Audio Watermarking and Fingerprinting: For Which Applications?

Leandro de C.T. Gomes[1], Pedro Cano[2], Emilia Gómez[2], Madeleine Bonnet[1], Eloi Batlle[2]

[1] InfoCom-Crip5, Université René Descartes, Paris, France
{tgomes, bonnet}@math-info.univ-paris5.fr, http://www.math-info.univ-paris5.fr/crip5/infocom/

[2] MTG-IUA, Universitat Pompeu Fabra, Barcelona, Spain
{pedro.cano, emilia.gomez, eloi.batlle}@iua.upf.es, http://www.iua.upf.es/mtg/

Although not a new issue, music piracy has acquired a new status in the digital era, as recordings can be easily copied and distributed. Watermarking has been proposed as a solution to this problem. It consists in embedding into the audio signal an inaudible mark containing copyright information. A different approach, called fingerprinting, consists in extracting a "fingerprint" from the audio signal. In association with a database, this fingerprint can be used to identify a recording, which is useful, for example, to monitor audio excerpts played by broadcasters and webcasters. There are far more applications to watermarking and fingerprinting. After a brief technical review, this article describes potential applications of both methodologies, showing which one is more suitable for each application.

1 Introduction

Music copyright enforcement is not a new issue. The recording industry has been fighting piracy since its very early times. However, the digital revolution in audio has brought this fight to a new level, as music in digital format can be copied and distributed easily and with no degradation. Electronic distribution means, particularly the Internet, associated with efficient compression algorithms (such as MP3) and peer-to-peer file-sharing systems (such as Napster (2002) and Gnutella (Gnutella wego, 2002) (Gnutella news, 2002)) create an environment that is propitious to music piracy.

Watermarking has been proposed as a potential solution to this problem. It consists in embedding a mark, the watermark, into the original audio signal. This mark should not degrade audio quality, but it should be detectable and indelible. Compliant devices should check for the presence of a watermark before proceeding to operations that could result in copyright infringement. Research in this field has been very active over the last years. In particular, the Secure Digital Music Initiative consortium (SDMI), which brings together the major actors in the recording and consumer-electronics industries, has recently released technology specifications intended to protect, by means of a watermark, the playing, storing and distribution of music in digital format (SDMI, 2002). This technology was submitted to public evaluation through a "challenge" inviting individuals to defeat SDMI's protection system, a goal that was shortly achieved, showing that the technology was not ready for commercial purposes (Boeuf & Stern, 2001) (Wu et al., 2001) (Craver, 2001).

Another approach to the copyright-protection problem, quite different from watermarking in its conception, consists in analyzing an audio signal and constructing a "fingerprint" that is uniquely associated with this signal[1]. *Automatic music recognition* or *fingerprinting* systems (Recording Industry Association of America, 2001) can identify a song by searching for its fingerprint in a previously constructed database. Such systems are being used, for example, to monitor music transfers in Napster-like file-sharing facilities, blocking transfers of copyrighted material or collecting the corresponding royalties and to track audio content played by broadcasters (as (AudibleMagic, 2001) (Music Reporter, 2001) and (Auditude, 2001)).

There are far more applications to watermarking and fingerprinting than just copyright protection. After a brief technical review, we describe potential applications of both methodologies, showing which approach would be more suitable for each application discussed.

[1] The term "fingerprinting" has been employed for many years as a special case of watermarking (consisting in uniquely watermarking each legal copy of a recording). 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 identify the audio signal (Craver & Liu, 2001). The latter is the meaning of the term "fingerprinting" in this article.

Accepted for publication on May 9th 2002

2 Watermarking

Watermarking printed documents in order to prevent counterfeiting has been common practice for centuries. This kind of watermark generally consists of a translucent drawing that becomes visible when the paper is held to the light. Currency bills, for example, are often watermarked as a proof of genuineness. The same term has been employed to designate the class of methods intended to imperceptibly mark digital documents (particularly images, audio and video).

Watermarking is often described as a subclass of *steganography* — a Greek word meaning ``hidden writing". The goal of cryptography is to render a message unintelligible, whereas steganography attempts to hide the very presence of a message by embedding it into another information. While the exchange of ciphered messages can arouse suspicion, a steganographic message can be hidden in an apparently innocent document (Petitcolas et al.,1999).

Initial research on audio watermarking dates back to the mid-nineties. The first techniques were directly inspired from previous research on image watermarking (Boney et al., 1996). The basic idea consists in adding a signal, the watermark, to the original audio signal. The resulting watermarked signal must be perceived by the listener as identical to the original one. The watermark carries data that can be retrieved by a detector and can be used for a multitude of purposes. This procedure is illustrated in Figure 1.

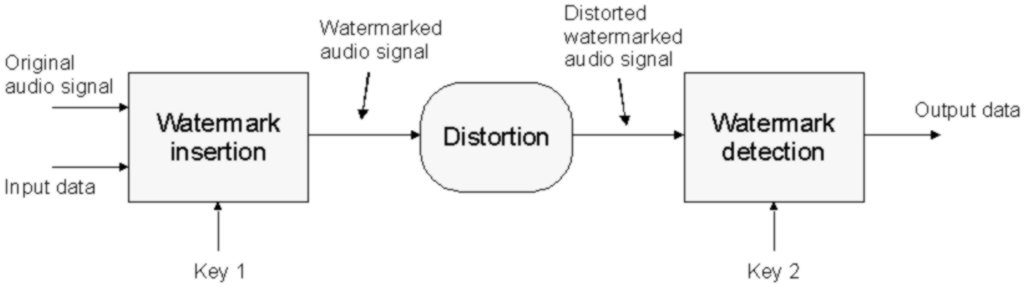

Figure 1: Watermarking as a communication channel.

As in cryptography, a *key* is generally used during the construction of the watermark, and another key (which may or may not be identical to the first one) is required for watermark detection. Despite this similarity, watermarking differs from cryptography in its very essence. While an encrypted audio file is useless without the corresponding decryption key, no such information is necessary in order to listen to a watermarked audio file. The important point is that the watermark is always present in the signal — even in illegal copies of it — and the protection that is offered by a watermarking system is therefore of a permanent kind. The same is not true for a cryptographic system, as audio files must be decrypted (and thus unprotected) in order to become usable.

Let us clarify the utilization of a watermarking system through an example. Audio content can be watermarked with a "copy-never" watermark. A compliant CD-writer device will analyze the input audio signal and check for the presence of the watermark before recording. If no watermark is found, the content is assumed to be copyright-free and the CD is recorded; otherwise, the equipment refuses to perform the requested operation. A more sophisticated system could admit multiple degrees of protection, ranging from "copy-never" to "copy-freely". For instance, audio marked as "copy-twice" could be duplicated, but the resulting copy would have its watermark set to the "copy-once" state. If a second copy were made from this first copy, it would be marked as "copy-never" and would not be reproducible. This would limit the number of generations in the duplication process — if you have an original CD, you can burn a copy for a friend, but he might not be able to do the same from the copy you gave him.

A watermarking system is called *symmetric* if the same key is used for both insertion and detection. When these keys are different from each other, the system is called *asymmetric.* Symmetric watermarking systems are suitable for *private* watermarking, where the key is kept secret; in contrast, asymmetric watermarking is appropriate for *public* watermarking, where a private key is used for watermark insertion and a public key for watermark detection. As in public encryption systems (in particular the RSA system (Boneh, 1999)), the idea of a non-invertible function is present: the public key is derived from the private key, but the private key cannot be deduced from the public key.

The requirements that an audio watermarking system must satisfy are application-dependent and often conflicting. As general requirements, we can mention:

- *Inaudibility*: watermarking should not degrade sound quality.
- *Robustness*: the watermark should resist any transformations applied to the audio signal, as long as

sound quality is not unacceptably degraded.

- *Capacity*: the watermark bit rate should be high enough for the intended application, which can be conflicting with inaudibility and robustness; a trade-off must be found.
- *Reliability*: data contained in the watermark should be extracted with acceptable error rates.
- *Low complexity*: for real-time applications, watermarking algorithms should not be excessively time-consuming.

All these requirements are to be respected to a certain extent, according to the application. Some applications (such as low bit-rate audio over the Internet) might admit the watermark to introduce a small level of sound quality degradation, while others (such as high bit-rate audio) would be extremely rigorous on that matter. Resistance to signal-processing operations such as filtering, resampling or coding is usually necessary. For copyright protection, resistance to malicious attacks aimed at preventing watermark detection is also required; for example, if a piece of the signal is deleted, the watermark should still be detectable. However, for integrity-verification applications (e.g., of testimonies recorded before a court), the watermark must no longer be recognized when audio content is modified in any way. In that case, robustness is no longer required; on the contrary, the watermark must be fragile.

2.1 How it works

Watermarking can be viewed as a communication system: the watermark is the signal carrying useful information and the audio signal plays the role of channel noise. In conventional communication systems, the useful signal is usually stronger than the noise, and the latter is often assumed to be Gaussian and white. This is not the case in watermarking. To avoid audible distortion, the watermark signal must be much weaker (some tens of decibels) than the audio signal. Furthermore, the audio signal is generally non-stationary and strongly colored.

