2000. Real-time recognition of orchestral instruments. In *Proceedings of the International Computer Music Conference.* 141–143.
[66] Jort F. Gemmeke, Daniel P.W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of the 42nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'17).*
[67] Asif Ghias, Jonathan Logan, David Chamberlin, and Brian C. Smith. 1995. Query by humming: Musical information retrieval in an audio database. In *Proceedings of the 3rd ACM International Conference on Multimedia.* ACM, New York, NY, 231–236.
[68] Dimitrios Giannoulis and Anssi Klapuri. 2013. Musical instrument recognition in polyphonic audio using missing feature approach. *IEEE Trans. Audio, Speech, Lang. Process.* 21, 9 (2013), 1805–1817.
[69] Emilia Gómez and Perfecto Herrera. 2004. Estimating the tonality of polyphonic audio files: Cognitive versus machine learning modelling strategies. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04).*
[70] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. 2003. RWC music database: Music genre database and musical instrument sound database. In *Proceedings of the 4th International Conference on MusicInformation and Retrieval (ISMIR'03).*
[71] Masataka Goto and Keiji Hirata. 2004. Recent studies on music information processing. *Acoust. Sci. Technol.* 25, 6 (2004), 419–425.
[72] Marco Grimaldi, Pádraig Cunningham, and Anil Kokaram. 2003. A wavelet packet representation of audio signals for music genre classification using different ensemble and feature selection techniques. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval.* ACM, New York, NY, 102–108.
[73] Zhiyuan Guo, Qiang Wang, Gang Liu, and Jun Guo. 2013. A query by humming system based on locality sensitive hashing indexes. *J. Signal Process.* 93, 8 (2013), 2229–2243.
[74] Philippe Hamel, Yoshua Bengio, and Douglas Eck. 2012. Building musically-relevant audio features through multiple timescale representations. In *Proceedings of the 13th Annual International Symposium on Music Information Retrieval (ISMIR'12).* 553–558.
[75] Philippe Hamel, Sean Wood, and Douglas Eck. 2009. Automatic identification of instrument classes in polyphonic and poly-instrument audio. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR'09).* 399–404.
[76] Byeong-Jun Han, Seungmin Ho, Roger B. Dannenberg, and Eenjun Hwang. 2009. SMERS: Music emotion recognition using support vector regression. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR'09).* 651–656.
[77] Byeong-Jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. 2010. Music emotion classification and context-based music recommendation. *Multimed. Tools Appl.* 47, 3 (2010), 433–460.

[78] Tae Sik Han, Seung-Kyu Ko, and Jaewoo Kang. 2007. Efficient subsequence matching using the longest common subsequence with a dual match index. In *Proceedings of the 5th International Conference on Machine Learning and Data Mining (MLDM'07)*. 585–600.

[79] Aki Harma and Unto K. Laine. 2001. A comparison of warped and conventional linear predictive coding. *IEEE Trans. Speech Audio Process.* 9, 5 (2001), 579–588.

[80] Toni Heittola, Anssi Klapuri, and Tuomas Virtanen. 2009. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR'09)*. 327–332.

[81] Marko Helen and Tuomas Virtanen. 2005. Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine. In *Proceedings of the 13th European Signal Processing Conference*. IEEE, Los Alamitos, CA, 1–4.

[82] Kate Hevner. 1936. Experimental studies of the elements of expression in music. *Am. J. Psychol.* 48, 2, 246–268.

[83] Matthew D. Hoffman, David M. Blei, and Perry R. Cook. 2009. Easy as CBA: A simple probabilistic model for tagging music. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR'09)*, Vol. 9. 369–374.

[84] Helge Homburg, Ingo Mierswa, Bülent Möller, Katharina Morik, and Michael Wurst. 2005. A benchmark dataset for audio classification and clustering. In *Proceedings of the 6th Annual International Symposium on Music Information Retrieval (ISMIR'05)*. 528–531.

[85] Chao-Ling Hsu and Jyh-Shing Roger Jang. 2010. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Trans. Audio, Speech, Lang. Process.* 18, 2 (2010), 310–319.

[86] Ning Hu, Roger B. Dannenberg, and Ann L. Lewis. 2002. A probabilistic model of melodic similarity. In *Proceedings of the International Computer Music Conference*.

[87] Xiao Hu and J. Stephen Downie. 2007. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR'07)*. 67–72.

[88] Yajie Hu, Xiaoou Chen, and Deshun Yang. 2009. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR'09)*. 123–128.

[89] Yin-Fu Huang, Sheng-Min Lin, Huan-Yu Wu, and Yu-Siou Li. 2014. Music genre classification based on local feature selection using a self-adaptive harmony search algorithm. *Data Knowl. Eng.* 92 (2014), 60–76.

