ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/265389987

Music information retrieval in compressed audio files: A survey

Article *in* New Review in Hypermedia and Multimedia · July 2014

DOI: 10.1080/13614568.2014.889223

CITATIONS

0

READS

127

2 authors:

Markos Zampoglou

The Centre for Research and Technology, Hellas

31 PUBLICATIONS **319** CITATIONS

SEE PROFILE

Athanasios G. Malamos

Hellenic Mediterranean University

70 PUBLICATIONS **245** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project

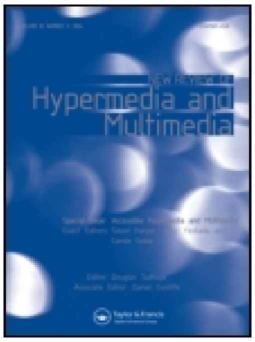
Web 3D and VR Technology in cultural applications View project

Project

Information Society:Collection, digitalization and documentation of material relating to the particular characteristics of Lasithi in culture, tourism, environment and economy" View project

All content following this page was uploaded by Markos Zampoglou on 11 November 2015.

The user has requested enhancement of the downloaded file.



New Review of Hypermedia and Multimedia

ISSN: 1361-4568 (Print) 1740-7842 (Online) Journal homepage: http://www.tandfonline.com/loi/tham20

Music information retrieval in compressed audio files: a survey

Markos Zampoglou & Athanasios G. Malamos

To cite this article: Markos Zampoglou & Athanasios G. Malamos (2014) Music information retrieval in compressed audio files: a survey, New Review of Hypermedia and Multimedia, 20:3, 189-206, DOI: 10.1080/13614568.2014.889223

To link to this article: http://dx.doi.org/10.1080/13614568.2014.889223

Published online: 14 Aug 2014.

Submit your article to this journal

Article views: 125

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=tham20

Download by: [95.92.242.12] **Date:** 05 November 2015, At: 02:23

New Review of Hypermedia and Multimedia, 2014
Vol. 20, No. 3, 189–206, http://dx.doi.org/10.1080/13614568.2014.889223

Taylor & Francis
Taylor & Francis Group

Music information retrieval in compressed audio files: a survey

MARKOS ZAMPOGLOU AND ATHANASIOS G. MALAMOS*
Department of Informatics Engineering, Technological Educational Institute of Crete, Crete, Greece

(*Received 26 March 2013; final version received 12 January 2014*)

In this paper, we present an organized survey of the existing literature on music information retrieval systems in which descriptor features are extracted directly from the compressed audio files, without prior decompression to pulse-code modulation format. Avoiding the decompression step and utilizing the readily available compressed-domain information can significantly lighten the computational cost of a music information retrieval system, allowing application to large-scale music databases. We identify a number of systems relying on compressed-domain information and form a systematic classification of the features they extract, the retrieval tasks they tackle and the degree in which they achieve an actual increase in the overall speed—as well as any resulting loss in accuracy. Finally, we discuss recent developments in the field, and the potential research directions they open toward ultra-fast, scalable systems.

Keywords: Music information retrieval; Audio compression; MPEG 7

1. Introduction

The term music information retrieval (MIR) encompasses a wide range of tasks, aimed at facilitating access and management of various aspects of musical information. After more than a decade of advances, (Casey et al., 2008; Grachten, Schedl, Pohle, & Widmer, 2009; Orio, 2006), the field has now grown to the point where specialized surveys can be found for specific MIR problems, such as emotion detection (Kim et al., 2010; Yang & Chen, 2012), query-by-humming (Kotsifakos, Papapetrou, Hollmen, Gunopulos, & Athitsos, 2012), multimodality (Mayer & Rauber, 2010), and adaptivity (Stober & Nürnberger, 2013).

However, one aspect that is often overlooked is that of scalability. MIR systems can be computationally demanding, while, at the same time, users often expect to see real-time applications with low response times. Considering the fact that online and local databases will most likely continue to grow in size over time, the issue of computational cost comes to the forefront. In this context, there are two processing steps found in virtually every MIR system, whose computational cost can be reduced, or avoided altogether.

*Corresponding author. Email: amalamos@epp.teicrete.gr

© 2014 Taylor & Francis

Downloaded by [95.92.242.12] at 02:23 05 November 2015

Downloaded by [95.92.242.12] at 02:23 05 November 2015

To begin with, practically every MIR method is based on the extraction of a number of descriptor features from the audio signal. Since, in today's world, audio databases almost universally store their files in compressed form; the straightforward approach favored by nearly every proposed MIR system so far is to decompress the audio data and perform feature extraction on the resulting pulse-code modulated (PCM) signal. If the descriptors could be extracted directly from the compressed files, the decompression step could be skipped entirely.

Furthermore, today's dominant compression formats commonly include at least one subband decimation step, which separates the original signal in a number of band-limited channels. However, many audio descriptors are also based on a decimation of the original audio signal into subbands (quite often the Short-Time Fourier Transform, or STFT). It then follows that, instead of decompressing the signal and then re-decimating it, the readily available compressed-domain information could be used for descriptor extraction.

Finally, compression formats often include other, potentially valuable types of information, such as the window sizes used in the MPEG-1/2 Audio Layer III (MP3) and the Advanced Audio Coding (AAC) formats. By taking advantage of the information contained in a compression, the computational benefits are twofold: we avoid both the cost of the decompression step, and (hopefully) some of the cost of forming the audio descriptors (Figure 1).

In this paper, we survey proposals made during the last decade to build MIR systems using descriptors extracted directly from the compressed-domain information. The rest of the paper is organized as follows: section 2 briefly describes the compression formats that we deal with in this survey; section 3 presents the state-of-the-art in the field of compressed-domain MIR; section 4 explores the degree at which the stated aims of the field,—that is, offering increased speed at a low or zero cost in precision—are being achieved; finally, section 5 presents the most recent developments and lays the groundwork for future research in the field.

2. Audio compression standards

While a large number of compression standards for audio files have been proposed over the years, in this paper we will solely focus on the ones that have drawn attention from the music information retrieval research community: MPEG-1/2 part 3 -and especially MP3 and AAC. The aim of drawing information directly from the compressed-domain data is to accelerate feature extraction for existing databases, whose audio files are already in compressed format; as a result, the popularity of certain standards over others has steered research interest toward them.