Several approaches for audio watermarking have been proposed in the literature. For example, we can mention:

- *Spread-spectrum watermarking*: As in spread-spectrum communication systems (Dixon, 1976) (Haykin, 1988) the idea consists in spreading the watermark in frequency to maximize its power while keeping it inaudible and increasing its resistance to attacks (Boney et al.,1996) (Garcia, 1999).
- *Echo-hiding watermarking*: Temporal masking properties are exploited in order to render the watermark inaudible. The watermark is an “echo” of the original signal (Bender et al., 1996) (Neubauer, 2000).
- *Bit stream watermarking*: The watermark is inserted directly in the bit stream generated by an audio coder. For example, in (Lacy et al., 1998), the watermark consists in the modification of scale factors in the MPEG AAC bit stream.

Many variations of these basic schemes have been proposed. For example, rather than adding the watermark to the audio signal in the time domain, some systems perform this operation in the frequency domain by directly replacing spectral components (Garcia, 1999). Other systems use different keys (i.e. different codebooks) for watermark insertion and detection; in such asymmetric schemes, detection methods more sophisticated than simple correlation calculations must be used (Gomes et al., 2000) (Furon et al., 2000).

A major difficulty in watermarking is the need for synchronization during detection. In general, the detector must know the starting and finishing times of each symbol contained in the watermark. In the context of copyright-protection applications, the system should resist desynchronization attacks such as the suppression (or addition) of samples to the audio signal. Some resynchronization methods have recently been proposed (Gomes et al., 2001) (Furon et al., 2000). These methods allow the system to resist a large class of desynchronization attacks.

Psychoacoustic models

Psychoacoustics is the study of the perception of sound. Through experimentation, psychoacousticians have established that the human ear presents several limitations. In particular, when two tones, close to each other in frequency, are played simultaneously, *frequency masking* may: if one of the tones is sufficiently loud, it *masks* the other one (Zwicker & Fastl, 1990).

Psychoacoustic models generalize the frequency-masking effect to non-tonal signals. From an audio signal $u(t)$, these models calculate a curve $M_u(f)$ called *masking threshold* that is homogeneous to a power spectral density (PSD) (Perreau, 1998). If the PSD $V(f)$ of a signal $v(t)$ is below $M_u(f)$ for all frequencies, then $v(t)$ is masked by $u(t)$. This means that the listener is unable to perceive any difference between $u(t)$ and $u(t) + v(t)$ (Figure 2). These models are widely used in lossy compression methods (such as MP3 or MPEG-AAC (International Organization for Standardization, 1997) (Bosi et al., 1997)) to render quantization noise inaudible, thus providing high quality audio at low bit rates.

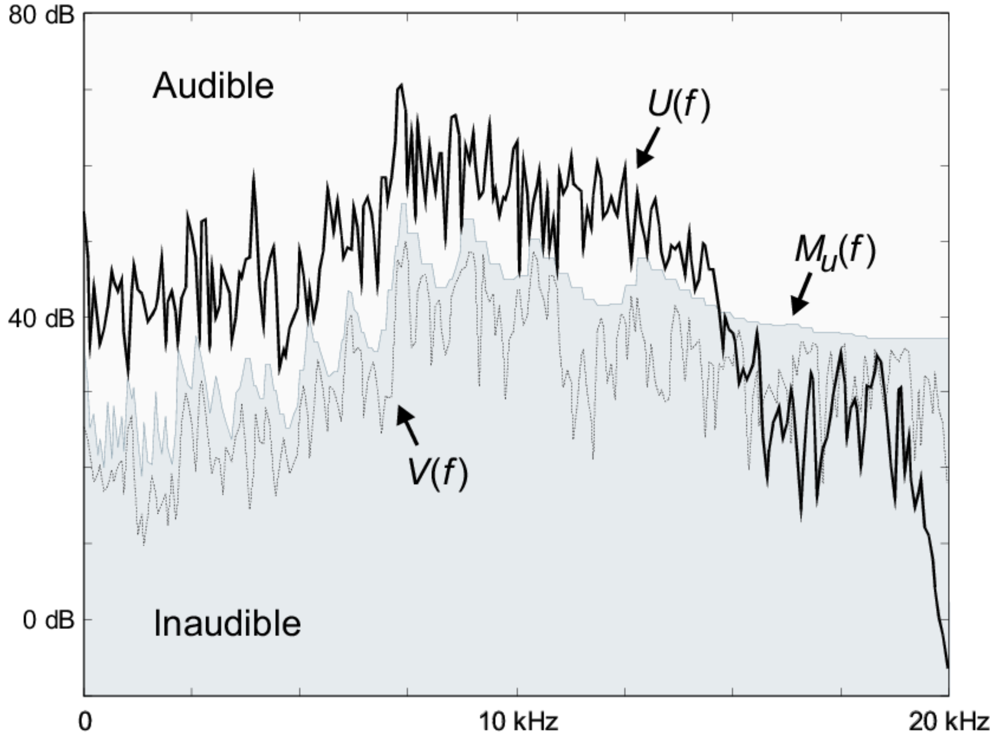

Figure 2: PSDs of the masking and masked signals (*U(f)*, continuous line, and *V(f),* dotted line, respectively), as well as the masking threshold $M_u(f)$. In the upper part of the spectrum, the masking threshold (and the masked signal) often surpasses the masking signal due to the low sensibility of the human ear to high frequencies.

In audio watermarking, psychoacoustic models are often used to ensure inaudibility of the watermark. The watermark is constructed by shaping in frequency a nearly-white signal according to the masking threshold. After this operation, the PSD of the watermark is always below the masking threshold and the watermark should not be heard in the presence of the original audio signal. Thanks to psychoacoustic models, inaudibility can be reached at signal-to-watermark power ratios of approximately 20 dB. In contrast, a white watermark would require much higher ratios to ensure inaudibility, thus rendering detection more difficult.

Masking can also occur in time domain with pre or post-masking. If two sounds are close to each other in time and one is sufficiently loud, it will mask the other one. This effect is exploited in lossy compression methods to further increase the compression rate (International Organization for Standardization, 1997).Post-masking is also used in “echo-hiding” watermarking systems: the watermark is a delayed and attenuated version of the audio signal, and the delay between the audio signal and this "echo" is used as a means of coding information. Many other applications have been proposed for psychoacoustic models. To name a few: echo cancellation, automatic audio quality evaluation and hearing aids for the deaf.

2.2 Case study

We present here a case study to illustrate a real-world watermarking system. This system was implemented at Paris V University in collaboration with ENST-Paris. The goal is to transport information between two computers through an acoustic channel by means of a watermarked piece of music. On the first computer, the user types a text that is converted into binary information. An audio file (also chosen by the user) is then read and the binary information is embedded into it through watermarking. Finally, the watermarked audio is played through a loudspeaker. At the receiver side, the second computer records the watermarked audio (along with ambient noise) through a microphone. It then performs the watermark-detection procedure on the sampled audio and retrieves the binary information, which is converted back into text and finally displayed on the second computer screen. This configuration is shown in Figure 3.

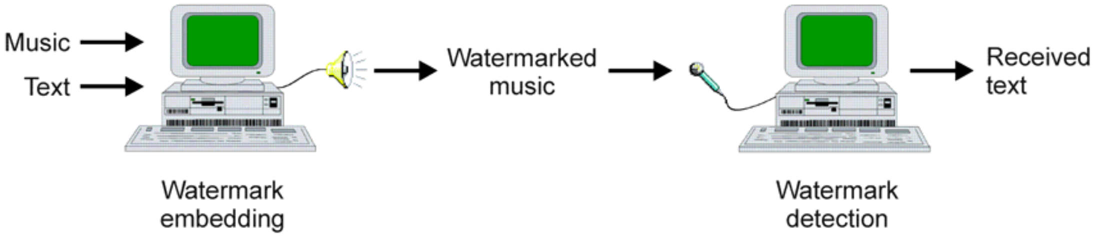

Figure 3: Transport information through an information channel.

In this application, a classical symmetric watermarking system (with a single key) was employed (Boney et al., 1996) (Gomes et al., 2001). This watermarking scheme is shown in Figure 4.

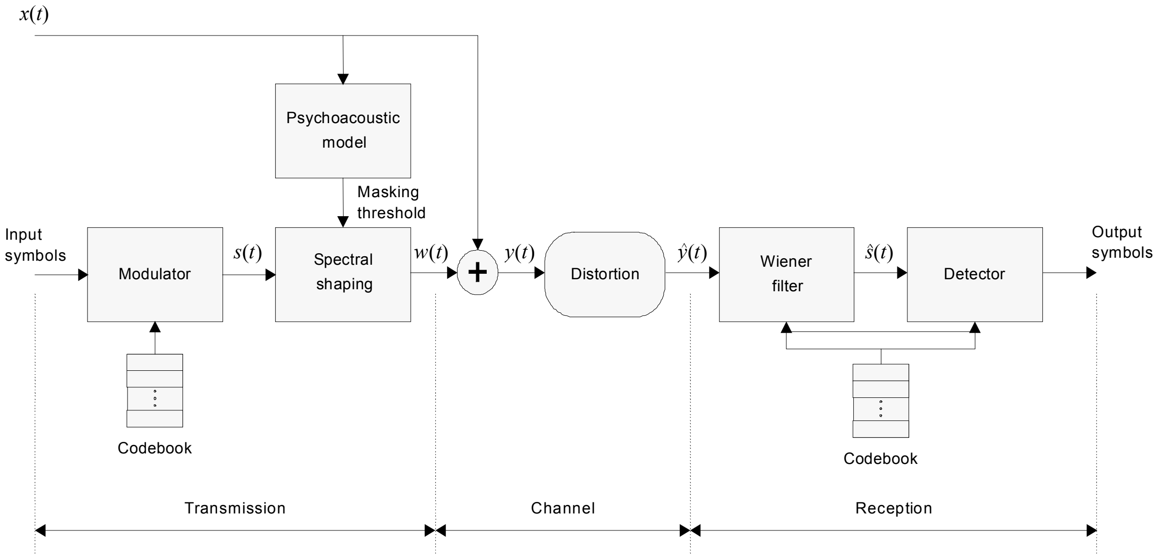

Figure 4: Employed watermarking scheme.

