Audio Fingerprinting: Concepts And Applications

Chapter · September 2005

DOI: 10.1007/10966518_17

5 authors, including:

Pedro Cano

47 PUBLICATIONS **1,599** CITATIONS

SEE PROFILE

Eloi Batlle

University Pompeu Fabra

30 PUBLICATIONS **943** CITATIONS

SEE PROFILE

Emilia Gómez

University Pompeu Fabra

164 PUBLICATIONS **3,232** CITATIONS

SEE PROFILE

Leandro de Campos Teixeira Gomes

University of Campinas

11 PUBLICATIONS **63** CITATIONS

SEE PROFILE

AUDIO FINGERPRINTING: CONCEPTS AND APPLICATIONS

Pedro Cano, Eloi Batlle, Emilia Gómez

MTG, Universitat Pompeu Fabra
Pg. Circumval.lació 8, 08003
Barcelona, Spain
{pcano, eloi, egomez}@iua.upf.es

Leandro de C.T. Gomes, Madeleine Bonnet

InfoCom-Crip5, Université René Descartes
45, rue des Saints-Pères, 75270
Paris cedex 06, France
{tgomes,bonnet}@math-info.univ-paris5.fr

ABSTRACT

An audio fingerprint is a compact digest derived from perceptually relevant aspects of a recording. Fingerprinting technologies allow the monitoring of audio content without the need of meta-data or watermark embedding. However, additional uses exist for audio fingerprinting and some are reviewed in this article.

1. INTRODUCTION

This paper aims to give a vision on *Audio Fingerprinting*. The rationale is presented along with the differences with respect to watermarking. The main requirements of fingerprinting systems are described. The basic modes of employing audio fingerprints, namely identification, authentication, content-based secret key generation for watermarking and content-based audio retrieval and processing are depicted. Some concrete scenarios and business models where the technology is used are presented as well as an example of an audio fingerprinting extraction algorithm which has been proposed for both identification and verification.

2. DEFINITION OF AUDIO FINGERPRINTING

An audio fingerprint is a content-based compact signature that summarizes an audio recording. Audio fingerprinting has attracted a lot of attention for its audio monitoring capabilities. Audio fingerprinting or content-based identification (CBID) technologies extract acoustic relevant characteristics of a piece of audio content and store them in a database. When presented with an unidentified piece of audio content, characteristics of that piece are calculated and matched against those stored in the database. Using fingerprints and matching algorithms, different versions of a single recording can be identified as the same music title [1].

The approach differs from an alternative existing solution to monitor audio content: Audio Watermarking. In audio watermarking [2], research on psychoacoustics is conducted so that an arbitrary message, the watermark, can be embedded in a recording without altering the perception of the sound. Compliant devices can check for the presence of the watermark before proceeding to operations that could result in copyright infringement. In audio fingerprinting, the message is automatically derived from the perceptually most relevant components of sound. Compared to watermarking, it is ideally less vulnerable to attacks and distortions since trying to modify this message, the fingerprint, means alteration of the quality of the sound. It is also suitable to deal with legacy content, that is, with audio material released without watermark. In addition, it requires no modification of the audio content. As a drawback, the complexity of fingerprinting is higher than watermarking and there is the need of a connection to a fingerprint repository. In addition, contrary to watermarking, the message is not independent from the content. It is therefore for example not possible to distinguish between perceptually identical copies of a recording. Just like with watermarking technology, there are more uses to fingerprinting than identification. Specifically, it can also be used for verification of content-integrity; similarly to fragile watermarks.

At this point, we should clarify that the term "fingerprinting" has been employed for many years as a special case of watermarking devised to keep track of an audio clip's usage history. Watermark fingerprinting consists in uniquely watermarking each legal copy of a recording. This allows to trace back to the individual who acquired it [3]. However, the same term has been used to name techniques that associate an audio signal to a much shorter numeric sequence (the "fingerprint") and use this sequence to e.g. identify the audio signal. The latter is the meaning of the term "fingerprinting" in this article. Other terms for audio fingerprinting are Robust Matching, Robust or Perceptual Hashing, Passive Watermarking, Automatic Music Recognition, Content-based Digital Signatures and Content-based Audio Identification. The areas relevant to audio fingerprinting include Information Retrieval, Pattern Matching, Signal Processing, Cryptography and Music Cognition to name a few.