2.1. MPEG-1/2 Part 3

MP3, formally MPEG-1 Part 3 Layer III (ISO/IEC 11172-3, 1993), is possibly the most popular compression standard for home audio collections and audio transfer. MP3 compression is a complex process, consisting of multiple steps (Pan, 1995). The coarse sequence of operations is (1) decimation of the PCM data into 32 frequency subbands through the use of a polyphase filterbank, (2) Modified

Downloaded by [95.92.242.12] at 02:23 05 November 2015

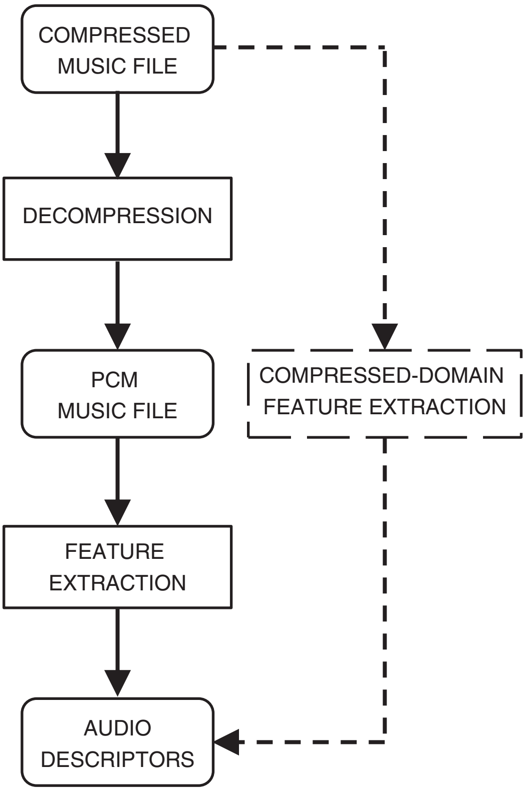


Figure 1. Block diagram of a typical feature extraction process and the compressed-domain alternative.

Discreet Cosine Transform (MDCT) of the subband samples, (3) manipulation of the MDCT coefficients according to a psychoacoustic model and the desired bit rate, and (4) Huffman coding of the resulting values.

This means that musical information goes through three distinct forms during MP3 compression and decompression: PCM, MDCT coefficients, and polyphase filter subband coefficients. When given a music file in MP3 format, the decompression process can be interrupted at any intermediate stage, giving us the corresponding data. Figure 2 shows the distribution of the time requirements for each MP3 decompression step.

The fundamental time unit of an MP3 file is a granule, corresponding to 576 PCM samples. During the polyphase filtering, the samples in a granule are filtered into 32 equally spaced frequency subbands of 18 samples each. The MDCT further subdivides each subband into 18 finer subbands, leading to a granule of 576 MDCT coefficients. Before MDCT, however, windowing takes place.

There are four types of windows in MP3: Normal (or Long), Short, Long-to-Short (or Start), and Short-to-Long (or Stop). Long windows are used in the majority of cases. In Long windows, MDCT is applied directly to the 18 samples, leading to a high frequency resolution of 18 sub-subbands. When the psychoacoustic model decides that the auditory content is changing abruptly, three Short

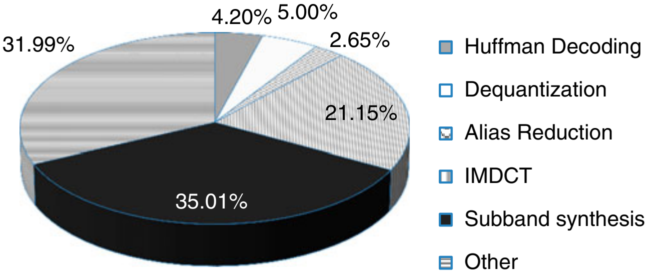


Figure 2. Time requirements for MP3 decoding [values taken from: Tsai and Chang (2009)].

windows are used instead. In this case, three independent MDCTs are run on sets of 6 samples each, giving reduced frequency resolution, but increased temporal one. Start and Stop windows are used as transitory between Long and Short ones.

MPEG-1 Audio Layers I & II are fundamentally similar to Layer III, the main differences being that they do not include an MDCT step, and the granule structure is slightly different. Finally, the differences between Parts 3 of MPEG-1 and MPEG-2 mainly concern the sampling rates and the number of channels supported. However, none of these considerations has a significant effect on the methods discussed here.

2.2. AAC

The Advanced Audio Coding format (found in MPEG-2 Part 7 (ISO/IEC 13818-7, 1997) and MPEG-4 Part 3 (ISO/IEC 14496-3, 2005) has fundamental similarities to MP3, as both algorithms are based on MDCT (Brandenburg, 1999). The main difference is the absence of a granule structure from AAC, and more importantly, the absence of polyphase filters. A Long-Short window structure similar to MP3 exists for AAC as well, with Long windows covering 1024 PCM samples, and Short windows covering 128 PCM samples, thus requiring 8 Short windows to make the equivalent of a Long one. While, much like MP3, a range of other tools are used to improve the perceived quality for a given bit rate, the fundamental components of the algorithm—and the aspects that potentially facilitate feature extraction—revolve around the MDCT and the window structure.

2.3. Compression components used for feature extraction

Systems based on descriptors extracted from compressed-domain data can take advantage of a range of different compression components. The most common ones found in literature are the MP3 and AAC MDCT coefficients and the MPEG-1/2 subband values derived from the polyphase filter banks.[1]

Other compression components that can give useful information are the window size patterns, the Huffman coding information, and the scale factors. A subband's scale factor is a weight that, during decoding, rescales samples to their original range. By definition, it carries information related to the maximum magnitude of the original signal. Table 1 presents the papers surveyed, organized by MIR task and compressed-domain feature source used.

Downloaded by [95.92.242.12] at 02:23 05 November 2015

Table 1. The surveyed approaches, organized by music information retrieval task and compressed-domain feature source.