Processing is done for individual signal windows (some tens of milliseconds long). In these time frames, the audio signal is considered almost-stationary. A sequence of input symbols (generated from the message to be transmitted) is converted into a modulated signal *s(n)* by means of a codebook that associates a unique vector to each symbol in the alphabet. This codebook corresponds to the key of the watermarking system. The vectors in the codebook can be obtained through traditional digital modulation techniques (such as PSK or FSK) followed by frequency spreading, or they can simply correspond to orthogonal realizations of a white random process. To ensure inaudibility, the modulated signal is frequency-shaped according to a masking threshold calculated through a psychoacoustic model, which gives the watermark *w(t).* The watermarked signal *y(t)* is obtained by adding *w(t)* to the original audio signal *x(t)*.

The audio signal $\hat{y}(t)$ received by the detector can be written as

$$\hat{y}(t) = y(t) + n(t) = w(t) + x(t) + n(t) \qquad (1)$$

where *n(t)* is an additive noise introduced by any process the watermarked signal might have undergone (such as filtering, compression/decompression or radio transmission). In some cases, this noise can be as strong as the watermark, or even stronger than it. Detection is then performed through correlation between the received signal (Wiener-filtered in order to improve the watermark-to-signal power ratio) and the vectors in the codebook, producing a sequence of output symbols from which the hidden information is retrieved.

The system must resist D/A-A/D conversion, transducer distortion, ambient noise and desynchronization between transmitter and receiver. Of course, there must be no perceptible difference between the original and the watermarked audio signals.

Transmission

In order to obtain a watermark that is spread in frequency, a codebook composed of white, orthogonal Gaussian vectors is used. The number of vectors is a function of the desired bit rate. Each codebook entry is associated with a symbol representing a specific binary pattern. The binary sequence to be transmitted, obtained from the input text through an ASCII-like code, is converted into a sequence of symbols. The latter is then converted into a modulated signal by concatenating the corresponding vectors from the codebook.

To allow for resynchronization in the receiver, synchronization symbols are regularly transmitted. In addition, a repetition code is used to improve detection and resynchronization performance. The resynchronization scheme is detailed later in this section.

To ensure watermark inaudibility, the modulated signal is spectrally shaped through filtering according to a masking threshold, which is obtained from a psychoacoustic model. This procedure, repeated for each window of the audio signal (≈ 12 ms), produces the watermark. The watermarked signal is obtained by simply adding together the original audio signal and the watermark.

Channel

The watermarked signal is converted into analog form and played through a loudspeaker. At the receiver side, a microphone captures the sound, producing an electric signal that is sampled and digitalized. This channel introduces significant distortion into the signal, namely:

- quantization noise due to analog-to-digital conversion;
- distortion due to non-flat frequency response of the transducers;
- distortion due to non-linearity in the elements of the channel (transducers and converters);
- ambient noise.

Additionally, the sampling rates of the D/A and A/D converters may not be strictly identical. Even a small difference in the sampling rates may cause synchronization problems, as its effect is cumulative.

Reception

The initial step in the reception is the synchronization procedure, described later in this section. After resynchronization, watermark detection is performed. For each window of the received signal, the watermark signal is strengthened by Wiener-filtering and correlation measures with each codebook entry are calculated. The symbol associated with the codebook entry that maximizes correlation is chosen as the received symbol. The so-constructed sequence of received symbols is converted into a binary sequence, which is then converted into text. Finally, this text is displayed on the screen.

Synchronization

The receiver does not know the exact start of the transmission; an initial-synchronization method is thus required. In addition, a difference in the sampling rate of the D/A and A/D converters must be continuously compensated.

Initial synchronization is implemented by introducing a known sequence of symbols at the beginning of the transmission. This sequence must be long enough to ensure a low possibility of incorrect initial synchronization, since this would compromise the entire detection process. A 36-bit synchronization sequence is used in the system.

Continuous resynchronization is performed by inserting synchronization symbols at the beginning of each *block* of symbols. Apart from the synchronization symbol, a block contains multiple copies of the same symbol. This repetition code is used to improve system robustness to distortion. A group of blocks is called a *frame*. This structure is illustrated in Figure 5.

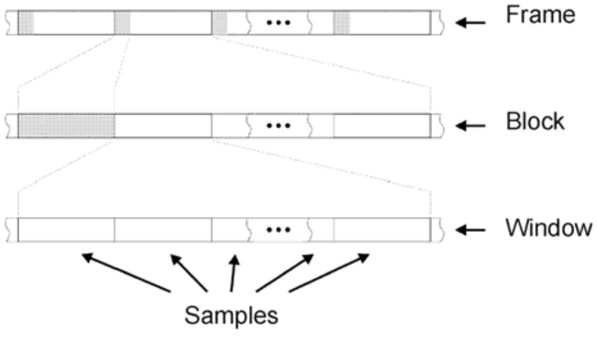

Figure 5: Data structure for synchronization and detection improvement. Dark gray corresponds to synchronization symbols; light gray corresponds to data symbols. Each symbol corresponds to a window

Accepted for publication on May 9th 2002

of the signal.

The exact starting point of the frame is initially unknown. The first step for resynchronization consists in averaging all blocks in a frame, as illustrated in Figure 5. (Note that the initial block in a frame is arbitrary.) As the synchronization symbol in the beginning of each block is always the same but data symbols are generally different among blocks, this operation will strengthen the synchronization symbol. Correlation measures between the averaged block and the synchronization symbol are then calculated for each possible beginning of this block. The exact starting point of the averaged block (and of the frame) is determined by maximizing this correlation. A frame should not contain too many blocks and a block should not contain too many windows, as the efficiency of the averaging operation would be reduced by cumulative desynchronization. In our experiments, 3 blocks per frame and 5 windows per block (512-sample windows at a sample rate of 44,1 kHz) seemed to be the optimal values, although this would depend on the exact desynchronization between the sound cards in the two PCs.

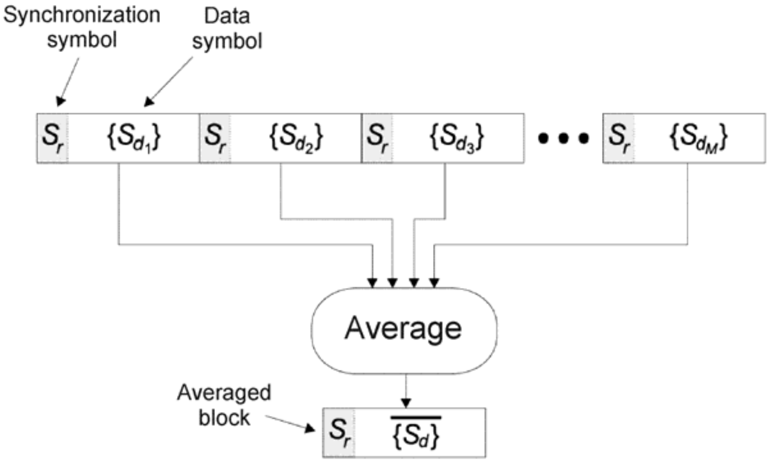

Figure 6: Calculation of the averaged block for resynchronization.

Once the beginning of the frame (and of the blocks in it) is determined, each block is treated separately. The symbols in a block (except the synchronization symbol) are averaged in order to improve the signal-to-noise ratio. Watermark detection is then performed as explained previously.

Results

In this experiment, ordinary PC sound cards, loudspeaker and microphone were used. The microphone was 1.5 m away from the loudspeaker. For an average signal-to-watermark power ratio of approximately 20 dB (which ensures watermark inaudibility) and for various kinds of music (sampled at 44.1 kHz), there were no detection errors when background noise was at reasonable levels (e.g., a person speaking in the room but away from the microphone). With strong background noise (e.g., a person speaking directly into the microphone), the error rate augmented significantly (from 10% with quiet voice to almost 100% with very loud voice).

3 Audio fingerprinting

In Watermarking, 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. In Audio Fingerprinting the message is automatically derived from the perceptually most relevant components of sound. This makes it less vulnerable to attacks since trying to remove this message, the fingerprint, would alter the quality of the sound (Craver et al., 2001).

Fingerprinting, or content-based identification (CBID), technologies work by extracting acoustic relevant characteristics of a piece of audio content and storing 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 complex matching algorithms and acoustic fingerprints different versions of a single recording can be identified as the same recording.(Recording Industry Association of America, 2001) (Napster, 2002).