[90] Ihab F. Ilyas, George Beskales, and Mohamed A. Soliman. 2008. A survey of top-*k* query processing techniques in relational database systems. *ACM Comput. Surv.* 40, 4 (2008), 11.

[91] Nazir A. Jairazbhoy. 1975. An interpretation of the 22 śrutis. *J. Asian Music* 6, 1–2 (1975), 38–59. http://www.jstor.org/stable/833842

[92] S. R. Janakiraman. 2008. *Essentials of Musicology in South Indian Music*. Vedams.

[93] J.-S. R. Jang and H.-R. Lee. 2008. A general framework of progressive filtering and its application to query by singing/humming. *IEEE Trans. Audio, Speech, Lang. Process.* 16, 2 (2008), 350–358.

[94] J.-S. Roger Jang and M.-Y. Gao. 2000. A query-by-singing system based on dynamic programming. In *Proceedings of the International Workshop on Intelligent System Resolutions (8th Bellman Continuum)*. 85–89.

[95] Jyh-Shing Roger Jang, Hong-Ru Lee, and Ming-Yang Kao. 2001. Content-based music retrieval using linear scaling and branch-and-bound tree search. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'01)*, Vol. 1. 289–292.

[96] J.-S. Roger Jang, Nien-Jung Lee, and Chao-Ling Hsu. 2006. Simple but effective methods for QBSH at MIREX 2006. *Proceedings of MIREX 2006*. 77.

[97] Jesper Hojvang Jensen, Mads Grasboll Christensen, Daniel P. W. Ellis, and Soren Holdt Jensen. 2009. Quantitative analysis of a common audio similarity measure. *IEEE Trans. Audio, Speech, Lang. Process.* 17, 4 (2009), 693–703.

[98] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. 2002. Music type classification by spectral contrast feature. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'02)*, Vol. 1. IEEE, Los Alamitos, CA, 113–116.

[99] Swarupa Joshi and Abhijit Chitre. 2015. Identification of indian musical instruments by feature analysis with different classifiers. In *Proceedings of the 6th International Conference on Computer and Communication Technology*. ACM, New York, NY, 110–114.

[100] Ian Kaminskyj. 2001. Multi-feature musical instrument sound classifier. *MikroPolyphonie* 6, 1–8.

[101] W.-T. Kao, C.-C. Wang, K. K. Chang, J.-S. R. Jang, and W. Liou. 2013. A two-stage query by singing/humming system on GPU. In *Proceedings of the IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA'13)*. IEEE, Los Alamitos, CA, 1–6.

[102] Trupti Katte. 2013. Multiple techniques for raga identification in Indian classical music. *Int. J. Electron. Commun. Comput. Eng.* 4, 6 (2013), 82–87.

[103] Roger A. Kendall. 1986. The role of acoustic signal partitions in listener categorization of musical phrases. *Music Percept.* 4, 185–213.

[104] Eamonn Keogh and Chotirat Ann Ratanamahatana. 2005. Exact indexing of dynamic time warping. *Knowl. Inform. Syst.* 7, 3 (2005), 358–386.

[105] Hyoung-Gook Kim and Thomas Sikora. 2004. Audio spectrum projection based on several basis decomposition algorithms applied to general sound recognition and audio segmentation. In *Proceedings of the 12th European Signal Processing Conference*. IEEE, Los Alamitos, CA, 1047–1050.

[106] Youngmoo E. Kim and Brian Whitman. 2002. Singer identification in popular music recordings using voice coding features. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'02)*. 164–169.

[107] Youngmoo E. Kim, Donald S. Williamson, and Sridhar Pilli. 2006. Towards quantifying the "album effect" in artist identification. In *Proceedings of the 7th International Symposium on Music Information Retrieval (ISMIR'06)*. 393–394.

[108] Tetsuro Kitahara. 2010. Mid-level representations of musical audio signals for music information retrieval. In *Advances in Music Information Retrieval*. Springer, 65–91.

[109] Tetsuro Kitahara, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. 2007. Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps. *EURASIP J. Appl. Signal Process.* 2007, 1 (2007), 155.

[110] Anssi P. Klapuri. 2003. Multiple fundamental frequency estimation based on harmonicity and spectral smoothness. *IEEE Trans. Speech Audio Process.* 11, 6 (2003), 804–816.

[111] Gopala Krishna Koduri, Sankalp Gulati, and Preeti Rao. 2011. A survey of raaga recognition techniques and improvements to the state-of-the-art. In *Proceedings of the International Conference of Sound and Music Computing (SMC'11)*.

[112] Gopala Krishna Koduri, Sankalp Gulati, Preeti Rao, and Xavier Serra. 2012. Rāga recognition based on pitch distribution methods. *J. New Music Res.* 41, 4 (2012), 337–350.