	Subband values	MDCT coefficients	Window sizes	Scale factors	Huffman bits
Music detection	Nakajima et al. (1999), Tzanetakis and Cook (2000), Shieh (2003), Rizzi et al. (2006), Jarina et al. (2004)	Kiranyaz et al. (2006)		Jarina et al. (2004)	
Song identification	Tsai and Wang (2004)	Tsai et al. (2006), Tsai and Chang (2009), Jiao et al. (2007), Li et al. (2010a), Li et al. (2010b), Liu and Chang (2011), Liu (2012)		Tsai and Wang (2004)	
Query-by-humming	Liu and Tsai (2001)				
Query-by-singing		Lie and Su (2004)			
Singer identification		Liu and Huang (2002)			
Chord recognition		Ravelli et al. (2010)			
Genre classification	Pye (2000), Chang et al. (2008), Rizzi et al. (2008)	Ravelli et al. (2010)			
Beat tracking		Wang and Vilermo (2001), Zhu and Wang (2008), Ravelli et al. (2010)	Wang and Vilermo (2001), D'Aguanno et al. (2006), D'Aguanno and Vercellesi (2007)		Zhu and Wang (2008)
Structure and summary	Liu and Yao (2004)	Shao et al. (2004)			

Downloaded by [95.92.242.12] at 02:23 05 November 2015

Downloaded by [95.92.242.12] at 02:23 05 November 2015

3. Compressed-domain music information retrieval

The prospect of extracting descriptors without having to go through the decoding and band-separation steps has been a particularly appealing one ever since the first steps of content-based audio retrieval. The first attempt to extract audio information from compressed data was published almost two decades ago (Patel & Sethi, 1996), and was an attempt to classify MPEG-1 movie segments into silence, dialog and non-dialog—in effect, a speech detection system with an added component for silence detection. During that first period, compressed-domain information was used in content-based retrieval applications, whose actual task—in the cases where it was applied to music—was the exact matching of a query musical piece in a database. Study (Pfeiffer & Vincent, 2001) offers a definitive survey of pre-2001 compressed-domain features used in audio classification and retrieval. Along the same track, study (Wang, Divakaran, Vetro, Chang, & Sun, 2003) surveys previous attempts to extract descriptors (including audio descriptors) from the compressed domain for the purposes of video indexing. In the latter, significant focus is put on attempts to extract MPEG-7 audio descriptors directly from the compressed domain.

During the last decade, the field of MIR has made tremendous progress, and complex descriptors have been proposed for various specialized tasks. In our paper, we classify the reviewed literature in five categories: music detection, song identification, non-specific retrieval (such as query by humming/singing, or genre classification), beat tracking, and structure analysis.

3.1. Music detection

While not always considered an MIR task, the detection of music in audio files, and its separation from other types of audio content (such as speech, silence, or sound effects) is most certainly a music-related task. A number of successful compressed-domain approaches have been proposed in the past, often as a pre-processing step for audio-based video retrieval.

Regardless of the type of compressed-domain information used by a system (subband values, MDCT coefficients or scale factors), in all works reported here silence is always detected through simple thresholding. However, the distinction between music and speech (and any other sound type considered by the model) can be achieved through a variety of methods.

One set of approaches are those based on the subband values (Nakajima et al., 1999; Rizzi, Buccino, Panella, & Uncini, 2006; Shieh, 2003; Tzanetakis & Cook, 2000). All these approaches form a feature vector based on simple statistics of the subband values (centroid, rolloff, spectral flux, RMS of values, central moments, number of subbands with nonzero values), averaged over a time interval, usually one sec. Another particularly powerful feature seems to be the percentage of MP3 frames or granules to have less than average power—"silent frames," while in (Shieh, 2003) the pitch slope from the subbands is also included as a feature. The resulting feature vectors can be classified through a Bayes discriminant function (Nakajima et al., 1999, Tzanetakis & Cook, 2000), k-nn (Rizzi et al., 2006; Tzanetakis & Cook, 2000) or the <Min-Max, PARC> algorithm (Rizzi et al., 2006).

Downloaded by [95.92.242.12] at 02:23 05 November 2015

A single attempt to discriminate between speech, music, and silence based on the MDCT coefficients is presented in (Kiranyaz, Farooq Qureshi, & Gabbouj, 2006). Similar to the subband-based methods, statistics of the MDCT coefficients (total energy, band energy ratio, and centroid frequency) are used to represent every audio frame. The proposed system also performs fundamental frequency estimation per frame. Classification begins on a per-frame basis, and proceeds by merging similar neighboring frames into segments via an iterative procedure. Extensions of the frame-level features to the segment level takes place, and segments are finally classified using a rule-based system.

Another approach (Jarina, O'Connor, Murphy, & Marlow, 2004) is to take advantage of the MPEG scale factor information. In these approaches, silence is detected by thresholding the sum of the scale factors in every granule, while speech/music differentiation is based on the concept of *peaks*. The sum of the granule scale factors over all subbands gives a function of time, which is thresholded to isolate regions of particularly high energy, called peaks. The width (duration) of peaks, and the rate of their occurrence in time are the features used for a simple but effective rule- based classification system. Besides the peak statistics, two further features are used in classification: a rhythm metric, and the harmonicity ratio. For the calculation of the rhythm feature, a normalized autocorrelation function is computed over the subband scale factors, and peaks in the function are taken to suggest the presence of a beat. The harmonicity ratio is a standard MPEG-7 audio descriptor, expressing the proportion of harmonic components in the file's power spectrum. It, too, can be computed from the subband values.

3.2. Song identification/near-duplicate retrieval

In song identification, a system is given a music file and is expected to find its exact match in the database. Of course, the task thus defined is trivial, and the only research problem is the time complexity of the system. There are two typical variants of the task. One is the introduction of distortions in the query file, such as echo, background noise or time/pitch variations. The other is the use of a short segment of the desired music track as a query, instead of the entire track. With these two extensions, the task applies to the realistic scenario where a user records a song segment in a real-world environment through a portable device and submits it to an on-line database. The database performs the matching and—to name a popular use case—responds with any metadata information available for the song.

The oldest attempt to use compressed-domain data for song identification is (Tsai & Wang, 2004). The MPEG scale factors and subband values of each frame are used as descriptors, following quantization and histogram binning. The statistical nature of the resulting descriptor gives the approach a certain degree of robustness to length variations. This is the only proposal to use the scale factors and the subband values. Most other approaches are based on the MDCT coefficients and are thus aimed at MP3 or AAC files. Thus, (Tsai, Wang, Hung, & Wey, 2006) segments a music file into "slots," and consecutively indexes sequences of the slots' MDCT energy coefficients to form the descriptor. A simple

Downloaded by [95.92.242.12] at 02:23 05 November 2015

melody-line contour descriptor is formed, by comparing each slot's energy content to that of the previous one. Study (Tsai & Chang, 2009) extends the same approach by extracting an approximation of Mel-Frequency Cepstral Coefficients (MFCCs) from the MDCT coefficients, in addition to the melody contour. In (Jiao, Yang, Li, & Niu, 2007), the ratio between the MDCT subband energy and the overall energy of a segment is calculated as a segment-level descriptor, and the difference between the descriptors of consecutive segments is used as the song-level descriptor. Despite its conceptual simplicity, the resulting descriptor demonstrates robustness in the face of transcoding, downsampling, echo addition, and equalization.