The areas relevant to audio fingerprinting include Information Retrieval, Signal Processing, Pattern Recognition, Databases, AI, Computer Music and Music Cognition (Dannenberg et al., 2001).

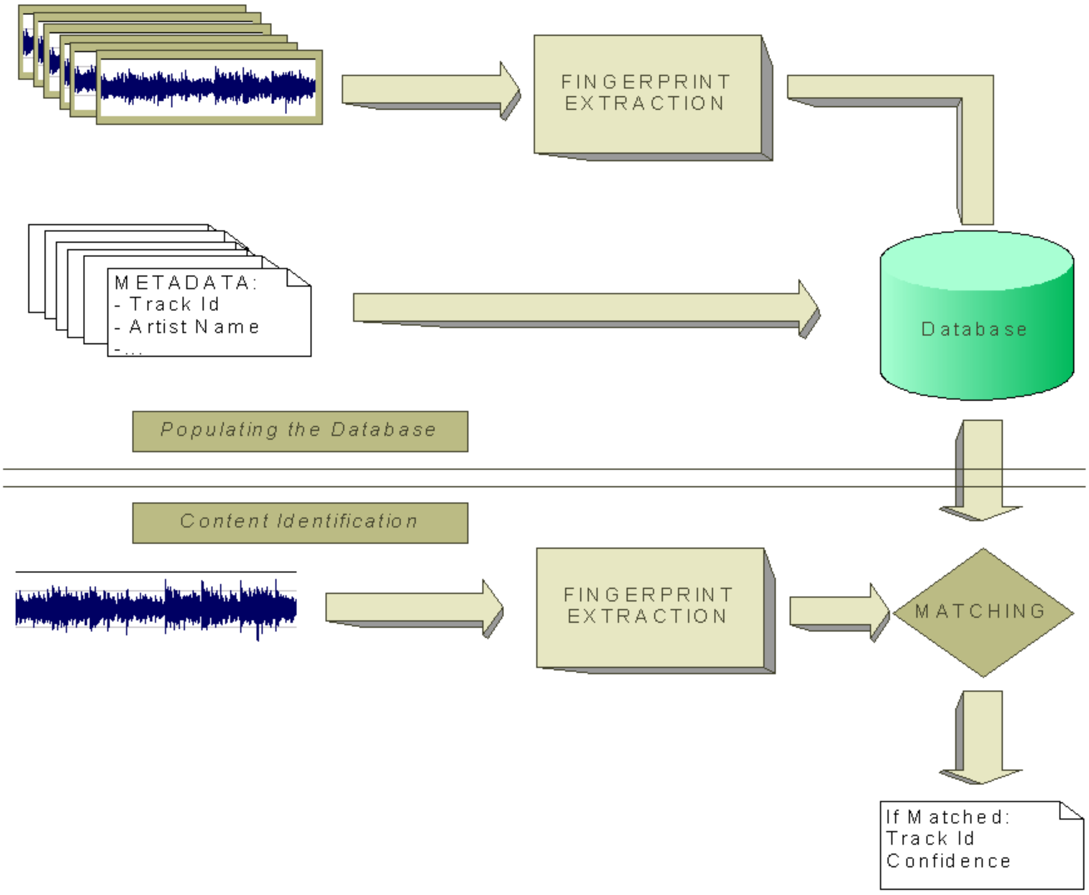

Figure 7: Fingerprinting overall functionality.

Requirements
The requirements depend obviously on the application but are useful in order to evaluate and compare different content-based identification technologies. In their call *Request for Information on Audio Fingerprinting Technologies* (Request for Audio Fingerprinting Technologies, 2001), the IFPI (International Federation of the Phonographic Industry, 2002) and RIAA (Recording Industry Association of America, 2001) were trying to evaluate several identification systems. For such a system it is mandatory to be computationally efficient and robust. A more detailed enumeration of requirements can help distinguish among the different approaches (Audio Identification Technology Overview, 2001).

- *Accuracy:* the number of correct identifications, non-identifications (false negatives), and wrong identifications (false positives). The issue of avoiding false positives is of great importance in applications like a monitoring system able to automatically generate play lists of registered songs for copyright enforcement organizations where a song, which has not been broadcast, should not be identified as a match, even at the cost of missing right matches. Approaches to deal with false positives have been treated in (Cano et al., 2001) and (Herre et al., 2001). In other applications, like automatic labeling of MP3 files (see the applications section) avoiding false positives is not such a mandatory requirement.
- *Security*: Vulnerability of the solution to hacking, tampering.
- *Versatility*: *Does the* identification of several formats requires different fingerprints? Can it identify streaming audio content? Will the technology work with previously released content (legacy content)?
- *Scalability*: Performance with very large databases of titles or a large number of concurrent identifications. This affects on accuracy and on fast matching of the database.
- *Robustness*: Ability to accurately identify regardless of the level of compression and distortion or interference in the transmission channel. Ability to identify from excerpts of songs.

3.1 How it works

Independently of the specific approach to extract the content-based compact signature, a common architecture can be devised to describe the functionality of fingerprinting (Request for Audio Fingerprinting Technologies, 2001).

From the Figure 7 it is possible to distinguish two operating modes. The overall functionality mimics the way humans perform the task. Off-line a memory of the works to be recognized is created; in the identification mode, unlabelled audio is presented to the system to look for a match.

- Building the database: The collection of works to be recognized is presented to the system. The system processes the audio signals extracting unique representations based on their acoustic characteristics. This compact and unique representation is stored in a database and can be linked with a tag or other metadata relevant to each recording.
- Actual Audio Identification: The unlabelled audio is processed in order to extract the fingerprint. The fingerprint is then compared to the fingerprints of the database. If a match is found, the tag associated with the work is obtained from the database. A confidence of the match can also be provided.

The actual implementations of audio fingerprinting normally follow the presented scheme with differences on the acoustic features observed and the modeling of audio and of a whole recording, as well as the matching and indexing algorithms.

The simplest approach would be direct file comparison. It consists on extracting a hash out of the bits of the binary file with MD5 (Message Digest 5) or CRC (Cyclic Redundancy Checking) methods to have a compact representation to store in the database. Of course this is not robust to compression or minimal distortions of any kind and in fact maybe cannot even be considered as content-based identification of audio since these methods do not consider the content, understood as information, of the file, just the bits. . This approach would also not be applicable when monitoring streaming audio or analog audio but it sets the basis for some fingerprinting research: Robust or Perceptual Hashing (Haitsma et al., 2001) (Mihçak & Venkatesan, 2001). The idea behind this higher-level hashing is that the hashing function incorporates acoustic features so that final hash code is robust to different audio manipulations as long as the content is preserved.

An important amount of features can be found in the literature: energy, loudness, spectral centroid, zero crossing rate, pitch, harmonicity, spectral flatness (Ismirli, 2000) (Allamanche et al., 2001), Mel-Frequency Cepstral Coefficients (Logan, 2000), etc. It is common to several methods to perform a filter bank analysis, do some transformation of the feature vector and in order to reduce the size of representation, extract some statistics: means or variances, over the whole recording (Etantrum, 2001), or extract a codebook for each song by means of unsupervised clustering (Herre et al., 2001). Other methods apply higher-level algorithms that try to go beyond the signal processing comparison and have notions of beat, harmonics, etc (Blum et al., 1999).

3.2 Case Study

A case study to illustrate with more detail an implementation of an audio fingerprinting solution is presented. The implementation was designed with one of the highest robustness requirements, identification of radio broadcasted songs. The difficulty inherent in the task of identifying broadcasted audio material is mainly due to the difference of quality of the original titles in the CD and the quality of the broadcasted ones. The song is transmitted partially, the speaker talks on top of different fragments, the piece is maybe playing faster and several manipulation effects are applied to increase the listener's psycho-acoustic impact (compressors, enhancers, equalization, bass-booster, etc). Moreover, in broadcasted audio stream there are no markers indicating the beginning and the end of the songs.
Yet the system also has to be fast because it must do comparisons with several thousand (of the order of 100.000's) songs online. This affects the memory and computation requisites since the system should observe several radio stations, give results online 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 sequence of timbres.

The 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 unlabelled audio and the set of songs are decomposed into these audio units ending up with a sequence of symbols for the unlabelled audio and a database of sequences representing the original songs. By approximate string matching, the song sequences that best resembles the sequence of the unlabelled 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 (Batlle & Cano, 2000). The extraction of ADU is shown in Figure 8.

The audio data is pre-processed by a front-end in a frame-by-frame analysis. In the first block a set of relevant feature vectors are extracted from the sound. Within the front end, a normalization of feature vectors as well as some other processing is done before the decoding block (Haykin, 1996). In the decoding block, the feature vectors are run against the statistical models of the ADU: HMM-ADU using the Viterbi algorithm (Viterbi, 2001). As a result, the most likely ADU sequence is produced.

Front-End Feature Extraction

The first stage in a classification system is the obtainment of a set of values that represent the main characteristics of the audio samples. A key assumption made at this step is that the signal can be regarded as stationary over an interval of a few milliseconds. Thus, the prime function of the front-end parameterization stage is to divide the input sound into blocks and from each block derive some features, like a smoothed spectral estimate.