[113] Sefki Kolozali, Mathieu Barthet, Gyorgy Fazekas, and Mark Sandler. 2013. Automatic ontology generation for musical instruments based on audio analysis. *IEEE Trans. Audio, Speech, Lang. Process.* 21, 10 (2013), 2207–2220.

[114] Mark D. Korhonen, David Clausi, M. Jernigan, et al. 2005. Modeling emotional content of music using system identification. *IEEE Trans. Syst., Man, Cybernetics, Part B: Cybern.* 36, 3 (2005), 588–599.

[115] Alexios Kotsifakos, Panagiotis Papapetrou, Jaakko Hollmén, and Dimitrios Gunopulos. 2011. A subsequence matching with gaps-range-tolerances framework: A query-by-humming application. *Proceedings of the VLDB Endowment* 4, 11 (2011), 761–771.

[116] Alexios Kotsifakos, Panagiotis Papapetrou, Jaakko Hollmén, Dimitrios Gunopulos, and Vassilis Athitsos. 2012. A survey of query-by-humming similarity methods. In *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments*. ACM, New York, NY, 5.

[117] A. G. Krishna and T. V. Sreenivas. 2004. Music instrument recognition from isolated notes to solo phrases. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Vol. 4. IEEE, Los Alamitos, CA, 265–268.

[118] T. M. Krishna and V. Ishwar. 2012. Carnatic music: Svara, gamaka, motif and raga identity. In *Proceedings of the 2nd CompMusic Workshop*.

[119] Arvindh Krishnaswamy. 2004. Melodic atoms for transcribing carnatic music. In *Proceedings of the 5th International Symposium on Music Information Retrieval (ISMIR'04)*. 63–345.

[120] Carol L. Krumhansl. 2001. *Cognitive Foundations of Musical Pitch*. Oxford University Press.

[121] V. Kumar, H. Pandya, and C. V. Jawahar. 2014. Identifying ragas in Indian music. In *Proceedings of the 22nd International Conference on Pattern Recognition (ICPR'14)*. IEEE, Los Alamitos, CA, 767–772.

[122] K. Kushima, M. Satoh, H. Akama, and M. Yamamuro. 2000. Integrating hierarchical classification and content-based image retrieval—image compass. In *Proceedings of the International Conference on Intelligent Information Processing*. 179–187.

[123] Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Hwai-San Lin. 2009. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *IEEE Trans. Multimed.* 11, 4 (2009), 670–682.

[124] Bernhard Lehner, Gerhard Widmer, and Sebastian Bock. 2015. A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks. In *Proceedings of the 23rd IEEE European Signal Processing Conference (EUSIPCO'15)*. IEEE, Los Alamitos, CA, 21–25.

[125] Kjell Lemström and Sami Perttu. 2000. SEMEX-An efficient music retrieval prototype. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR'00)*.

[126] Kjell Lemström and Esko Ukkonen. 2000. Including interval encoding into edit distance based music comparison and retrieval. In *Proceedings of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (AISB'00)*. 53–60.

[127] Pierre Leveau, David Sodoyer, and Laurent Daudet. 2007. Automatic instrument recognition in a polyphonic mixture using sparse representations. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR'07)*. 233–236.

[128] Guohui Li and Ashfaq A. Khokhar. 2000. Content-based indexing and retrieval of audio data using wavelets. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'00)*, Vol. 2. IEEE, Los Alamitos, CA, 885–888.

[129] Tao Li and Mitsunori Ogihara. 2003. Detecting emotion in music. In *Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR'03)*, Vol. 3. 239–240.

[130] Tao Li and Mitsunori Ogihara. 2005. Music genre classification with taxonomy. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Vol. 5. IEEE, Los Alamitos, CA, 197–200.

[131] Tao Li and Mitsunori Ogihara. 2006. Toward intelligent music information retrieval. *IEEE Trans. Multimed.* 8, 3 (2006), 564–574.

[132] Tao Li, Mitsunori Ogihara, and Qi Li. 2003. A comparative study on content-based music genre classification. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 282–289.

[133] Yipeng Li and DeLiang Wang. 2006. Singing voice separation from monaural recordings. In *Proceedings of the 7th International Symposium on Music Information Retrieval (ISMIR'06)*. 176–179.

[134] Thomas Lidy and Andreas Rauber. 2005. Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR'05)*. 34–41.

[135] Thomas Lidy and Andreas Rauber. 2005. MIREX 2005: Combined fluctuation features for music genre classification. In *Proceedings of the 6th Annual International Symposium on Music Information Retrieval (ISMIR'05)*.

[136] Thomas Lidy, Andreas Rauber, Antonio Pertusa, and José Manuel Iñesta Quereda. 2007. Improving genre classification by combination of audio and symbolic descriptors using a transcription systems. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR'07)*. 61–66.

[137] David Little and Bryan Pardo. 2008. Learning musical instruments from mixtures of audio with weak labels. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR'08)*, Vol. 8. 127–132.