3. PROPERTIES OF AUDIO FINGERPRINTING

The requirements depend heavily on the application but are useful in order to evaluate and compare different audio fingerprinting technologies. In their *Request for Information on Audio Fingerprinting Technologies* [1], the IFPI (International Federation of the Phonographic Industry) and the RIAA (Recording Industry Association of America) tried to evaluate several identification systems. Such systems have to be computationally efficient and robust. A more detailed enumeration of requirements can help to distinguish among the different approaches [4][5]:

Accuracy: The number of correct identifications, missed identifications, and wrong identifications (false positives).

Reliability: Methods for assessing that a query is present or not in the repository of items to identify is of major importance in play list generation for copyright enforcement organizations. In such cases, if a song has not been broadcast, it should not be identified as a match, even at the cost of missing actual matches. Approaches to deal with false positives have been treated for instance in [6]. In other applications,

like automatic labeling of MP3 files (see the applications section), avoiding false positives is not such a mandatory requirement.

Robustness: Ability to accurately identify an item, regardless of the level of compression and distortion or interference in the transmission channel. Ability to identify whole titles from excerpts a few seconds long (known as cropping or granularity), which requires methods for dealing with lack of synchronization. Other sources of degradation are pitching, equalization, background noise, D/A-A/D conversion, audio coders (such as GSM and MP3), etc.

Security: Vulnerability of the solution to cracking or tampering. In contrast with the robustness requirement, the manipulations to deal with are designed to fool the fingerprint identification algorithm.

Versatility: Ability to identify audio regardless of the audio format. Ability to use the same database for different applications.

Scalability: Performance with very large databases of titles or a large number of concurrent identifications. This affects the accuracy and the complexity of the system.

Complexity: It refers to the computational costs of the fingerprint extraction, the size of the fingerprint, the complexity of the search, the complexity of the fingerprint comparison, the cost of adding new items to the database, etc.

Fragility: Some applications, such as content-integrity verification systems, may require the detection of changes in the content. This is contrary to the robustness requirement, as the fingerprint should be robust to content-preserving transformations but not to other distortions (see the integrity verification section).

Improving a certain requirement often implies losing performance in some other. Generally, the fingerprint should be:

- A perceptual digest of the recording. The fingerprint must retain the maximum of acoustically relevant information. This digest should allow the discrimination over a large number of fingerprints. This may be conflicting with other requirements, such as complexity and robustness.
- Invariant to distortions. This derives from the robustness requirement. Content-integrity applications, however, relax this constraint for content-preserving distortions in order to detect deliberate manipulations.
- Compact. A small-sized representation is interesting for complexity, since a large number (maybe millions) of fingerprints need to be stored and compared. An excessively short representation, however, might not be sufficient to discriminate among recordings, affecting thus accuracy, reliability and robustness.
- Easily computable. For complexity reasons, the extraction of the fingerprint should not be excessively time-consuming.

4. USAGE MODES

4.1. Identification

Independently of the specific approach to extract the content-based compact signature, a common architecture can be devised to describe the functionality of fingerprinting when used for identification [1].

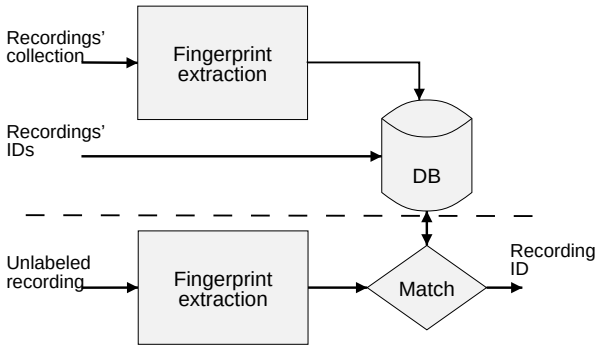

Figure 1: Content-based audio identification framework

The overall functionality mimics the way humans perform the task. As seen in Figure 1, a memory of the works to be recognized is created off-line (top); in the identification mode (bottom), unlabeled audio is presented to the system to look for a match.

Database creation: The collection of works to be recognized is presented to the system for the extraction of their fingerprint. The fingerprints are stored in a database and can be linked to a tag or other meta-data relevant to each recording.

Identification: The unlabeled audio is processed in order to extract the fingerprint. The fingerprint is then compared with the fingerprints in the database. If a match is found, the tag associated with the work is obtained from the database. A reliability measure of the match can also be provided.