Two recent methods in this field are (Li, Liu, & Xue, 2010a) and (Li, Liu, & Xue, 2010b). In the former, the MDCT coefficients are aligned in a 2-D "auditory image," with granule groups forming the *x*-axis and groups of neighboring MDCT bands forming the *y*-axis (Figure 3). The resulting image, which conceptually resembles a STFT spectrogram, is used for the calculation of a Zernlike Moments descriptor. Zernlike Moments is a well-established image texture descriptor (Khotanzad & Hong, 1990) that, in the context described here, can give excellent retrieval results in the face of echo and noise addition, pitch shift, equalization, time-scale modifications, and transcoding. Similarly, in (Li et al., 2010b), the MDCT coefficients are grouped in nine wide bands, and for each granule group, the energy of these bands is normalized to resemble a Probability Mass Function.

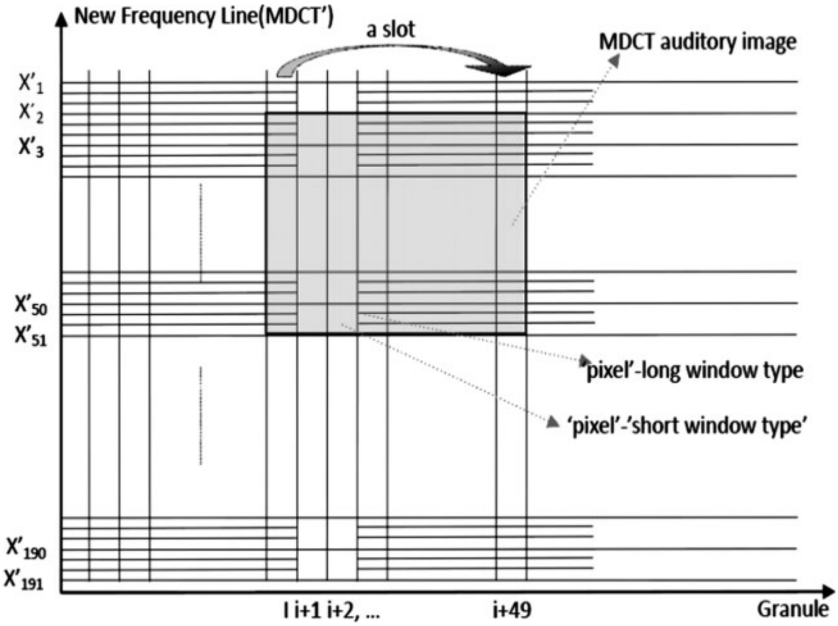


Figure 3. The auditory image of Li et al. (2010a). The 576 MDCT coefficients are summed in 191 subbands, of which only 2–51 are taken into account, and form the *y*-axis. Groups of 50 consecutive granules form the *x*-axis.

The function's entropy is then calculated for each group, and the stability of the entropy descriptor is demonstrated over a range of distortions and alterations.

An important issue to keep in mind, when attempting song identification, is that, the problem has been—to a large extent—tackled. At this point, a multitude of algorithms exist—including the ones mentioned in this section—that can identify song segments in the face of alterations, at reasonable speeds—often in real-time. In this sense, any new proposal at this point should offer extensive comparisons to other identification algorithms. Furthermore, the basic contribution that a compressed-domain algorithm could make to the field is further speed increase, beyond the limitations of PCM-based algorithms. However, concerning both these points, little information is offered. Study (Li et al., 2010a) is the only work to report a comparison to other results, and the comparison is made with the reported success rates of past compressed-domains over different datasets. Overall, no comparison is offered over the same dataset, and no time/complexity analysis is offered for any of the reported works.

A more recent approach is presented in (Liu & Chang, 2011) and (Liu, 2012). In (Liu & Chang, 2011), a set of features consisting of MFCCs, MPEG-7 descriptors and Chroma vectors are all drawn from the MP3 MDCT coefficients, and are used as segment fingerprints, in addition to the original values of 24 MDCT coefficients. Furthermore, PCA dimensionality reduction is proposed for matching, to deal with the high dimensionality of the descriptor vector. This descriptor vector is then used in (Liu, 2012) for copyright infringement detection in network file transfer: streams being transferred between users can be mirrored by a server, and a stream segment can be matched against a database of copyrighted MP3 files. By identifying the track being transferred, it can be evaluated whether the users have permission to exchange the particular file. In order to perform successful matching against a large database, the process is split into two steps: in the first, a coarser fingerprint is formed, based on the Attack (Onset)-Decay-Sustain-Release pattern (see Section 3.3 for details on ADSR). After forming a Mel-spectrum from the MDCT coefficients, nine specific time/frequency points, matching aspects of the ADSR pattern, are isolated. The Mel-spectrum coefficients at these points and various bands, for Mel spectra of three different granularities, are used as a first-level fingerprint in order to identify a candidate subset of the database for matching (Figure 4). Then, exact matching is performed within the subset, using the approach of (Liu & Chang, 2011).

3.3. Non-specific retrieval tasks

A number of MIR tasks are based on performing similarity matching between files that contain different recordings, but which share certain common features: in the case of compressed-domain algorithms, efforts have been made toward query-by-humming, query-by-singing, singer identification, and genre classification.

Query-by-singing and query-by-humming are two very similar tasks, in which the user vocally produces an approximation of a desired music track, and the system attempts to recognize and retrieve the desired track (or its metadata) from the database. The main difference between the two tasks is that in the former the user attempts to reproduce the track lyrics (thus referring to songs and not

Downloaded by [95.92.242.12] at 02:23 05 November 2015

Downloaded by [95.92.242.12] at 02:23 05 November 2015

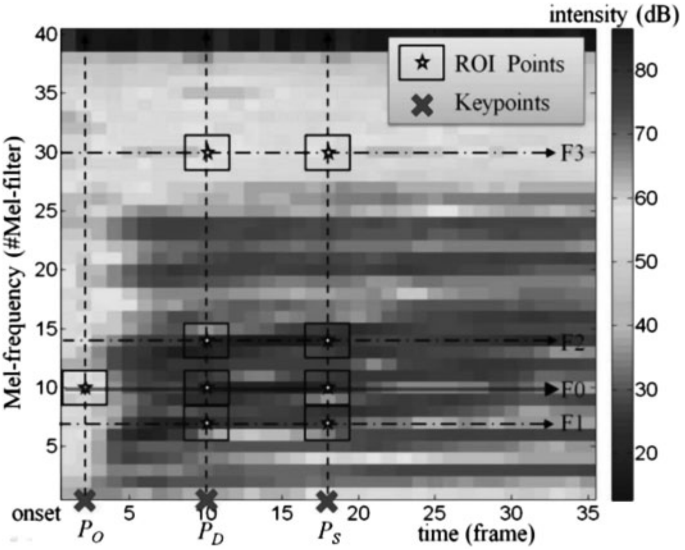


Figure 4. The nine interest points in the high-resolution Mel spectrum for first-level identification. At each point, the nine surrounding coefficients are used as a descriptor [Image from Liu (2012)].

instrumental tracks) in addition to the actual melody, while in the latter the user hums inarticulately, thus attempting to emulate only the melody. As a result, methods for the two tasks differ in their input data: query-by-humming assumes that only a crude version of the melody is given, while query-by-singing also assumes the presence of articulate language phonemes, which can serve as useful surplus information, but can also increase the possibility of errors and misguiding information on the part of the user.

