



An Audio Watermarking Method Based On Molecular Matching Pursuit

Mathieu Parvaix¹, Sridhar (Sri) Krishnan², Cornel Ioana³

¹: Institute of Speech Communication, Polytechnic Institute of Grenoble,
46 Av. Félix Viallet, 38000 Grenoble – FRANCE, E-mail : Mathieu.Parvaix@gipsa-lab.inpg.fr

²: SAR Laboratory, Dept. of Electrical and Computer Engineering, Ryerson University,
245 Church Street, Toronto – CANADA, E-mail : krishnan@ee.ryerson.ca

³: Gipsa-Lab, Dept. of Images and Signals, Polytechnic Institute of Grenoble,
38402, Saint Martin d’Heres – FRANCE, E-mail : Cornel.Ioana@lis.inpg.fr

ABSTRACT

In this paper we introduce a new watermarking model combining a joint time frequency (TF) representation using the molecular matching pursuit (MMP) algorithm and a psychoacoustic model. We take advantage of the notion of structure of the signal introduced by the MMP to get a precise representation of audio signals, and then by using a psychoacoustic model we can embed a watermark efficiently on the signal. By selecting atoms of TF components that are not perceptible by the human ear we ensure the security and imperceptibility of the watermark. Then by judicious selection of the watermark host spots we ensure the robustness of the watermark to main kind of signal attacks, including lossy compression. The robustness of the proposed method proves the potential of joint TF representation techniques as viable watermarking schemes.

Index Terms — audio coding, time-frequency analysis, matching pursuits, human auditory system.

1. INTRODUCTION

We live in a digital world where multimedia files are omni-present and ever increasing. The development of the Internet brought with it the possibility to share tremendous amounts of audio or video files. But in the same time, it has become easy to illegally copy and distribute the multimedia files. Each time a file is downloaded from a peer to peer server, there is possibility of copyright violation. Since the past few years majority of entertainment industries have to face considerable loss of revenue. The interest to protect the copyright of multimedia content became obviously indispensable. To ensure this protection the watermarking technology appears as a viable tool, but beyond the simple

proof of ownership it provides, it can also be used to embed data within multimedia content. To guarantee efficient protection the watermark should satisfy three main constraints and should be :

- imperceptible i.e. inaudible in the case of audio files
- recoverable to prove the ownership of the watermarked file
- resistant to all kind of signal manipulations.

The method we propose in this paper will strive to respect those three essential requirements.

To respect the first assumption of efficiency, a watermark embedded on a signal should not decrease the audio quality of this signal. To ensure this inaudibility we will try to benefit by taking advantage of the human auditory system and use the phenomena of temporal and frequency masking. Using these masking effects implies an accurate knowledge of both time and frequency characteristics of the signal we want to watermark. For this reason we decided to implement the decomposition algorithm developed by *Daudet* in [1] named Molecular Matching Pursuit (MMP). This algorithm showed very promising results for audio signals presenting fast variations. The TF information we get after MMP decomposition will be combined with a psychoacoustic model to embed the watermark on parts of the signal which can be sparingly modified without being perceptible by the human ear.

This paper is organized as follows. In Section 2. a description of the MMP and its application in audio signal processing is presented. We present the results of the embedding and recovering schemes in Section 3. Finally, the conclusions are presented in Section 4.

2. Methodology

A. Molecular Matching Pursuit

The first assumption we make is that an audio signal can be seen as the sum of two structures: a tonal part composed by sinusoids varying slowly both in amplitude and frequency, and a transient part represented as narrow picks located at the attack of notes. The MMP algorithm is a signal decomposition algorithms derived from the famous Matching Pursuit (MP) developed by *Mallat* and *Zhang* [2]. It enables to decompose a complex signal whose time and frequency properties are unknown in a linear combination of well known elementary signals gathered in dictionaries. If we consider a signal x of dimension N , we will use, to decompose it, a two-times redundant dictionary $D = C \cup W$, where $C = \{c_i\}_{i=1\dots N}$ is an N -element dictionary composed of orthogonal Modified Discrete Cosine Transform (MDCT) coefficients [3] and $W = \{w_j\}_{j=1\dots N}$ another N elements dictionary composed of Discrete Wavelet Transform (DWT) coefficients. The MDCT dictionary will permit to decompose the tonal part and the DWT one will be used to reconstruct the transient part of the signal.