The spacing between blocks is around 10 ms and blocks are normally overlapped to give a longer analysis window, typically 25 ms. As with all processing of this type, a tapered window function (e.g., Hamming) is applied to each block so as to minimize the signal discontinuities at the beginning and end of each frame (Oppenheim & Schafer, 1989).

It is well known that the human ear performs some kind of signal processing before the audio signal enters the brain (Ruggero, 1992). Since this processing has proven to be robust in front of several kinds of noises and distortions in the area of speech recognition, it seems reasonable to use a similar signal front-end processing for music in the system. The required spectral estimates are computed via Fourier analysis and there are a number of additional transformations that can be applied in order to generate the final acoustic vectors. To illustrate one typical arrangement, Figure 8 shows the front-end to generate Mel-Frequency Cepstral Coefficients (MFCCs).

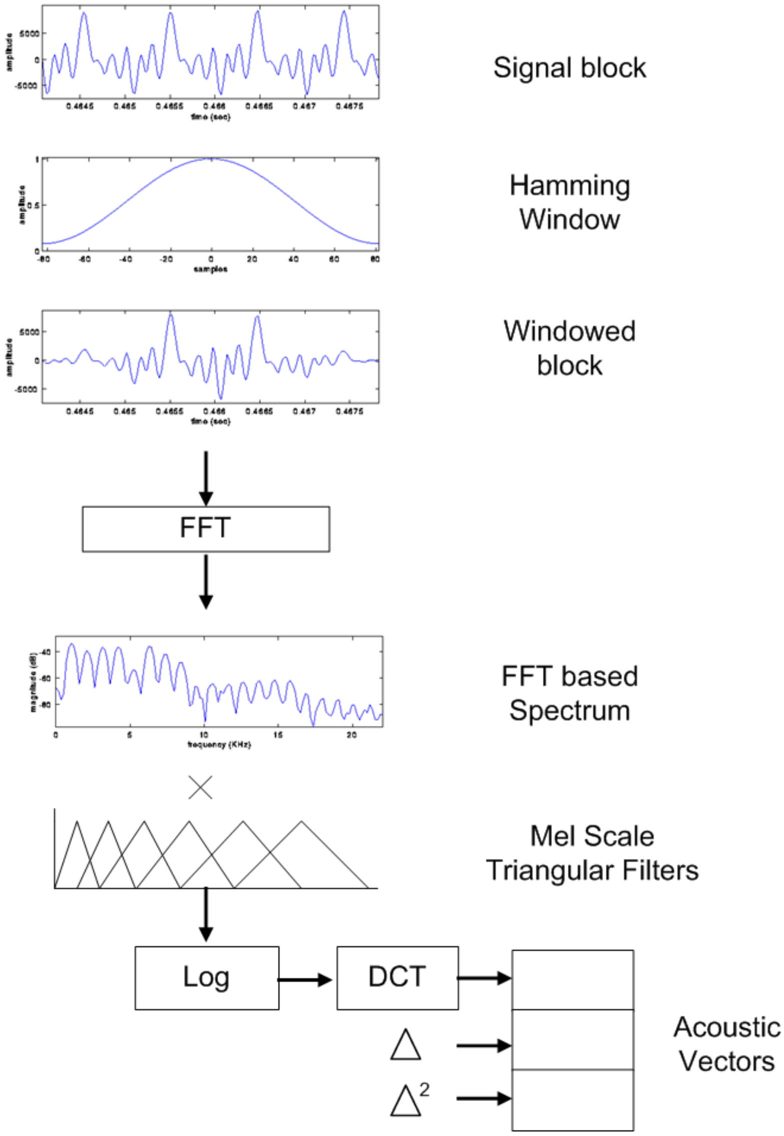

Figure 8: Front-end feature extraction.

To compute MFCC coefficients, the Fourier spectrum is smoothed by integrating the spectral coefficients within triangular frequency bins arranged on a non-linear scale called the Mel-scale. The Mel-scale is designed to approximate the frequency resolution of the human ear (Ruggero, 1992) being linear up to 1000 Hz and logarithmic thereafter. In order to make the statistics of the estimated song power spectrum approximately Gaussian, logarithmic (compression) conversion is applied to the filter-bank output. The final processing stage is to apply the Discrete Cosine Transform to the log filter-bank coefficients. This has the effect of compressing the spectral information into the lower order coefficients and it also de-correlates them (Batlle et al., 1998).

The acoustic modeling based on HMM assumes that each acoustic vector is independent of its neighbors. This is a rather poor assumption since physical constraints of the musical instruments ensure that there is continuity between successive spectral estimates. However, appending the first and second order differentials to the basic static coefficients will greatly reduce the problem. Obviously, there are more features that can be extracted, like tonality center, rhythm, etc to feed the models.

Acoustic Modeling: HMM ADU

The purpose of the acoustic models is to provide a method of calculating the likelihood of any sequence of ADU given a vector sequence Y. Each individual ADU is represented by a Hidden Markov model (HMM) (Rabiner, 1989). An HMM is most easily understood as a generator of vector sequences. It is a finite state machine which changes state once every time unit and each time t that a state j is entered, an n acoustic vector y_t is generated with probability density $b_j(y_t)$. Furthermore, the transition from state i to state j is also probabilistic and governed by the discrete probability a_{ij}.

The joint probability of a vector sequence Y and state sequence X given some model M is calculated simply as the product of the transition probabilities and the output probabilities. The joint probability of an acoustic vector sequence Y and some state sequence $X = x(1), x(2), x(3),\dots,x(T)$ is:

$$P(Y, X|M) = a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(y_t) a_{x(t)x(t+1)}$$

In practice only the observation sequence Y is known and the underlying state sequence X is hidden. This is why it is called *Hidden Markov Model.*

Viterbi Decoding

For the decoding of ADU sequences, the trained models are run against the feature vectors using the Viterbi algorithm (Viterbi, 2001). As a result, the most probable path through the models is found, providing a sequence of ADU Models and the points in time for every transition from one model to the following.

The Viterbi algorithm is an efficient algorithm to find the state sequence that most likely produced the observations. Let $\phi_j(t)$ represent the maximum likelihood of observing acoustic vectors y_1 to y_t and being in state j at time t. This partial likelihood can be computed using the following recursion

$$\phi_j(t) = \max_i \left\{ \phi_j(t-1) \cdot a_{ij} \right\} b_j(y_t)$$

where $\phi_1(1) = 1$, $\phi_j(1) = a_{1j} b_j(y_1)$, for $1<j<N$. The maximum likelihood P'(Y|M) is then given by

$$\phi_N(T) = \max_i \left\{ \phi_j(T) a_{iN} \right\}$$

By keeping track of the state j giving the maximum value in the above recursion formula, it is possible, at the end of the input sequence, to retrieve the states visited by the best path, thus obtaining the most probable ADU sequence given the input frames.

As it has been shown, the algorithm performs the backtracking at the end of the audio data. When dealing with streaming audio data, like when observing a radio broadcast, it is necessary to provide the ADU sequence in real time (plus a little latency time). Fortunately, it is possible to modify the algorithm to work online (Loscos et al., 1999). To do so, the backtracking is adapted to determine the best path at each frame iteration instead of waiting until the end of the utterance. Doing so it is possible to detect that the sequence of ADU up to a point has converged, that is to say, it has become stationary and included in the best path from a certain time on, they can be extracted. The time of convergence is variable and depends on the probabilistic modeling of the sound with the current HMMs. The fingerprint has an average bitrate of 50 bps.

4 Copyright-related applications

Enforcement of copyright is the application that first motivated the development of audio watermarking, and it is also a major application of fingerprinting systems. We discuss below how these methodologies can be used to prevent and confront piracy. Among other applications related to audio tracking included in this section, we name:

- Playlist generation of radio stations and music television channels from an independent and neutral source.
- Statistical summaries, like audience characteristics and audience preferences, which can be used for in-house programming and broadcast policy.
- Exact accounting and appropriate refund for rightholders.

4.1 Proof of ownership

Content creators are often concerned about the possibility of their work to be “appropriated” by other people. Let us imagine the following scenario: (1) artist A records a song and makes it available on his website; (2) artist B gets a copy of this song and releases it as his own (possibly for financial profit); (3)

artist A sues artist B but is unable to prove that he is the actual author of the song. This situation is more likely to take place when the content creator is not widely known to the public: while no one would have doubts about the ownership of a song released by a major pop artist, a lesser known artist might have trouble proving he is the actual author of a song if someone else manages to appropriate it.

This situation can be resolved by the introduction of a “trusted third party” (TTP) that acts as a repository of audio content. This is actually the case in many countries. Before releasing new work, the artist registers it with a TTP (possibly a governmental agency) that keeps a copy of it on file. The artist is then in position to successfully sue anyone who tries to appropriate his work without permission.

We can conceive a procedure relying on watermarking to prove rightful ownership. A unique secret key, the owner’s *signature,* is used to generate a watermark embedded into the audio signal. The signature is registered with a TTP. The presence of the watermark must be accepted by a court of law as evidence of ownership. This procedure avoids the need of transferring the audio content itself to the TTP, and new audio content is automatically protected if watermarked with the same key.