[138] Chih-Chin Liu and Chuan-Sung Huang. 2002. A singer identification technique for content-based classification of MP3 music objects. In *Proceedings of the 11th International Conference on Information and Knowledge Management*. ACM, New York, NY, 438–445.

[139] Arie A. Livshin and Xavier Rodet. 2004. Instrument recognition beyond separate notes—indexing continues recordings. In *Proceedings of the International Computing Music Conference*.

[140] Arie A. Livshin and Xavier Rodet. 2004. Musical instrument identification in continuous recordings. In *Proceedings of the 7th International Conference on Digital Audio Effects*. 1–5.

[141] Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Symposium on Music Information Retrieval (ISMIR'00)*.

[142] Lie Lu, Dan Liu, and Hong-Jiang Zhang. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio, Speech, Lang. Process.* 14, 1 (2006), 5–18.

[143] Laura Williams Macy. 2001. *Grove Music Online*. Macmillan.

[144] Namunu Chinthaka Maddage, Changsheng Xu, and Ye Wang. 2004. Singer identification based on vocal and instrumental models. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, Vol. 2. IEEE, Los Alamitos, CA, 375–378.

[145] Mrinal K. Mandal, Tyseer Aboulnasr, and Sethuraman Panchanathan. 1998. Fast wavelet histogram techniques for image indexing. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*. IEEE, Los Alamitos, CA, 68–72.

[146] M. I. Mandel and D. P. W. Ellis. 2005. Song-level features and support vector machines for music classification. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR'05)*. 594–599.

[147] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze et al. 2008. *Introduction to Information Retrieval*. Vol. 1. Cambridge University Press.

[148] Ugo Marchand and Geoffroy Peeters. 2016. The extended ballroom dataset. In the *Late-Breaking-Demo Session of the 17th International Society for Music Information Retrieval Conference (ISMIR'16)*.

[149] Janet Marques and Pedro J. Moreno. 1999. A study of musical instrument classification using Gaussian mixture models and support vector machines. *Technical Report Series CRL* 4. Cambridge Research Laboratory.

[150] Yossi Matias, Jeffrey Scott Vitter, and Min Wang. 1998. Wavelet-based histograms for selectivity estimation. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, Vol. 27. ACM, New York, NY, 448–459.

[151] Matthias Mauch, Hiromasa Fujihara, Kazuyoshi Yoshii, and Masataka Goto. 2011. Timbre and melody features for the recognition of vocal activity and instrumental solos in polyphonic music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11).* 233–238.

[152] Dominic Mazzoni and Roger B. Dannenberg. 2001. Melody matching directly from audio. In *Proceedings of the 2nd Annual International Symposium on Music Information Retrieval (ISMIR'01).* 17–18.

[153] Stephen McAdams, Suzanne Winsberg, Sophie Donnadieu, Geert De Soete, and Jochen Krimphoff. 1995. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychol. Res.* 58, 3 (1995), 177–192.

[154] Cory McKay, Daniel McEnnis, and Ichiro Fujinaga. 2006. A large publicly accessible prototype audio database for music research. In *Proceedings of the 7th Annual International Symposium on Music Information Retrieval (ISMIR'06).* 160–163.

[155] Martin F. McKinney and Jeroen Breebaart. 2003. Features for audio and music classification. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'03)*, Vol. 3. 151–158.

[156] M. Mellody and G. H. Wakefield. 2000. Signal analysis of the singing voice: Low-order representations of singer identity. In *Proceedings of the International Computer Music Conference (ICMC'00).* 98–101.

[157] Anders Meng, Peter Ahrendt, and Jan Larsen. 2005. Improving music genre classification by short time feature integration. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)*, Vol. 5. IEEE, Los Alamitos, CA, 497.

[158] Anders Meng, Peter Ahrendt, Jan Larsen, and Lars Kai Hansen. 2007. Temporal feature integration for music genre classification. *IEEE Trans. Audio, Speech, Lang. Process.* 15, 5 (2007), 1654–1664.

[159] Aanamaria Mesaros and Jaakko Astola. 2005. Inter-dependence of spectral measures for the singing voice. In *Proceedings of the International Symposium on Signals, Circuits, and Systems (ISSCS'05).* IEEE, Los Alamitos, CA, 307–310.

[160] Annamaria Mesaros and Jaakko Astola. 2005. The mel-frequency cepstral coefficients in the context of singer identification. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR'05).* 610–613.

[161] Annamaria Mesaros, Tuomas Virtanen, and Anssi Klapuri. 2007. Singer identification in polyphonic music using vocal separation and pattern recognition methods. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR'07).* 375–378.

[162] Luca Mion and Giovanni De Poli. 2008. Score-independent audio features for description of music expression. *IEEE Trans. Audio, Speech, Language Process.* 16, 2 (2008), 458–466.