4.2. Integrity verification

Integrity verification aims at detecting the alteration of data. The overall functionality (see Figure 2) is similar to identification. First, a fingerprint is extracted from the original audio. In the verification phase, the fingerprint extracted from the test signal is compared with the fingerprint of the original. As a result, a report indicating whether the signal has been manipulated is output. Optionally, the system can indicate the type of manipulation and where in the audio it occurred. The verification data, which should be significantly smaller than the audio data, can be sent along with the original audio data (e.g. as a header) or stored in a database. A technique known as *self-embedding* avoids the need of a database or a specially dedicated header, by embedding the content-based signature into the audio data using watermarking (see Figure 3). An example of such a system is described in [7].

4.3. Watermarking support

Audio fingerprinting can assist watermarking. Audio fingerprints can be used to derive secret keys from the actual content. As described by Mihçak el al. [8], using the same secret key for a number of different audio items may compromise security, since each item may leak partial information about the key. Perceptual hashing can help generate input-dependent keys for each piece of audio. Haitsma et al. [9] suggest audio fingerprinting to enhance the security of watermarks in the context of copy attacks. Copy attacks estimate a watermark from watermarked content and transplant it to unmarked content. Binding the watermark to the content can help to defeat this type of attacks. In addition, fingerprinting can

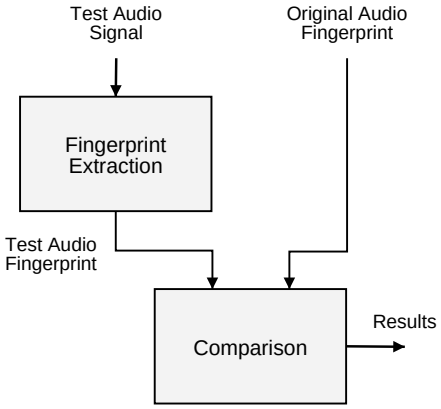

Figure 2: Integrity verification framework

be useful against insertion/deletion attacks that cause desynchronization of the watermark detection: by using the fingerprint, the detector is able to find anchor points in the audio stream and thus to resynchronize at these locations [8].

4.4. Content-based audio retrieval and processing

Deriving compact signatures from complex multimedia objects is an essential step in Multimedia Information Retrieval. Fingerprinting can extract information from the audio signal at different abstraction levels, from low level descriptors to higher level descriptors. Especially, higher level abstractions for modeling audio hold the possibility to extend the fingerprinting usage modes to content-based navigation, search by similarity, content-based processing and other applications of Music Information Retrieval. In a query-by-example scheme, the fingerprint of the song can be used to retrieve not only the original version but also "similar" ones [10].

5. APPLICATION SCENARIOS

Most of the applications presented in this section are particular cases of the identification usage mode described above. They are therefore based on the ability of audio fingerprinting to link unlabeled audio to corresponding metadata, regardless of audio format.

5.1. Audio Content Monitoring and Tracking

5.1.1. Monitoring at the distributor end

Content distributors may need to know whether they have the rights to broadcast the content to consumers. Fingerprinting can help identify unlabeled audio in TV and Radio channels repositories. It can also identify unidentified audio content recovered from CD plants and distributors in anti-piracy investigations (e.g. screening of master recordings at CD manufacturing plants) [1].

5.1.2. Monitoring at the transmission channel

In many countries, radio stations must pay royalties for the music they air. Rights holders need to monitor radio transmissions in order to verify whether royalties are being properly paid. Even in countries where radio stations can freely air music, rights holders are interested in monitoring radio transmissions for statistical purposes. Advertisers also need to monitor radio and TV transmissions to verify whether commercials are being broadcast as agreed. The same is true for web broadcasts. Other uses include chart compilations for statistical analysis of program material or enforcement of "cultural laws" (e.g. French titles in France). Fingerprinting-based monitoring systems are being used for this purpose. The system "listens" to the radio and continuously updates a play list of songs or commercials broadcast by each station. Of course, a database containing fingerprints of all songs and commercials to be identified must be available to the system, and this database must be updated as new songs come out. Examples of commercial providers of this service are: Broadcast Data System (www.bdsonline.com), Music Reporter (www.musicreporter.net), Audible Magic (www.audiblemagic.com).