A query-by-humming method based on compressed-domain features is proposed in (Liu & Tsai, 2001). The descriptor is drawn from the subband values. The song is first segmented into musical sentences, and each sentence is further divided in 8 "slots." Each slot is described by a 32-element vector, containing the average energy of each band, which is used for matching on a slot-sequence basis. However, the song relies on a successful segmentation into meaningful sentences, which, in the absence of a reliable algorithm, is performed manually. In a similar task, (Lie & Su, 2004) presents a query-by-singing algorithm. The system is based on a number of statistical descriptors calculated from the MDCT coefficients (centroid, spread, flux, energy distribution, and tone). The system includes a "pure music" detection module, based on these features and Support Vector Machines. The purely musical parts are removed from the database songs prior to the final descriptor formation. Finally, the query item's descriptor is classified through a simple k-means approach followed by time-warping to solve for possible tempo variation.

For singer identification, a method has been proposed in (Liu & Huang, 2002), based on the MDCT coefficients although the authors claim it could be applied to the subband samples as well. The system attempts to isolate phonemes by following the typical Attack-Decay-Sustain-Release pattern occurring in speech and song (i.e. steep energy increase— steep decrease—smooth decrease—steep decrease, Figure 5). Having isolated the phonemes, a per-phoneme descriptor is formed, which is the sum of MDCT coefficients per band, for all phoneme frames. The phonemes of the query song are then classified individually, and the final result is calculated by voting.

In the case of genre classification, three compressed-domain approaches have been proposed in the past. Study (Pye, 2000), and more recently (Chang, Yu, Wan, & Yao, 2008), attempt to form the equivalent of Mel-Frequency Cepstral Coefficients from the subband data instead of forming them from a STFT. Finally, in (Rizzi, Buccino, Panella, & Uncini, 2008), a wide range of timbral, energy and rhythm features are extracted from the subbands. In order to tackle the fundamental drawback of the polyphase filter decimation, that is, the low frequency resolution, the lowest subband is wavelet-transformed, thus giving two sub-subbands. Timbral, energy and rhythm features are all computed from the subbands, making this the most complete attempt at extracting descriptors from the compressed-domain information. Feature selection is performed through Genetic Algorithms, and finally genre classification is achieved through the use of the <Min-Max, PARC> algorithm.

3.4. Beat tracking/Tempo Induction

Beat Tracking aims at detecting the actual beat attack times in a song. It is commonly a first step in the process of Tempo Induction, which aims primarily at calculating the period, that is, the time interval between two successive beats, at the quarter-note level, either for the entire song, or a designated section. The task

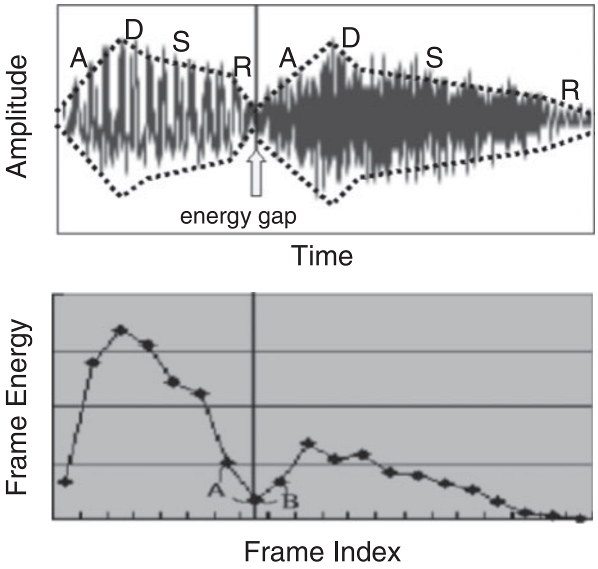


Figure 5. Two consecutive phonemes and their ADSR pattern. Top: the waveform of the two phonemes. Down: the corresponding Frame Energy, as derived from the MDCT coefficients [Image from Liu and Huang (2002)].

Downloaded by [95.92.242.12] at 02:23 05 November 2015

difficulty varies greatly depending upon a number of factors, including the presence of tempo changes, the complexity of the rhythm, and the presence or absence of drums (which is often related to the genre of the song being analyzed).

The oldest approach to estimate tempo from compressed-domain information is (Wang & Vilermo, 2001). It a fast approach based on the MP3 window size patterns and MDCT coefficients. The window size information is by definition directly related to the beat attack times: short windows are employed when the sound signal changes abruptly and increased time resolution is required. The authors note that, around beat times, a specific window size pattern appears: Long => Long-to-Short => Short => Short-to-Long => Long (Figure 6). The presence of this pattern (sometimes expressed as 01230, from the corresponding window size codes) at regular intervals can signify the presence of beats, with high temporal accuracy (constrained, of course, by the window size). In order to improve detection performance, an MDCT-based feature is used as well. The MDCT energy coefficients are summed in a small number of bands, and local energy maxima in time are detected. Following a simple statistical processing of the Inter-Onset Intervals (IOI) thus calculated, an estimate is returned for the song's tempo period.