[163] Riccardo Miotto and Gert Lanckriet. 2012. A generative context model for semantic music annotation and retrieval. *IEEE Trans. Audio, Speech, Language Process.* 20, 4 (2012), 1096–1108.

[164] Marcel Mongeau and David Sankoff. 1990. Comparison of musical sequences. *Comput. Humanit.* 24, 3 (1990), 161–175.

[165] Fabian Mörchen, Alfred Ultsch, Michael Thies, and Ingo Löhken. 2006. Modeling timbre distance with temporal statistics from polyphonic music. *IEEE Trans. Audio, Speech, Lang. Process.* 14, 1 (2006), 81–90.

[166] Y. V. Srinivasa Murthy and Shashidhar G. Koolagudi. 2015. Classification of vocal and non-vocal regions from audio songs using spectral features and pitch variations. In *Proceedings of the 28th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE'15).* IEEE, Los Alamitos, CA, 1271–1276.

[167] G. R. K. Murty. 2014. India's romance with monsoon rains: A peep into poetic expressions and personal experiences. *IUP J. Engl. Stud.* 9, 3 (2014), 54–73.

[168] Oxford Music. 2010. *Benjamin West's "Jonah": A previously overlooked illustration for the first oratorio composed in the New World. Am. Art J.* 28, 1–2 (2010), 122–137.

[169] Juhan Nam, Jorge Herrera, Malcolm Slaney, and Julius O. Smith. 2012. Learning sparse feature representations for music annotation and retrieval. In *Proceedings of the 13th International Symposium on Music Information Retrieval (ISMIR'12).* 565–570.

[170] Steven R. Ness, Anthony Theocharis, George Tzanetakis, and Luis Gustavo Martins. 2009. Improving automatic music tag annotation using stacked generalization of probabilistic SVM outputs. In *Proceedings of the 17th ACM International Conference on Multimedia.* ACM, New York, NY, 705–708.

[171] Nicola Orio. 2006. *Music Retrieval: A Tutorial and Review.* Now Publishers.

[172] Nicola Orio, David Rizo, Riccardo Miotto, Markus Schedl, Nicola Montecchio, and Olivier Lartillot. 2011. MusiCLEF: A benchmark activity in multi-modal music information retrieval. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11).* 603–608.

[173] Francois Pachet and Jean-Julien Aucouturier. 2004. Improving timbre similarity: How high is the sky. *J. Negat. Results Speech Audio Sci.* 1, 1 (2004), 1–13.

[174] François Pachet, Daniel Cazaly, et al. 2000. A taxonomy of musical genres. In *Proceedings of the Content-Based Multimedia Information Access Conference (RIOA'00).* 1238–1245.

[175] Elias Pampalk, Arthur Flexer, Gerhard Widmer, et al. 2005. Improvements of audio-based music similarity and genre classification. In *Proceedings of 6th International Symposium on Music Information Retrieval (ISMIR'05)*, Vol. 5. 634–637.

[176] Elias Pampalk, Andreas Rauber, and Dieter Merkl. 2002. Content-based organization and visualization of music archives. In *Proceedings of the 10th ACM International Conference on Multimedia.* ACM, New York, NY, 570–579.

[177] Ioannis Panagakis, Emmanouil Benetos, and Constantine Kotropoulos. 2008. Music genre classification: A multi-linear approach. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR'08).* 583–588.

[178] Hemant A. Patil, Purushotam G. Radadia, and T. K. Basu. 2012. Combining evidences from mel cepstral features and cepstral mean subtracted features for singer identification. In *Proceedings of the International Conference on Asian Language Processing (IALP'12).* IEEE, Los Alamitos, CA, 145–148.

[179] Steffen Pauws. 2002. CubyHum: A fully operational query by humming system. In *Proceedings of the 3rd International Symposium on Music Information Retrieval (ISMIR'02).* 187–196.

[180] Geoffroy Peeters. 2003. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. In *Proceedings of the Audio Engineering Society Convention 115.*

[181] Geoffroy Peeters. 2004. *A Large Set of Audio Features for Sound Description (Similarity and Classification) in the CUIDADO Project.* Technical Report. Ircam.

[182] David Perrot and Robert Gjerdigen. 1999. Scanning the dial: An exploration of factors in the identification of musical style. In *Proceedings of the Society for Music Perception and Cognition.* 88.

[183] Reinier Plomp. 1989. The perception of timbre of steady-state complex tones. *J. Acoust. Soc. Am.* 86, S1 (1989), S57.

[184] Arthur N. Popper and Richard R. Fay. 2014. *Perspectives on Auditory Research.* Vol. 50. Springer.

[185] Harold Stone Powers. 1963. *The Background of the South Indian Raga-System.* University Microfilms.