Audio fingerprinting can also be used in a procedure to prove rightful ownership. Instead of registering the complete piece of music with a TTP, only its fingerprint is registered, and that fingerprint must be accepted as evidence of ownership by a court of law. The advantage of this approach in comparison with the traditional one is the possibility of easily verifying if a specific piece of music is present in the TTP’s database (as long as this database is available to the public).

Some points regarding these approaches must be stressed:

- The watermarking or fingerprinting procedure must be recognized by authorities as valid for legal purposes, which raises questions concerning system security and reliability that have not yet been satisfactorily answered.
- If a watermark is to be accepted as *proof* (not only evidence) of ownership, the watermarking process has to be entirely controlled by the TTP, which assigns and holds secret keys (unknown even to signal owners).
- The availability of different signals watermarked with the same key might help pirates isolate and remove the watermark through averaging.
- The watermarking system has to avoid ambiguity attacks, which consist in choosing keys such that false positives are induced during watermark detection for a specific audio signal. If keys are assigned by the TTP, this attack can be easily prevented.
- The fingerprinting system should never associate identical fingerprints to different pieces of music, even if someone intentionally tries to force this situation (for illicit purposes). Again, if the TTP is responsible for the extraction of the fingerprints, this attack can be easily avoided.
- Watermarking and fingerprinting can protect audio content only against unfounded claims of ownership regarding specific recordings; they cannot protect audio content from illicit acts such as plagiarism. This kind of protection requires two steps: (1) watermarking or fingerprinting is used to prove ownership of a specific original recording, and (2) the original recording is compared with a suspicious one through listening tests.

A watermarking-based system for proving ownership would be difficult to implement at the present time, as no audio watermarking system can be considered sufficiently robust to malicious attacks. While a similar system based on audio fingerprinting might seem more feasible, the advantages it presents in comparison with the traditional system (i.e., registering the complete music piece with a TTP) might not suffice to justify the use of this technology.

4.2 Monitoring at the consumer end

What renders the watermarking-based proof-of-ownership application unrealistic is the requirement of absolute robusteness to malicious attacks. If this constraint is relaxed, the problem becomes feasible. A robust yet not indestructible watermark might not be accepted as evidence before a court of law, but it can be used by rights holders to discourage illicit usage of audio signals. If copyright infringement is detected, the problem is dealt with through conventional legal channels.

In usage-policy monitoring applications, the goal is to avoid misuse of audio signals by the consumer. The watermark contains information that dictates the behavior of compliant devices (e.g., CD players and recorders, MP3 players or even computers) in accordance with the usage policy. As the typical end user does not have the necessary skills to erase the watermark, such a system should prevent most home piracy (“keep honest people honest”). In contrast, professional pirates would probably be able to overcome this protection system, although the cost of doing so might render piracy a less attractive activity.

In principle, fingerprinting is not appropriate for usage monitoring at the consumer end, as it does not add information to the signal. This limitation can be circumvented by conceiving a system where a piece of music is identified by means of a fingerprint and a database is then contacted to retrieve information about user rights. Compliant devices are required to be connected to a network in order to access the database. This may sound unrealistic at the present time but should not be troublesome in the near future, as the tendency of connecting electronic equipments other than computers to the Internet seems to strengthen. The practical utility of this system can be compromised if the rights granted to the consumer depend not only on the piece of music itself, but also on the price paid for it by the consumer.

4.3 Monitoring at the distribution end

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 and a new file-filtering system relying on it was to be released as of this writing (TRM 2001). Other systems are also commercially available (Baytsp, 2001) (AudibleMagic, 2001) (Music Reporter, 2001) and (Auditude, 2001). Such systems require a database containing fingerprints of every recording that shall not be freely transferred. Before a file transfer between users A and B is initiated, the fingerprinting of the requested recording is extracted locally on user A's machine by the client software and sent to a server that searches for it in the database. If the fingerprinting is not found, the recording is considered to be copyright-free and the file transfer to user B's machine is promptly authorized; otherwise, a fee must be paid in order for the transfer to be initiated.

A similar system based on audio watermarking can also be conceived. If all copyrighted recordings are watermarked before release, the server can check for the presence of the watermark and act accordingly. If no watermark is found, the recording is considered to be copyright-free. The advantage of this approach is the absence of a database, as all the information necessary for system operation is carried by the watermarked signal itself. This means that new recordings are automatically protected (as long as they are watermarked), while a fingerprinting-based monitoring system requires the databases of all distributors to be updated before protection becomes effective. The complexity of fingerprinting match also augments as the database grows, whereas watermarking-based systems present constant complexity. However, watermarking is in general much more sensitive to malicious attacks than fingerprinting, thus reducing system reliability. In addition, legacy content (i.e. old non-watermarked recordings) would not be protected at all.

4.4 Identification of broadcast audio

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.

Fingerprinting-based monitoring systems are being used for this purpose (Music Reporter, 2001), (Auditude, 2001). The system "listens" to the radio and continuously updates a playlist 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.

A similar watermarking-based system can be conceived. All songs and commercials that shall be identified must be watermarked. The watermark contains information that uniquely identifies the song or commercial. Previously released non-watermarked recordings cannot be identified by the system, which can be a major drawback in this application. Watermarking is also more sensitive to distortion. However, watermarking does not require a database, and no update is necessary when new songs come out.

4.5 Tracking of illicit copies

Unauthorized use of copyrighted material (texts, images, sounds) has been common practice on the World-Wide Web since its very beginning. Web crawlers can be used to automatically search the Web for copyrighted material. In what concerns audio files, an automated direct comparison between material found on web pages and recordings contained in a database is difficult to implement, as there may be several variations of the same recording (e.g., different formats, different sample rates, cropped versions, compressed versions). This comparison would also be inefficient in terms of speed, as audio files tend to be large even for rather short recordings.

Watermarking and fingerprinting can be used in audio-file tracking systems. The watermarking-based approach consists in watermarking recordings to be protected before distribution. A web crawler will then search the Web and check for the presence of the watermark on each audio file it finds. If a watermarked recording is found, the system notifies the rights holder, who will contact the transgressor after manually confirming the infringement. The system might automatically send infringement notices to transgressors (without manual confirmation), but in this case the probability of false positives must be very low. An intermediate solution is also possible: the system asks for manual confirmation only when the watermark is detected with low reliability (i.e., a weak watermark is detected, which might constitute a false positive).

A fingerprinting-based approach would be quite similar, but the system would extract a fingerprint from each audio file found on the Internet and would search for it in a database of fingerprints. Again, the false-positive rate will dictate the need for manual confirmation of copyright infringement.

The comparative advantages of each approach are the same as presented in the previous sections: watermarking is more sensitive to malicious attacks than fingerprinting, and previously released non-watermarked recordings could not be tracked; on the other hand, in fingerprinting-based systems, tracking of new releases requires the database to be updated, while watermarking-based systems do not require any update. When compared to the direct match between audio files on the Internet and copyrighted recordings, both approaches should lead to a significant gain in speed, as much less information has to be matched (the watermark and the fingerprint correspond to only a small fraction of the amount of data required to store the original recording). A watermarking-based system would probably be faster than a corresponding fingerprinting-based system, as there is no database search.

4.6 Determination of the origin of illicit copies

When an illicit copy of a recording is found on a website, it may be possible for the rights holder to prosecute the website owner. But this might be just the final link in a piracy chain: if we could trace back towards the beginning of this chain, we would eventually reach an *original legal copy* of the recording. This original was legally purchased from an authorized distributor, but it was used in a way that infringed the usage policy defined by the rights holder.

If each legal copy of a recording is watermarked with different information, such as a unique serial number, it is possible to determine from which legal copy an illegal copy has been made[2]. When music is distributed on-line, this kind of watermark is generally embedded in the bit stream domain due to real time constraints. If the distributor keeps on file the identities of its clients and the serial numbers of all recordings each client purchased, the individual in the beginning of the piracy chain can be identified and prosecuted. This kind of protection system is particularly well-suited to web-based music distribution. When the user downloads a recording, the latter is watermarked in real time and the operation is registered in the distributor's files.

This kind of protection system would severely discourage non-professional piracy (again, "keeping honest people honest"). However, professional pirates might prevent watermark detection by averaging several legal copies of the same recording, which is called *collusion attack.* Since the watermark is unique in each legal copy, this attack would tend to recover the original non-watermarked recording, or at least would weaken individual watermarks, possibly to the point of rendering them undetectable. Fingerprinting is not appropriate for this application, since all copies of a recording have exactly the same fingerprint.

5 Added value services

Copyright-related applications based on watermarking require a certain degree of robustness to intentional attacks, which adds considerably to the complexity of the system. However, for the transport of information not related to copyright, malicious attacks are not an issue, since the information actually *adds value* to the signal. The watermark should still be robust to licit operations, which depend on the application (e.g., filtering, compression, D/A-A/D conversion, transmission through radio waves).

[2] This kind of watermark is sometimes called *fingerprint* in the literature. In this article, this term will not be used in this sense.

5.1 Content-related 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: rhythmic, timbral, melodic or harmonic description.
- 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 or artist biography.

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* (MDS), 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 (MPEG Working Documents) providing for easy data interchange between different equipments.