The MDCT beat tracking subsystem of (Wang & Vilermo, 2001) drew little further attention. The window size patterns, on the other hand, have proven to be a very popular ultra-fast feature for beat tracking. Two recent approaches (D'Aguanno, Haus, & Vercellesi, 2006; D'Aguanno & Vercellesi, 2007) explore

Downloaded by [95.92.242.12] at 02:23 05 November 2015

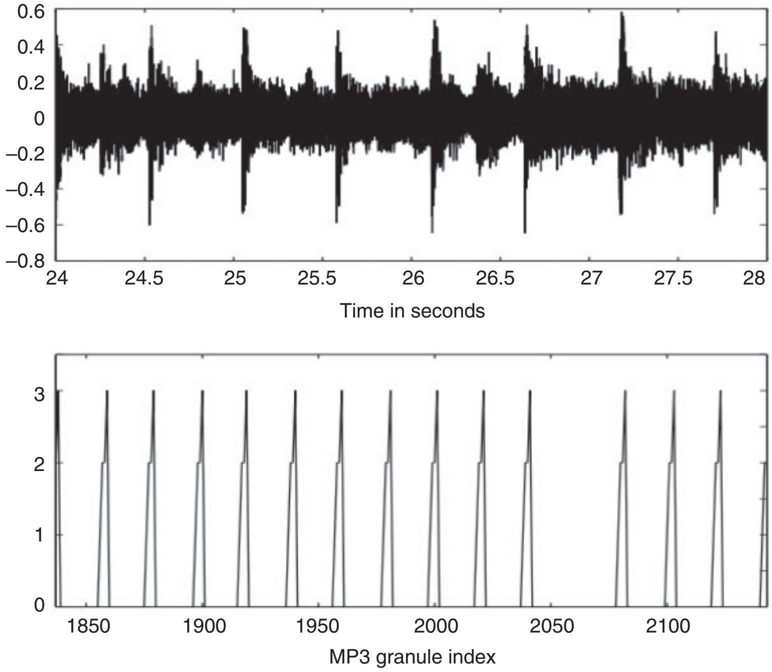


Figure 6. Beat tracking based on MP3 window sizes [Image from Wang and Vilermo (2001)].

the potential of a system based exclusively on window size patterns, and conclude that the information it offers is sufficient for beat tracking, and at high speeds.

An even faster approach is presented in (Zhu & Wang, 2008). The only information extracted from the MP3 compressed files are the Huffman code lengths, without any decoding. The resulting algorithm is shown to combine reliability with high speed.

3.5. Structure analysis and summarization

Another important set of music information retrieval tasks concern the internal structure of music files and its representation. Music summarization deals with the automatic segmentation of a song into its structural elements (verse, bridge, chorus/refrain etc.) and its consecutive representation by some of these elements.

Study (Liu & Yao, 2004) proposes an MPEG-1/2 summarization method based on locating the repeating non-trivial parts, which are expected to correspond to refrains/choruses. Following removal of the non-vocal parts of the song, the energy of each frame is calculated as the sum of squares of all samples in a frame, from all 32 subbands. The sequence of frame energies is then used to detect phonemes using the Attack-Decay-Sustain-Release heuristic, similar to (Liu & Huang, 2002). Sequences of consecutive phonemes can then be grouped in phrases. By clustering phrases, a song can be represented by a small number of clusters. An RP-tree is finally used to model the cluster repetition patterns in the song, and from the tree structure, a summary can be generated. A different, MDCT-based approach is that of (Shao, Xu, Wang, & Kankanhalli, 2004), which attempts the extraction of significantly more complex features. Besides the RMS of the MDCT coefficients, the MDCT spectral centroid is also computed, while additionally an attempt is made to emulate MFCCs from the MDCT coefficients. A per-frame feature vector is formed by the mean and variance of the amplitude envelope, the spectral centroid and 10 cepstral coefficients, which is used to cluster the frames into segments.

4. Computational gains and retrieval performance

The fundamental aim of any method attempting to extract descriptor features directly from the subband domain is the bypassing of two steps of significant computational cost: the decoding of the already-compressed files, and the pre-processing required in order to extract features (e.g. frequency decimation using STFT). On the other hand, compressed-domain algorithms impose specific constraints on the decimation algorithms and parameters we can use, while, when starting with PCM data, we can use custom-tailored feature extraction algorithms. It follows, then, that any gain in computational speed from the use of the compression information might come at a cost in retrieval performance. In this sense, a proposed system ought to be evaluated both in terms of retrieval performance and computational cost compared to state-of-the-art systems designed for PCM data, in order to evaluate the speed/performance tradeoff. Unfortunately, the vast majority of papers surveyed here are lacking in both these

Downloaded by [95.92.242.12] at 02:23 05 November 2015

Downloaded by [95.92.242.12] at 02:23 05 November 2015

aspects. Few papers offer either computational cost or performance comparisons to PCM-based systems.

In terms of speed, a common observation is that the greatest speed gain comes not from the bypassing of the decompression step, but from the ready extraction of subbands without the need for STFT or another decimation method. In (Nakajima et al., 1999) the authors calculate a sixfold increase in feature extraction speed (16.1% of the time required) compared to a PCM-based approach. The same factor of six is also reported in (Pye, 2000)—in this case, the authors actually demonstrate that the bulk of the computational cost for the PCM system is due to the STFT. Since both these methods are based on the subband information, their feature extraction includes decompression up to and including Inverse MDCT. On the other hand, the Huffman code method of (Zhu & Wang, 2008) is compared both to a system based on the MDCT coefficients and a PCM-based system. In this case, the reported speed increase of the beat detection subsystem is 60-fold for MDCT and 800-fold for the Huffman codes, compared to PCM. Finally, (Ravelli, Richard, & Daudet, 2010) reports a sixfold gain in beat tracking, a fivefold increase in genre classification, and a 40% gain for chord recognition. The reason that chord recognition shows a smaller gain is that the machine learning algorithm used in both systems for chord recognition (a Hidden Markov Model trained through the EM algorithm) is particularly costly, thus reducing the significance of the speed difference during feature extraction.

While few, mostly and small-scale, comparative performance evaluations, when given, have been encouraging. In terms of music detection, in (Tzanetakis & Cook, 2000) the subband domain information performs only slightly worse than STFT features, while (Jarina et al., 2004) was submitted in a TRECVid evaluation, and gave better results than some PCM-based algorithms. Study (Pye, 2000) demonstrates equal performance to a PCM-based genre classification system, albeit for a dataset that seems too small by today's standards. Study (Shao et al., 2004) gives comparable results for PCM-based and MP3-based descriptors for music summarization, while (Ravelli et al., 2010) reports comparable performance between MP3, AAC, and PCM-based algorithms for beat-tracking and genre classification, but reduced performance for the compressed-domain descriptors in chord recognition.

5. Conclusions, current trends and future directions

In over two decades of music information retrieval research, almost every popular MIR task has been approached, at least once, with the use of compressed-domain features. Even the tasks not considered in this paper (such as, for example, mood classification) can be treated as variants of the work presented here, and very similar compressed-domain features could be used to tackle them.