[186] Yingyong Qi. 1992. Time normalization in voice analysis. *J. Acoust. Soc. Am.* 92, 5 (1992), 2569–2576.

[187] Lawrence Rabiner and Biing-Hwang Juang. 1993. *Fundamentals of Speech Recognition.* Prentice Hall.

[188] Mathieu Ramona, Gaël Richard, and Bertrand David. 2008. Vocal detection in music with support vector machines. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08).* IEEE, Los Alamitos, CA, 1885–1888.

[189] H. G. Ranjani, S. Arthi, and T. V. Sreenivas. 2011. Carnatic music analysis: Shadja, swara identification and raga verification in alapana using stochastic models. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'11).* IEEE, Los Alamitos, CA, 29–32.

[190] Jeremy Reed and Chin-Hui Lee. 2009. On the importance of modeling temporal information in music tag annotation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09).* IEEE, Los Alamitos, CA, 1873–1876.

[191] Lise Regnier and Geoffroy Peeters. 2009. Singing voice detection in music tracks using direct voice vibrato detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09).* IEEE, Los Alamitos, CA, 1685–1688.

[192] Douglas A. Reynolds. 1994. Experimental evaluation of features for robust speaker identification. *IEEE Trans. Speech Audio Process.* 2, 4 (1994), 639–643.

[193] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *Int.J. Comput.Vis.* 40, 2 (2000), 99–121.

[194] Justin Salamon, Sankalp Gulati, and Xavier Serra. 2012. A multi-pitch approach to tonic identification in Indian classical music. In *Proceedings of the 13th International Symposium on Music Information Retrieval (ISMIR'12).*

[195] Pichu Sambamoorthy. 1958. *South Indian Music.* Vol. 3. Indian Music Publishing House.

[196] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek. 2006. Automatic genre classification of music content: A survey. *IEEE Signal Process. Mag.* 23, 2 (2006), 133–141.

[197] Markus Schedl, Emilia Gómez, and Julián Urbano. 2014. Music information retrieval: Recent developments and applications. *Found. Trends Inform. Retriev.* 8, 2–3 (2014), 127–261.

[198] Markus Schedl, Peter Knees, and Gerhard Widmer. 2005. *Interactive Poster: Using CoMIRVA for Visualizing Similarities between Music Artists.* Department of Computational Perception, Johannes Kepler University.

[199] Christian Schörkhuber and Anssi Klapuri. 2010. Constant-Q transform toolbox for music processing. In *Proceedings of the 7th Sound and Music Computing Conference.* 3–64.

[200] Klaus Seyerlehner, Gerhard Widmer, and Tim Pohle. 2010. Fusing block-level features for music similarity estimation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10).* 225–232.

[201] Hiteshwari Sharma and Rasmeet S. Bali. 2014. Raga identification of hindustani music using soft computing techniques. In *Proceedings of the Recent Advances in Engineering and Computational Sciences (RAECS'14).* IEEE, Los Alamitos, CA, 1–6.