Some fields of an MDS can be automatically extracted from audio recordings, with greater or lesser success. This extraction is performed by means of signal processing techniques and psychoacoustic analysis. For example, melodic information can be derived from pitch data, rhythmic information can be derived from onset detection, and low-level timbre information can be derived from low-level audio descriptors. Nevertheless, many features, which are important for certain applications cannot be automatically extracted, such as the title, author, performer or edition year of a musical work. This information is usually stored externally to the audio data, either in a database or inserted in the structure of the audio file or transmitted frame, depending on the communication channel or storage support.

Some systems store MDSs in a database which is accessible through the Internet. Fingerprinting can then be used to identify a recording and retrieve the corresponding MDS, regardless of support type, file format or any other particularity of the audio data. Some commercial services are:

- Labeling unlabeled or mislabeled collections of music. For example, MusicBrainz (Musicbrainz, 2001), Id3man (Id3man, 2001) and Moodlogic (Moodlogic Inc., 2001) automatically label collections of audio files; the user can download a compatible player that extracts fingerprints and submits them to a central server where metadata associated to the recordings is downloaded. Some companies like Gracenote (formerly CDDB) (Gracenote, 2001) offer a similar service but use the table of contents of a CD as a hash to obtain metadata associated to it.
- Identification of a tune through a cell phone, offering the possibility of purchasing the song. This is one of the most demanding situations in terms of robustness, as the audio signal goes through radio distortion, speaker + microphone and GSM coding.

If the need of a database is an inconvenience, watermarking may be a better choice than fingerprinting: the watermark carries the MDS, which is immediately available to the reading device. An example of this approach can be found in (Packman & Kurth, 2000), where song lyrics are stored directly in the audio signal. The advantage of the watermarking-based approach is that content information is not dissociable from audio data, even after format conversions or change in support type; if the information were stored in a header, for example, it might not survive format conversion, and it would obviously not survive digital-to-analog conversion.

We can also imagine a system where both watermarking and fingerprinting are used. A fingerprinting-based search system retrieves a detailed MDS from a remote database. If this search fails (due, for example, to Internet connection problems), a watermarking-based system is used as a backup. The watermark contains basic information about the recording not as in-depth as the remote database, but enough for basic user needs.

5.2 Transport of general-purpose information

Watermarking can also be employed as a means for transmitting any kind of information, not necessarily related to the audio content. Radio stations, for example, can watermark their audio just before transmission, embedding text into it. This text then scrolls on a small screen on the receiver. Here are some examples of information that may be transmitted through the "watermarking channel":

- news;
- weather forecast;

- stock quotes;
- advertisement;
- identification of the station.

The last item can be useful, for example, in systems for automatic measurement of audience: the radio (or TV) station that is tuned is detected from the watermarked audio by a specific device. This information is registered and periodically transmitted to the company responsible for audience statistics.

Fingerprinting systems are not appropriate to the transport of general-purpose information, since their goal is to identify the signal without adding information to it.

6 Integrity-verification applications

Watermarking and fingerprinting can also be used to verify the *integrity* of a recording, i.e., to check whether a recording has been modified since it was watermarked or since its fingerprinting was extracted. Verification of integrity is important, for example, when a previously recorded testimony is used as evidence before a court: it is desirable to assure that the recording has not been edited in any way.

6.1 Fragile watermarks

Integrity of an audio recording can be controlled by means of a *fragile* watermark. If the watermarked signal is edited, the watermark must no longer be detectable. By “edited”, we understand any modification that could corrupt the *meaning* of a recording. For example, the deletion (or addition) of segments in a recorded testimony would potentially modify the meaning of phrases; therefore, such a modification must render the watermark undetectable. In contrast, lossy compression with reasonable bit rates should not introduce enough distortion to corrupt the significance of phrases in a testimony; in this case, the watermark should still be detectable.

Fingerprinting can also be used for integrity verification. The fingerprint of the original recording is extracted and stored. To verify integrity, a new fingerprint is extracted and compared with the original one; if they are identical, the recording has not been severely modified. The disadvantage of this approach is the need of storing the original fingerprint in a safe place, which adds both complexity and risk to the whole procedure.

Very fragile watermarks can also be used to verify whether a signal has been manipulated in any way, even without audible distortion. For example, a recording company can watermark the content of its CDs with a very fragile watermark. If songs from this CD are compressed in MP3 format, then decompressed and recorded in a new CD, the watermark would not be detected in the new recording, even if the latter sounds exactly as the original one to the listener. A CD player can then check for the presence of this watermark; if no watermark is found, the recording has necessarily undergone illicit manipulations and the CD is refused.

6.2 A combined watermarking/fingerprinting system

Robustness of integrity verification could be increased by combining watermarking and fingerprinting in a single system. First, the fingerprint of the original recording is extracted; this fingerprint, viewed as a sequence of bits, is then used as the information to be embedded into the signal through watermarking. This information is repeated as many times as possible in the watermarked signal, in order to minimize the possibility of error in the reconstruction of the original fingerprint from the watermark. As the watermark signal is weak, the watermarked recording should have the same fingerprint as the original recording. Thus, the integrity of a recording can be verified by extracting its fingerprint and comparing it with the original one (reconstructed from the watermark) (This method is currently under investigation at IUA-UPF and Paris V University (Gómez et al., 2002)).

7 Major differences and similarities

In this section, we summarize the major differences and similarities between audio watermarking and automatic music recognition.

7.1 Modification of the audio signal

Audio watermarking modifies the original audio signal by embedding a mark into it, whereas fingerprinting does not change the signal at all but rather analyzes it and constructs a hash (the fingerprint) uniquely associated with this signal. In watermarking, there is a trade-off between watermark power (and audibility) and detection performance. In fingerprinting, there is no such trade-off: the system “listens” to the music, constructs a description of it and searches for a matching description in its database.

7.2 Requirement of a fingerprint repository

A human listener can only identify a piece of music if he has heard it before, unless he has access to more information than just the audio signal. Similarly, fingerprinting systems require previous knowledge of the audio signals in order to identify them, since no information other than the audio signal itself is available to the system in the identification phase. Therefore, a musical knowledge database must be built. This database contains the fingerprints of all the songs the system is supposed to identify. During detection, the fingerprint of the input signal is calculated and a matching algorithm compares it to all fingerprints in the database. The knowledge database must be updated as new songs come out. As the number of songs in the database grows, memory requirements and computational costs also grow; thus, the complexity of the detection process increases with the size of the database.

In contrast, no database is required for detection in a watermarking system, as all the information associated with a signal is contained in the watermark itself. The detector checks for the presence of a watermark and, if one is found, it extracts the data contained therein. Hence, watermarking requires no update as new songs come out, and the complexity of the detection process is not changed when new audio signals are watermarked.

7.3 Requirement of previous processing

For several applications, the need of previously processing audio signals is a severe disadvantage of watermarking systems. For example, watermarking-based distribution-monitoring systems would only be able to detect copyright infringements if the copyrighted signals had been previously watermarked, which means that old non-watermarked material would not be protected at all. Additionally, new material would have to be watermarked in all its distribution formats, as even the availability of a small number of non-watermarked copies might compromise system security. This is not an issue for audio fingerprinting systems, since no previous processing is required.

7.4 Robustness

In watermark detection, the signal that contains useful information corresponds to a small fraction of the input power, as the watermark is much weaker than the original audio signal due to the inaudibility constraint. In addition, noise that might be added to the watermarked signal (by MP3 compression or analog transmission, for example) can be as strong, or even stronger, as the watermark. In case of severe channel perturbation or piracy attack, the watermark may no longer be detectable.

In contrast, detection in fingerprinting systems is based on the audio signal itself, which is strong enough to resist most channel perturbations and is less susceptible to piracy attacks. Such systems are thus inherently more robust. As long as the original audio in the knowledge database sounds approximately the same as the piece of music that the system is "listening" to, their fingerprints will also be approximately the same. The definition of "approximately" depends on the fingerprint extraction procedure; therefore, the robustness of the system will also depend on it. Most fingerprinting systems use a psychoacoustic front-end approach to derive the fingerprint. By doing so, the audio to analyze (and identify) can be strongly distorted with no decrease in system performance.

7.5 Independence between signal and information

The information contained in the watermark may have no direct relationship with the carrier audio signal. For example, a radio station could embed the latest news into the songs it airs through a watermark; at reception, the news would appear on a small screen while the songs are played. In contrast, a fingerprint is correlated with the audio signal from which it was extracted; any change in the audio signal that is perceivable to a human listener should cause a change in the fingerprint. This fact is behind most differences in applications between the two approaches: while watermarks can carry any kind of information, fingerprints always represent the audio signal.

This independence between signal and information is derived from the fact that watermarking systems only deal with information that has been previously added, given that no connection to a database is provided. This information can be either related or not with the audio signal in which it has been embedded. Fingerprinting can extract information from the audio signal at different abstraction levels, depending on the application and the usage scenario. The higher level abstractions for modeling the audio and thus the fingerprinting hold the possibility to extend the applications to content-based navigation, search by similarity and other applications of Music Information Retrieval.