The issue of descriptor features, in fact, is what differentiates the methods presented here from any other MIR method in literature: compressed domain data, at first glance, restrict our options compared to PCM audio. However, many of the methods presented in this paper attempt to emulate, quite successfully, STFT-based descriptors with MDCT-based ones. This is a major observation we can draw after reviewing the literature presented here: there are particularly strong similarities between the MDCT coefficient spectrum and the STFT spectrogram which—as demonstrated by the large number of methods successfully replacing STFT-based descriptors with MDCT-based ones—are far from diminished by the MDCT's relative inflexibility.

In this sense, to the extent in which the MDCT can prove to be a suitable substitute for the STFT, compressed-domain information can, in fact, be said to offer extra description options, in the form of Huffman bits, window sizes and scale factors, all readily available at a very low computational cost. The application of such features, as the approaches presented here demonstrate, and their combination with the popular MDCT-based ones, can lead to powerful, efficient solutions to many open MIR problems. What, then, the field appears to be lacking, in order to move forward and establish its contribution to music information retrieval, is a common, open framework for a systematic and rigorous comparative evaluation between compressed-domain features and PCM-based features, that would demonstrate the relative strengths of each.

Perhaps the most complete large-scale attempt to study and evaluate compressed- domain music retrieval is a recent one found in (Ravelli et al., 2010). While the aim of the paper is the demonstration of a novel compression algorithm, referred to as 8 × MDCT, a set of comprehensive evaluations are run on a database over three different music information retrieval tasks (beat tracking, chord recognition, and genre classification) for four different file encodings (PCM, MP3, AAC and 8 × MDCT). A set of mid-level features are extracted from the MDCT coefficients for the three compression formats and from the Short-Time Fourier Transform coefficients for PCM. Thus, an onset-detection function is calculated for beat- tracking, a chromagram for chord recognition and MFCCs, organized in texture windows, for genre classification.

The reported experimental results demonstrate near-excellent performance on all encodings for beat-tracking and genre classification, but significantly reduced performance for chord recognition, both for MP3 and AAC, compared to PCM and 8 × MDCT. The purpose of the experiments is to demonstrate the superiority of 8 × MDCT for the task. However, in a world where most existing audio databases are stored in MP3 or AAC, our interest focuses on the potential of the mainstream formats to give powerful descriptors.

The authors of (Ravelli et al., 2010) attribute the reduced performance of MP3 and AAC in chord recognition to the low frequency resolution of the MDCT. However, the subband and MDCT resolution do not have to be restrictive. On the one hand, the subband values can be further analyzed, such as in (Rizzi et al., 2008), where a wavelet transform is used to split a subband in two. On the other hand, there is significant research toward conversion of subband representations into different decimations. In one recent such work (Schuller, Gruhne, & Friedrich, 2011), which is directly related to our field of interest, a fast algorithm for decimation conversion is presented, and used to extract MPEG-7 descriptors from MP3 or AAC audio files. While the aim in (Schuller et al., 2011) is not specifically music-related, its application could easily be extended to MIR. Were such a line of research to prove fruitful, compressed-domain MIR algorithms

Downloaded by [95.92.242.12] at 02:23 05 November 2015

would have overcome their most significant limitation: the inflexibility resulting from the fact that the frequency analysis parameters are embedded in each encoding.

Even if, however, compressed-domain algorithms ultimately prove unable to achieve satisfactory performance in pitch recognition tasks, there exists a multitude of other MIR challenges where could prove invaluable. Only a small range of attempts has been made thus far and the field remains, to a large extent, unexplored.

Note

[1] While the term “subband” is occasionally used to refer to the MDCT coefficients as well, in this paper we will use the term to refer exclusively to the output of the polyphase filter bank, unless explicitly stated otherwise.

References

Brandenburg, K. (1999). MP3 and AAC explained. In *Proceedings of the Audio Engineering Society 17th International Conference on High Quality Audio Coding* (pp. 139–146). Signa: AES.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, *96*(4), 668–696. doi:10.1109/JPROC.2008.916370

Chang, L., Yu, X., Wan, W., & Yao, J. (2008). Research on fast music classification based on SVM in compressed domain. In *International Conference on Audio, Language and Image Processing (ICALIP08)* (pp. 638–642). Shanghai: IEEE.

D’Aguanno, A., Haus, G., & Vercellesi, G. (2006). MP3 window-switching pattern analysis for general purposes beat tracking on music with drums. In *The 2006 audio engineering society convention* (pp. 20–23). Paris: AES.

D’Aguanno, A., & Vercellesi, G. (2007). Tempo induction algorithm in MP3 compressed domain. In *The international workshop on multimedia information retrieval* (pp. 153–158). Augsburg: ACM.

Grachten, M., Schedl, M., Pohle, T., & Widmer, G. (2009). The ISMIR cloud: A decade of ISMIR conferences at your fingertips. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR09)* (pp. 63–68). Kobe: ISMIR.

ISO/IEC 11172-3. (1993). Information technology – Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s — Part 3: Audio.

ISO/IEC 13818-7. (1997). Information technology – Generic coding of moving pictures and associated audio information – Part 7: Advanced Audio Coding (AAC).

ISO/IEC 14496-3. (2005). Information technology – Coding of audio-visual objects – Part 3: Audio.

Jarina, R., O’Connor, N. E., Murphy, N., & S. Marlow, (2004). An experiment in audio classification from compressed data, *International Workshop on Systems, Signals and Image Processing, Ambient Multimedia*, *11*, 307–310.

Jiao, Y., Yang, B., Li, M., & Niu, X. (2007). MDCT-based perceptual hashing for compressed audio content identification, In *IEEE 9th Workshop on Multimedia Signal Processing (MMSP07)* (pp. 381–384). Chania: IEEE.

Khotanzad, A., & Hong, Y. H. (1990). Invariant image recognition by Zernike moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(5), 489–497. doi:10.1109/34.55109

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson P., Scott, J., … Turnbull, D. (2010). Music emotion recognition: A state of the art review, In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR10)* (pp. 255–266). Utrecht: ISMIR.

Kiranyaz, S., Farooq Qureshi, A., & Gabbouj, M. (2006). A generic audio classification and segmentation approach for multimedia indexing and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(3), 1062–1081. doi:10.1109/TSA.2005.857573

Kotsifakos, A., Papapetrou, P., Hollmen, J., Gunopulos, D., & Athitsos, V. (2012). A survey of query-by-humming similarity methods, In *Proceedings, conference on Pervasive Technologies Related to Assistive Environments (PETRA)* (p. 5). Heraklion, Crete: ACM.