[202] Rahul Sharma, Y. V. Srinivasa Murthy, and Shashidhar G. Koolagudi. 2016. Audio songs classification based on music patterns. In *Proceedings of the 2nd International Conference on Computer and Communication Technologies*. 157–166.
[203] J. S. Shawe-Taylor and A. Meng. 2005. An investigation of feature models for music genre classification using the support vector classifier. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR'05)*. 604–609.
[204] Jialie Shen, John Shepherd, Bin Cui, and Kian-Lee Tan. 2009. A novel framework for efficient automated singer identification in large music databases. *ACM Trans. Inform. Syst.* 27, 3 (2009), 18.
[205] Jialie Shen, John Shepherd, and Anne H. H. Ngu. 2006. Towards effective content-based music retrieval with multiple acoustic feature combination. *IEEE Trans. Multimed.* 8, 6 (2006), 1179–1189.
[206] Arun Shenoy, Yuansheng Wu, and Ye Wang. 2005. Singing voice detection for karaoke application. In *Proceedings of the Conference on Visual Communications and Image Processing*. 596028–596028.
[207] S. Shetty and K. K. Achary. 2009. Raga mining of Indian music by extracting arohana-avarohana pattern. *Int. J. Recent Trends Engin.* 1, 1 (2009), 362–366.
[208] Carlos Nascimento Silla Jr., Alessandro L. Koerich, and Celso A. A. Kaestner. 2008. The Latin music database. In *Proceedings of the 9th Annual International Symposium on Music Information Retrieval (ISMIR'08)*. 451–456.
[209] Andrew J. R. Simpson, Gerard Roma, and Mark D. Plumbley. 2015. Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network. In *Proceedings of the 11th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'15)*. 429–436.
[210] Malcolm Slaney. 2002. Semantic-audio retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02)*, Vol. 4. IEEE, Los Alamitos, CA, 4108–4111.
[211] John R. Smith and Chung-Sheng Li. 1999. Image classification and querying using composite region templates. *Comput. Vis. Image Underst.* 75, 1 (1999), 165–174.
[212] Yading Song, Simon Dixon, and Marcus Pearce. 2012. A survey of music recommendation systems and future perspectives. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR'12)*.
[213] Yangqiu Song and Changshui Zhang. 2008. Content-based information fusion for semi-supervised music genre classification. *IEEE Trans. Multimedi.* 10, 1 (2008), 145–152.
[214] Rajeswari Sridhar and T. V. Geetha. 2009. Raga identification of carnatic music for music information retrieval. *Int. J. Recent Trends Eng.* 1, 1 (2009), 571–574.
[215] P. Sriram. 1990. *A Carnatic Music Primer*. Paramus: Carnatic Music Association of North America.
[216] Hans Werner Strube. 1980. Linear prediction on a warped frequency scale. *J. Acoust. Soc. Am.* 68, 4 (1980), 1071–1076.
[217] Bob L. Sturm. 2013. The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. arXiv:1306.1461.
[218] S. M. Suma and S. G. Koolagudi. 2015. Raga classification for carnatic music. In *Proceedings of the Conference on Information Systems Design and Intelligent Applications (INDIA'15)*. 865–875.
[219] Johan Sundberg and Thomas D. Rossing. 1990. The science of singing voice. *J. Acoust. Soc. Am.* 87, 1 (1990), 462–463.
[220] Gordon N. Swift. 1990. Ornamentation in South Indian music and the violin. *J. Soc. Asian Music* 21, 2 (1990), 71–89.
[221] Auke Tellegen, David Watson, and Lee Anna Clark. 1999. On the dimensional and hierarchical structure of affect. *Psychol. Sci.* 10, 4 (1999), 297–303.
[222] Robert E. Thayer. 1989. *The Bio-Psychology of Mood and Arousal*. Oxford University Press.
[223] Derek Tingle, Youngmoo E. Kim, and Douglas Turnbull. 2010. Exploring automatic music annotation with acoustically-objective tags. In *Proceedings of the 11th International Symposium on Multimedia Information Retrieval (ISMIR'10)*. 55–62.
[224] Tero Tolonen and Matti Karjalainen. 2000. A computationally efficient multi-pitch analysis model. *IEEE Trans. Speech Audio Process.* 8, 6 (2000), 708–716.
[225] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P. Vlahavas. 2008. Multi-label classification of music into emotions. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR'08)*, Vol. 8. 325–330.
[226] Wei-Ho Tsai, Shih-Jie Liao, and Catherine Lai. 2008. Automatic identification of simultaneous singers in duet recordings. In *Proceedings of the 9th International Symposium on Music Information Retrieval (ISMIR'08)*. 115–120.
[227] W.-H. Tsai and H.-M. Wang. 2004. Automatic detection and tracking of target singer in multi-singer music recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, Vol. 4. IEEE, Los Alamitos, CA, 221.
[228] W.-H. Tsai and H.-M. Wang. 2006. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *IEEE Trans. Audio, Speech, Lang. Process.* 14, 1 (2006), 330–341.
[229] Emiru Tsunoo, George Tzanetakis, Nobutaka Ono, and Shigeki Sagayama. 2009. Audio genre classification using percussive pattern clustering combined with timbral features. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'09)*. IEEE, Los Alamitos, CA, 382–385.

[230] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. 2007. Towards musical query-by-semantic-description using the CAL500 dataset. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, New York, NY, 439–446.

[231] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. 2008. Semantic annotation and retrieval of music and sound effects. *IEEE Trans. Audio, Speech, Lang. Process.* 16, 2 (2008), 467–476.

[232] Rainer Typke, Panos Giannopoulos, Remco C. Veltkamp, Frans Wiering, and René Van Oostrum. 2003. Using transportation distances for measuring melodic similarity. In *Proceedings of the 4th Annual International Symposium on Music Information Retrieval (ISMIR'03).* 107–114.

[233] Rainer Typke, Frans Wiering, and Remco C. Veltkamp. 2005. A survey of music information retrieval systems. In *Proceedings of the 6th International Symposium on Music Information Retrieval (ISMIR'05).* ISMIR, 153–160.

[234] George Tzanetakis and Perry Cook. 2002. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.* 10, 5 (2002), 293–302.

[235] George Tzanetakis, Randy Jones, and Kirk McNally. 2007. Stereo panning features for classifying recording production style. In *Proceedings of the 8th International Symposium on Music Information Retrieval (ISMIR'07).* 441–444.

[236] George Tzanetakis, Luis Gustavo Martins, Kirk McNally, and Randy Jones. 2010. Stereo panning information for music information retrieval tasks. *J. Audio Eng. Soc.* 58, 5 (2010), 409–417.