Napster and Web-based communities alike, where users share music files, have been excellent channels for music piracy. After a court battle with the recording industry, Napster was enjoined from facilitating the transfer of copyrighted music. The first measure taken to conform with the judicial ruling was the introduction of a filtering system based on file-name analysis, according to lists of copyrighted music recordings supplied by the recording companies. This simple system did not solve the problem, as users proved to be extremely creative in choosing file names that deceived the filtering system while still allowing other users to easily recognize specific recordings. The large number of songs with identical titles was an additional factor in reducing the efficiency of such filters. Fingerprinting-based monitoring systems constitute a well-suited solution to this problem. Napster actually adopted a fingerprinting technology (see www.relatable.com) and a new file-filtering system relying on it. Additionally, audio content can be found in ordinary web pages. Audio fingerprinting combined with a web crawler can identify this content and report it to the corresponding right owners (e.g. www.baytsp.com).

5.1.3. Monitoring at the consumer end

In usage-policy monitoring applications, the goal is to avoid misuse of audio signals by the consumer. We can conceive a system where a piece of music is identified by means of a fingerprint and a database is contacted to retrieve information about the rights. This information dictates the behavior of compliant devices (e.g. CD and DVD players and recorders, MP3 players or even computers) in accordance with the usage policy. Compliant devices are required to be connected to a network in order to access the database.

5.2. Added-value services

Content information is defined as information about an audio excerpt that is relevant to the user or necessary for the intended application. Depending on the application and the user profile, several levels of content information can be defined. Here are some of the situations we can imagine:

- Content information describing an audio excerpt, such as rhythmic, timbrical, melodic or harmonic descriptions.
- Meta-data describing a musical work, how it was composed and how it was recorded. For example: composer, year of composition, performer, date of performance, studio recording/live performance.
- Other information concerning a musical work, such as album cover image, album price, artist biography, information on the next concerts, etc.

Different user profiles can be defined. Common users would be interested in general information about a musical work, such as title, composer, label and year of edition; musicians might want to know which instruments were played, while sound engineers could be interested in information about the recording process. Content information can be structured by means of a music description scheme (MusicDS), which is a structure of meta-data used to describe and annotate audio data. The MPEG-7 standard proposes a description scheme for multimedia content based on the XML metalanguage [12], providing for easy data interchange between different equipments.

Some systems store content information in a database that is accessible through the Internet. Fingerprinting can then be used to identify a recording and retrieve the corresponding content information, regardless of support type, file format or any other particularity of the audio data. For example, MusicBrainz, Id3man or Moodlogic (www.musicbrainz.org, www.id3man.com, www.moodlogic.com) automatically label collections of audio files; the user can download a compatible player that extracts fingerprints and submits them to a central server from which meta data associated to the recordings is downloaded. Another example is the identification of a tune through mobile devices, e.g. a cell phone; this is one of the most demanding situations in terms of robustness, as the audio signal goes through radio distortion, D/A-A/D conversion, background noise and GSM coding, and only a few seconds of audio are available.

5.3. Integrity verification systems

In some applications, the integrity of audio recordings must be established before the signal can actually be used, i.e. one must assure that the recording has not been modified or that it is not too distorted. If the signal undergoes lossy compression, D/A-A/D conversion or other content-preserving transformations in the transmission channel, integrity cannot be checked by means of standard hash functions, since a single bit flip is sufficient for the output of the hash function to change. Methods based on fragile watermarking can also provide false alarms in such a context. Systems based on audio fingerprinting, sometimes combined with watermarking, are being researched to tackle this issue. Among some possible applications [7], we can name: Check that commercials are broadcast with the required length and quality, verify that a suspected infringing recording is in fact the same as a recording whose ownership is known, etc.