Downloaded by [95.92.242.12] at 02:23 05 November 2015

Li, W., Liu, Y., & Xue, X. (2010a). Robust audio identification for MP3 popular music. In *Proceedings of the ACM 33rd International SIGIR Conference on Research and Development in Information Retrieval* (pp. 627–634). Geneva: ACM.

Li, W. Y., Liu, Y., & Xue, X. (2010b). Robust music identification based on low-order Zernike moment in the compressed domain. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 739–740). Geneva: ACM.

Lie, W.-N., & Su, C.-K. (2004). Content-based retrieval of MP3 songs based on query by singing. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP04)* (vol. 5, pp. 929–932), Montreal: IEEE.

Liu, C.-C. (2012). MP3 sniffer: A system for online detecting MP3 music transmissions. In *The 10th International Conference on Advances in Mobile Computing and Multimedia (MoMM12)* (pp. 93–96). Bali: ACM.

Liu, C.-C., & Chang, P. F. (2011). An efficient audio fingerprint design for MP3 music. In *9th ACM International Conference on Advances in Mobile Computing and Multimedia (MoMM)* (pp. 190–193). Ho Chi Minh City: ACM.

Liu, C.-C., & Huang, C.-S. (2002). A singer identification technique for content-based classification of MP3 music objects. In *Eleventh International Conference on Information and Knowledge Management* (pp. 438–445). McLean, VA: ACM.

Liu, C.-C. & Tsai, P.-J. (2001). Content-based retrieval of MP3 music objects. In *Tenth International Conference on Information and knowledge management (CIKM11)* (pp. 506–511). Atlanta: ACM.

Liu, C.-C., & Yao, P.-C. (2004). Automatic summarization of MP3 music objects. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP04)* (vol. 5, pp. 921–924), Montreal: IEEE.

Mayer, R., & Rauber, A. (2010). Multimodal aspects of music retrieval: Audio, song lyrics and beyond?. *Studies in Computational Intelligence: Advances in Music Information Retrieval*, *274*, 333–363.

Nakajima, Y., Lu, Y., Sugano, M., Yoneyama, A., Yamagihara, H., & Kurematsu, A. (1999). A fast audio classification from MPEG coded data, In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP99)* (vol. 6, pp. 3005–3008). Phoenix: IEEE.

Orio, N. (2006). Music retrieval: A tutorial and review. *Foundations and Trends in Information Retrieval*, *1*(1), pp. 1–96. doi:10.1561/1500000002

Pan, D. (1995). A tutorial on MPEG/audio compression. *IEEE Multimedia Magazine*, *2*(2), pp. 60–74. doi:10.1109/93.388209

Patel, N. V., & Sethi, I. K. (1996). Audio characterization for video indexing. In *Storage and retrieval for still image and video databases* (vol. 2670, pp. 373–384). San Diego/La Jolla, CA: SPIE.

Pfeiffer, S., & Vincent, T. (2001). Formalisation of MPEG-1 compressed domain audio features, *CSIRO Mathematical and Information Sciences, Australia, Technical Report*, *1*(196), pp. 1–18.

Pye, D. (2000). Content-based methods for the management of digital music. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP00)* (pp. 2437–2440). Istanbul: IEEE.

Ravelli, E., Richard, G., & Daudet, L. (2010). Audio signal representations for indexing in the transform domain. IEEE Transactions on Audio, Speech, *and Language Processing*, *18*(3), pp. 434–446. doi:10.1109/TASL.2009.2025099

Rizzi, A., Buccino, M., Panella, M., & Uncini, A. (2006). Optimal short-time features for music/speech classification of compressed audio data. In *Proceedings, International Conference on Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce* (p. 210). Sydney: IEEE.

Rizzi, A., Buccino, N. M., Panella, M. & Uncini, A. (2008). Genre classification of compressed audio data. In *IEEE 10th Workshop on Multimedia Signal Processing* (pp. 654–659). Queensland, Cairns: IEEE.

Schuller, G., Gruhne, M., & Friedrich, T. (2011). Fast audio feature extraction from compressed audio data. *IEEE Journal of Selected Topics in Signal Processing*, *5*(6), 1262–1271. doi:10.1109/JSTSP.2011.2158802

Shao, X., Xu, C., Wang, Y., & Kankanhalli, M. S. (2004). Automatic music summarization in compressed domain. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP04)* (vol. 4, pp. iv–261). Montreal: IEEE.

Shieh, J.-R. J. (2003). Audio content based feature extraction on subband domain. In *37th Annual International Carnahan Conference on Security Technology* (pp. 552–555). Taipei: IEEE.

Stober, S., & Nürnberger, A. (2013). Adaptive music retrieval – a state of the art, *Multimedia Tools and Applications*, *65*(3), 467–494. doi:10.1007/s11042-012-1042-z

Tsai, T.-H., & Chang, W.-C. (2009). Two-stage method for specific audio retrieval based on MP3 compression domain. In *IEEE International Symposium on Circuits and Systems (ISCAS09)* (pp. 713–716). Taipei: IEEE.

Downloaded by [95.92.242.12] at 02:23 05 November 2015

Tsai, T.-H., & Wang, Y.-T. (2004). Content-based retrieval of audio example on MP3 compression domain. In *IEEE 6th Workshop on Multimedia Signal Processing* (pp. 123–126). Siena: IEEE.

Tsai, T.-H., Wang, Y.-T., Hung, J. H., & Wey, C.-L. (2006). Compressed domain content-based retrieval of mp3 audio example using quantization tree indexing and melody-line tracking method. In *IEEE International Symposium on Circuits and Systems (ISCAS06)* (p. 4). Kos: IEEE.

Tzanetakis, G., & Cook, F. (2000). Sound analysis using MPEG compressed audio. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP00)* (vol. 2, pp. II761–II764). Istanbul: IEEE.

Wang, H., Divakaran, A., Vetro, A., Chang, S.-F., & Sun, H. (2003). Survey of compressed-domain features used in audio-visual indexing and analysis. *Journal of Visual Communication and Image Representation*, *14*(2), 150–183. doi:10.1016/S1047-3203(03)00019-1

Wang, Y., & Vilermo, M. (2001). A compressed domain beat detector using MP3 audio bitstreams. In *The Ninth ACM International Conference on Multimedia* (pp. 194–202). Ottawa, ON: ACM.

Yang, Y.-H. & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(3), 40.

Zhu, J., & Wang, Y. (2008). Complexity-scalable beat detection with MP3 audio bitstreams. *Computer Music Journal*, *32*(1), 71–87. doi:10.1162/comj.2008.32.1.71

Downloaded by [95.92.242.12] at 02:23 05 November 2015

View publication stats