[237] Karthikeyan Umapathy, Sridhar Krishnan, and Shihab Jimaa. 2005. Multi-group classification of audio signals using time-frequency parameters. *IEEE Trans. Multimed.* 7, 2 (2005), 308–315.

[238] Erdem Unal, Elaine Chew, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. 2008. Challenging uncertainty in query by humming systems: A fingerprinting approach. *IEEE Trans. Audio, Speech, Lang. Process.* 16, 2 (2008), 359–371.

[239] Remco C. Veltkamp and Mirela Tanase. 2001. Content-based image retrieval systems: A survey. In *Proceedings of the Dagstuhl Seminar on State-of-the-Art in Content-Based Image and Video Retrieval.* 196–203.

[240] Shankar Vembu and Stephan Baumann. 2005. Separation of vocals from polyphonic audio recordings. In *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR'05).* 337–344.

[241] T. Allen Viswanathan, Matthew Harp T. Viswanathan, and Matthew Harp Allen. 2004. *Music in South India: The Karnatak Concert Tradition and Beyond: Experiencing Music, Expressing Culture.* Cambridge University Press..

[242] H. M. Vyas, S. M. Suma, S. G. Koolagudi, and K. R. Guruprasad. 2015. Identifying gamakas in carnatic music. In *Proceedings of the 8th International Conference on Contemporary Computing (IC3'15).* IEEE, Los Alamitos, CA, 106–110.

[243] Chung-Che Wang, Jyh-Shing Roger Jang, and Wennen Wang. 2010. An improved query by singing/humming system using melody and lyrics information. In *Proceedings of the 10th Annual International Symposium on Music Information Retrieval (ISMIR'10)*, Vol. 4571. 45–50.

[244] Qiang Wang, Zhiyuan Guo, Baoxiang Li, Gang Liu, and Jun Guo. 2012. Tempo variation based multi-layer filters for query by humming. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR'12).* IEEE, Los Alamitos, CA, 3034–3037.

[245] Shuo-Yang Wang, Ju-Chiang Wang, Yi-Hsuan Yang, and Hsin-Min Wang. 2014. Towards time-varying music auto-tagging based on CAL500 expansion. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'14).* IEEE, Los Alamitos, CA, 1–6.

[246] Claus Weihs, Uwe Ligges, Fabian Mörchen, and Daniel Müllensiefen. 2007. Classification in music research. *Adv. Data Anal. Classif.* 1, 3 (2007), 255–291.

[247] Kris West. 2008. *Novel Techniques for Audio Music Classification and Search.* Ph.D. Dissertation. University of East Anglia.

[248] Brian Whitman, Gary Flake, and Steve Lawrence. 2001. Artist detection in music with Minnowmatch. In *Proceedings of the IEEE Signal Processing Society Workshop on Neural Networks for Signal Processing.* IEEE, Los Alamitos, CA, 559–568.

[249] Kay Wolter, Christoph Bastuck, and Daniel Gärtner. 2008. Adaptive user modeling for content-based music retrieval. In *Proceedings of the 6th Workshop on Adaptive Multimedia Retrieval: Identifying, Summarizing, and Recommending Image and Music.* 40–52.

[250] Xiao Wu, Ming Li, Jian Liu, Jun Yang, and Yonghong Yan. 2006. A top-down approach to melody match in pitch contour for query by humming. In *Proceedings of 5th International Symposium on Chinese Spoken Language Processing.*

[251] Jingzhou Yang, Jia Liu, and Weiqiang Zhang. 2010. A fast query by humming system based on notes. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH'10).* 2898–2901.

[252] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen. 2008. A regression approach to music emotion recognition. *IEEE Trans. Audio, Speech, Lang. Process.* 16, 2 (2008), 448–457.

[253] Atsuo Yoshitaka and Tadao Ichikawa. 1999. A survey on content-based retrieval for multimedia databases. *IEEE Trans. Knowl. Data Eng.* 11, 1 (1999), 81–93.

[254] Alejandro Zentner. 2003. *Measuring the Effect of Online Music Piracy on Music Sales.* Technical Report. University of Chicago.

[255] Tong Zhang. 2003. Automatic singer identification. In *Proceedings of the International Conference on Multimedia and Expo (ICME'03)*, Vol. 1. IEEE, Los Alamitos, CA, 33–36.

[256] T. Zhang and H. Packard. 2003. System and method for automatic singer identification. In *Proceedings of the IEEE International Conference on Multimedia and Expo.* 756.

[257] Yunyue Zhu and Dennis Shasha. 2003. Warping indexes with envelope transforms for query by humming. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data (SIGMOD'03).* ACM, New York, NY, 181–192.

Received August 2016; revised October 2017; accepted January 2018