6. AUDIO FINGERPRINTING SYSTEM: AN EXAMPLE

The actual implementations of audio fingerprinting usually follow the presented scheme, with differences on the acoustic features observed and the modeling of audio, as well as the matching and searching algorithms. The simplest approach would be direct file comparison. A way to efficiently implement this idea consists in using a hash method, such as MD5 (Message Digest 5) or CRC (Cyclic Redundancy Checking), to obtain a compact representation of the binary file. In this setup, one compares the compact signatures instead of the whole files. Of course, this approach is not robust to compression or distortions of any kind, and might not even be considered as content-based identification of audio, since it is not based on an analysis of the content (understood as perceptual information) but just on manipulations performed on binary data. This approach would not be appropriate for monitoring streaming audio or analog audio; however, it sets the basis for a class of fingerprinting methods: Robust or Perceptual hashing [8][9]. The idea behind Robust hashing is the incorporation of acoustic features in the hash function, so that the final hash code is robust to audio manipulations as long as the content is preserved. Many features for audio characterization can be found in the literature, such as energy, loudness, spectral centroid, zero crossing rate, pitch, harmonicity, spectral flatness [11] and Mel-Frequency Cepstral Coefficients (MFCC's). Several methods perform a filter bank analysis, apply a transformation to the feature vector and, in order to reduce the representation size, extract some statistics: means or variances over the whole recording, or a codebook for each song by means of unsupervised clustering. Other methods apply higher-level algorithms that try to go beyond signal processing comparisons and make use of notions such as beat and harmonics [13].

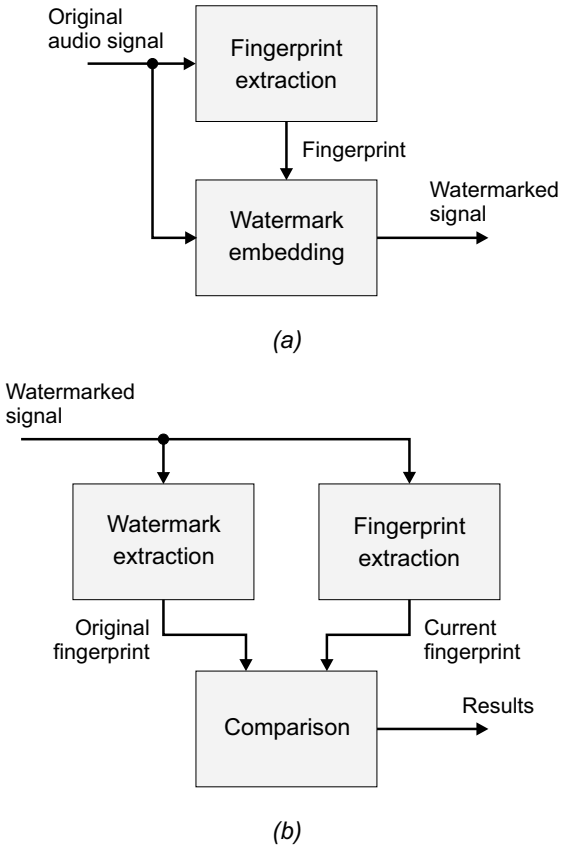

Figure 3: Self-embedding integrity verification framework: (a) fingerprint embedding and (b) fingerprint comparison.

A case study to illustrate in more detail an implementation of an audio fingerprinting solution is presented. The implementation was designed with high robustness requirements: identification of radio broadcast songs [14], and has been tested for content integrity verification [7]. The inherent difficulty in the task of identifying broadcast audio is mainly due to differences between the original titles (as available on CDs) and the broadcast ones: a song may be partially transmitted, the speaker may talk on top

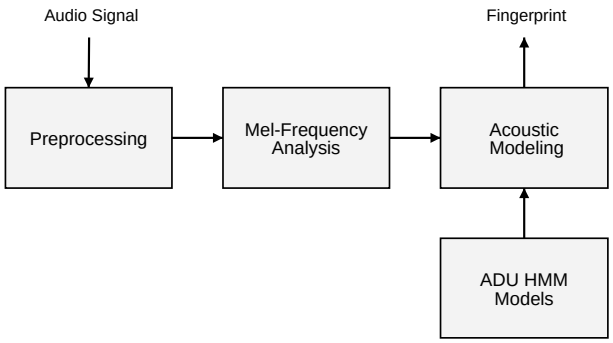

Figure 4: Fingerprint extraction case study

of different segments of the song, the piece may be played faster and several manipulation effects may be applied to increase the listener's psychoacoustic impact (compressors, enhancers, equalization, bass-booster, etc.).

Yet the system also has to be fast because it must do comparisons with several thousand (of the order of 100,000's) songs on-line. This affects memory and computation requisites, since the system should observe several radio stations, give results on-line and should not be very expensive in terms of hardware. In this scenario, a particular abstraction of audio to be used as robust fingerprint is presented: audio as a sequence of basic sounds. The whole identification system works as follows. An alphabet of sounds that best describe the music is extracted in an off-line process out of a collection of music representative of the type of songs to be identified. These audio units are modeled with Hidden Markov Models (HMM). The unlabeled audio and the set of songs are decomposed into these audio units, ending up with a sequence of symbols for the unlabeled audio and a database of sequences representing the original songs. By approximate string matching, the song sequence that best resembles the sequence of the unlabeled audio is obtained.