8 Conclusions

We have presented two methodologies that have many applications in common and many specific ones as well. In audio watermarking we embed information into an audio signal. Although initially intended for copyright protection, watermarking is useful for a multitude of purposes, particularly for transport of general-purpose information. Audio fingerprinting does not add any information to the signal, since it uses the significant acoustic features to extracts a unique fingerprint from it. In conjunction with a database, this fingerprint can be used to identify the audio signal, which is useful in many applications (copyright-related or not).

While information retrieved from a database by means of a fingerprint is always related to a specific piece of music, information embedded into the signal by means of a watermark may be of any kind. Watermarking can even be used as a replacement (or a complement) for cryptography in secure communications. Watermarking has therefore a broader range of applications than fingerprinting. On the other hand, fingerprinting is inherently more robust than watermarking: while the fingerprint extraction procedure makes use of the full audio signal power, watermark detection is based on a fraction of the watermarked signal power (the watermark, which is several times weaker than the original audio signal due to the inaudibility constraint). This means that fingerprinting will resist distortion at higher levels than watermarking, which is a particularly attractive characteristic in copyright-related applications. When both techniques apply, robustness may be a strong argument in favor of fingerprinting. Another issue is that information does not need to be embedded in the audio in order to be identified. This allows tracking and identification of audio already released without watermarks in many different formats by presenting one example of the audio excerpt to the system. Assuming the cost of higher computational requirements and the need of a repository of the fingerprints, this approach represents a flexible solution for copyright-related and content related applications.

An important lesson has been (re)learned from recent research on audio watermarking: absolute protection against piracy is nothing more than an illusion. Sooner or later (probably sooner), pirates will find their way into breaking new protection schemes. The actual goal is to render piracy a less attractive (i.e., more expensive) activity, and to “keep honest people honest”. Neither fingerprinting-based protection systems shall claim absolute invulnerability.

Copyright-related applications are still central to the research on both watermarking and fingerprinting. However, recently-proposed added-value applications tend to become more and more prominent in the years to come.

Both technologies presented here have potential to respond to these needs. As their strong points are often complementary, their combined use could lead to interesting solutions.

Acknowledgments
The authors would like to thank Nicolas Moreau and Lazar Bibin for their participation in the research on audio watermarking described in this article. Leandro de C. T. Gomes would like to thank CNPq for financial support. The audio fingerprinting related work has been partially funded by the IST European project RAA (http://raa.joanneum.ac.at).

References

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

Audible Magic (2001). http://www.audiblemagic.com.

Auditude (2001). http://www.auditude.com.

Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Batlle, E. & Cano, P. (2000). Automatic Segmentation for Music Classification using Competitive Hidden Markov Models. *Proceedings Of International Symposium on Music Information Retrieval.*

Batlle, E., Nadeu, C., & Fonollosa, J. (1998). Feature Decorrelation Methods in Speech Recognition. A Comparative Study. *Proceedings of International Conference on Speech and Language Processing.*

Baytsp (2001). http://www.baytsp.com.

Bender, W., Gruhl, D., Morimoto, N., & Lu, A., (1996). Techniques for data hiding. *IBM System Journal* vol. 35, pp. 313-336.

Blum, T. L., Keislar, D. F., Wheaton, J. A., & Wold, E. H. (1999). *Method and Article of Manufacture for Content-Based Analysis, Storage, Retrieval and Segmentation of Audio Information*. USA patent number 5,918,223.

Boeuf, J. & Stern, J. P. (2001). *An analysis of one of the SDMI candidates.*

http://www.julienstern.org/sdmi/files/sdmiF/sdmiF.html.

Boneh, D. (1999). Twenty years of attacks on the RSA cryptosystem. *American Mathematical Society*.

Boney, L., Tewfik, A., & Hamdy, K. (1996). Digital watermarks for audio signals. *IEEE Proceedings Multimedia*.

Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., Dietz, M., Herre, J., Davidson, G., Oikawa, Y. (1997). ISO/IEC MPEG-2 Advanced Audio Coding. *Journal of the AES*, Vol. 45, No. 10, pp. 789-814.

Cano, P., Kaltenbrunner, M., Mayor, O., & Batlle, E. (2001). Statistical Significance in Song-Spotting in Audio. *Proceedings International Symposium on Music Information Retrieval*.

Craver, S. A, Wu. M., & Liu, B. (2001). What Can We Reasonably Expect From Watermarks?. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Craver, S. A. (2001). Reading between lines: lessons from the SDMI challenge. *Proceedings of the 10th USENIX Security Symposium*.

Dannenberg, R., Foote, J., Tzanetakis, G., & Weare, C. (2001). Panel: New Directions in Music Information Retrieval. *Proceedings of the International Computer Music Conference*.

Dixon, R. (1976). *Spread-spectrum systems*. John Wiley & Sons.

Etantrum (2001). http://www.etantrum.com, http://sourceforge.net/projects/freetantrum/

Furon, T., Moreau , N., and Duhamel, P. (2000). Audio public key watermarking technique. *Proceedings of the ICASSP*.

Garcia, R. A. (1999). Digital watermarking of audio signals using a psychoacoustic auditory model and spread-spectrum theory. *107th AES Convention*.

Gnutella news (2002). http://www.gnutellanews.com.

Gnutella wego (2002). http://gnutella.wego.com.

Gomes, L. de C. T., Gómez, E., and Moreau, N. (2001). Resynchronization methods for audio watermarking. *111th AES Convention*.

Gomes, L. de C. T., Mboup, M., Bonnet, M., & Moreau , N. (2000). Cyclostationarity-based audio watermarking with private and public hidden data. *109th AES Convention*.

Gómez, E., Cano, P., Gomes, L., Batlle, E., & Bonnet, M. (2002). Mixed Watermarking-Fingerprinting Approach for Integrity Verification of Audio Recordings. To appear in *Proceedings of IEEE International Telecommunications Symposium. Natal, Brazil*

Gracenote (2001). http://www.gracenote.com.

Haitsma, J.A., Kalker T., Oostveen, J. (2001). Robust Audio Hashing for Content Identification. *Second International Workshop on Content Based Multimedia and Indexing*.

Haykin, S. (1988). *Digital Communications*. Prentice Hall.

Haykin, S. (1996). *Adaptive Filter Theory*. Prentice Hall.

Herre, J., Allamanche, E., & Helmuth, O. (2001). Robust Matching of Audio Signals Using Spectral Flatness Features. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

Id3man (2001). http://www.id3man.com.

International Federation of the Phonographic Industry (2002). http://www.ifpi.org.

International Organization for Standardization (1997). ISO/IEC 13818-7 (MPEG2 Advanced Audio Coding, AAC).

Ismirli, O. (2000). Using a Spectral Flatness Based Feature for Audio Segmentation and Retrieval. *Proceedings of the International Symposium on Music Information Retrieval*.

Lacy, J., Quackenbush, S., Reibman, A., Shur D., & Snyder, J. (1998). On combining watermarking with perceptual coding. *Proceedings of the International Conference on Acoustic Speech and Signal Processing*, Seattle.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modelling. *Proceedings of the International Symposium on Music Information Retrieval*.

Loscos, A., Cano, P., & Bonada, J. (1999). Low-Delay Singing Voice Alignment To Text. *Proceedings of the International Computer Music Conference*.

Mihçak, M. K. & Venkatesan, R. (2001). A Perceptual Audio Hashing Algorithm: A Tool For Robust Audio Identification and Information Hiding. *4th Workshop on Information Hiding*.

Moodlogic Inc. (2001). http://www.moodlogic.com.

MPEG Working Documents. http://www.cselt.it/mpeg/working_documents.htm.

Music Reporter (2001). http://www.musicreporter.net.

Musicbrainz (2001). http://www.musicbrainz.org.

Napster (2002). http://www.napster.com.

Neubauer, C. & Herre, J. (2000). Audio watermarking of MPEG2 AAC bit stream. *108th AES*

Convention, Paris.

Oppenheim, A. V. & Schafer, R. W. (1989). *Discrete-Time Signal Processing*. Prentice Hall.

Packman, N. & Kurth, F. (2000). Transport of Content-based Information in Digital Audio Data. *109th AES Convention*.

Perreau Guimarães, M. (1998). Optimisation de l'allocation de ressources binaires et modelisation psychoacoustique pour le codage audio. PhD Thesis, Université Paris V.

Petitcolas, F. A. P., Anderson, R. J., & Kuhn, M. G. (1999). Information hiding --- a survey. *Proceedings IEEE*, special issue on protection of multimedia content, 87(7):1062-1078.

Rabiner, L. R. (1998). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceeding of the IEEE*.

Recording Industry Association of America (2001). http://www.riaa.com/.

Request for Audio Fingerprinting Technologies (2001). News (June 15) http://www.riaa.org/

Ruggero, M. A., (1992). Physiology and Coding of Sounds in the Auditory Nerve. from "*The Mammalian Auditory Pathway: Neurophysiology*", Springer-Verlang.

SDMI (2001). http://www.sdmi.org.

TRM (2001). http://www.relatable.com.

Viterbi, A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Informational Theory*, vol. 13 (2).

Wu, M., Craver, S. A., Felten, E. W., Liu, B. (2001). Analysis of Attacks on SDMI Audio Watermarks. *Proceedings of the ICASSP*.

Zwicker, E. & Fastl, H. (1990). *Psychoacoustics, Facts and Models*. Springer Verlag.

Accepted for publication on May 9th 2002