The approach to learn the relevant acoustic events, the Audio Descriptor Units (ADU), is performed with unsupervised training, that is, without any previous knowledge of music events, through a modified Baum-Welch algorithm. The audio data is pre-processed by a front-end in a frame-by-frame analysis. Then a set of relevant feature(s) vectors is extracted from the sound. In the acoustic modeling block, the feature vectors are run against the statistical models of the ADU: HMM-ADU using the Viterbi algorithm. As a result, the most likely ADU sequence is produced. The fingerprint consists then in a sequence of symbols with their associate start and end times. The average output is around 20 symbols per second out of a vocabulary of 32 ADUs.

7. SUMMARY

We have presented an introduction to concepts related to Audio Fingerprinting along with some possible usage scenarios and application contexts. We have reviewed the requirements desired in a fingerprinting scheme acknowledging the existence of a trade-off between them. Most applications profit the ability to link content to unlabeled audio but there are more uses. The list of applications provided is necessarily incomplete since many uses are likely to come up in the near future.

8. REFERENCES

[1] Request for Information on Audio Fingerprinting Technologies. http://www.riaa.org/pdf/RIAA_IFPI_Fingerprinting_RFI.pdf. 2001.

[2] L. Boney, A. Tewfik, and K. Hamdy, "Digital Watermarks for Audio Signals," *IEEE Proceedings Multimedia*, 473-480, 1996.

[3] S.A. Craver, Wu. M, and B. Liu, "What Can We Reasonably Expect from Watermarks?," *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY, Oct. 2001.

[4] Audio Identification Technology Overview. http://www.audiblemagic.com/about. 2001.

[5] T. Kalker, "Applications and Challenges for Audio Fingerprinting", presentation at the *111th AES Convention*, New York, 2001.

[6] P. Cano, M. Kaltenbrunner, O. Mayor, and E. Batlle, "Statistical Significance in Song-Spotting in Audio," *Proceedings of the International Symposium on Music Information Retrieval*, Bloommington, IN, Oct. 2001.

[7] E. Gómez, P. Cano, L. de C.T. Gomes, E. Batlle, and M. Bonnet, "Mixed Watermarking-Fingerprinting Approach for Integrity Verification of Audio Recordings," *Proceedings of the International Telecommunications Symposium*, Natal, Brazil, Sept. 2002.

[8] M.K. Mihçak and R. Venkatesan, "A Perceptual Audio Hashing Algorithm: a Tool for Robust Audio Identification and Information Hiding," *Proceedings of the 4th Workshop on Information Hiding*, Pittsburg, PA, Ap. 2001.

[9] J. Haitsma, T. Kalker, and J. Oostveen, "Robust Audio Hashing for Content Identification," *Proceedings of the International Workshop on Content-Based Multimedia Indexing*, Brescia, Italy, Sept. 2001.

[10] P. Cano, M. Kaltenbrunner, F. Gouyon, and E. Batlle, "On the Use of FastMap for Audio Information Retrieval," *Proceedings of the International Symposium on Music Information Retrieval*, Paris, France, Oct. 2002.

[11] E. Allamanche, J. Herre, O. Helmuth, B. Fröba, T. Kasten, and M. Cremer, "Content-Based Identification of Audio Material Using MPEG-7 Low Level Description," *Proceedings of the International Symposium of Music Information Retrieval*, Bloommington, IN, Oct. 2001.

[12] Moving Picture Experts Group (MPEG), "MPEG Working Documents", http://mpeg.telecomitalialab.com/working_documents.htm. 2002.

[13] T.L. Blum, D.F. Keislar, J.A. Wheaton, and E.H. Wold, "Method and Article of Manufacture for Content-Based Analysis, Storage, Retrieval and Segmentation of Audio Information," U.S. Patent 08/897,662[5,918,223], 1999.

[14] P. Cano, E. Batlle, H. Mayer, and H. Neuschmied, "Robust Sound Modeling for Song Detection in Broadcast Audio," *Proceedings of the Audio Engineering Society*, Munich, Germany, May 2002.