The main difference between the MMP algorithm and the classical MP (and the derived algorithms [4]) is that the MMP uses the structural information of the signal. It aims at grouping atoms that have TF similarities, in other words atoms located in a neighboring area in the time-frequency plane. This group of atoms is called molecule. We will distinguish the Tonal and the Transient molecules.

The MMP algorithm is described as follows:

(1) Initialization: compute all the inner products between the initial signal x and the elements of the dictionaries C and W . We obtain a matrix of MDCT coefficients and one of DWT coefficients. The MDCT coefficients are computed with overlapping windows of 256 samples. The residual is equal to x .

(2) Threshold the MDCT coefficients according to an energy criterion: we keep coefficients that represent 99% of the energy of the residual.

(3) Selection of the wavelet subspace that minimizes the entropy to find the DWT coefficients close to the transient part of the signal. We only keep the corresponding DWT coefficients.

(4) We keep among those coefficients the one representing 99% of the energy of the corresponding subspace.

(5) Gathering the main MDCT coefficients located in the same time-frequency area by applying on all the MDCT coefficients a time-frequency window centered on the most significant MDCT coefficient. They constitute the Tonal molecule.

(6) Gathering the DWT coefficients, among the ones selected at stage (4), close to the tonal molecule by windowing. They constitute the transient molecule.

(7) Subtraction of the tonal and the transient molecules to the residual.

The MDCT coefficients are computed using the Fast Fourier Transform for only $N/4$ points which allows an $N \log(N)$ complexity.

The effectiveness of MMP algorithm for music signal representation has been demonstrated. Figure 1 shows the results of reconstruction on a test signal of 5 seconds duration sampled at 44.1 kHz. As this signal is extracted from a rock music piece it presents both tonal part and strongly marked transient part.

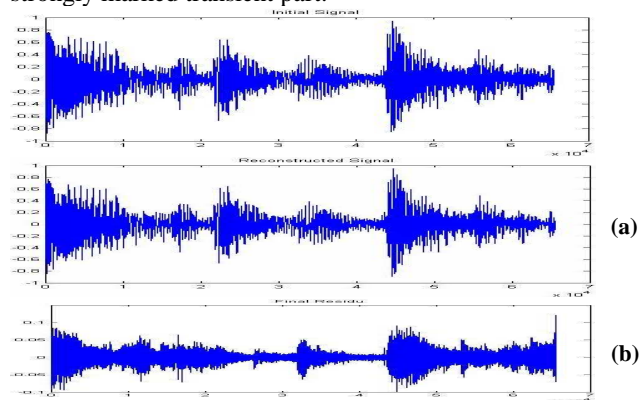


Fig. 1. (a) Reconstructed signal (tonal + transient part) after 315 iterations of the MMP algorithm. (b) Residual (note the different scale)

The following table shows the results obtained for 6 5-seconds music signals, each of them sampled at 44.1kHz which represents 220500 samples by signal. We note that pop and rock signals that possess the largest number of transients are more difficult to reconstruct.

	Class	Count	Folk	Jazz	Pop	Rock
Energy	99.43	98.12	98.76	98.53	96.93	96.59
Correlation	0.9972	0.991	0.9938	0.9927	0.9846	0.9828

Fig. 2. Results after 300 iterations of the MMP algorithm showing the energy of the reconstructed signal and its correlation compared to the initial signal's ones.

B. The Psychoacoustic Model

The TF characteristics provided by the MMP algorithm should be combined with psychoacoustic model for watermark representation.

Time and Frequency Masking

To introduce a watermark on an audio signal we will

first take advantage of the time masking effect. If one sound occurs at the time t , during the 100 to 200ms following t , a sound of lower amplitude located under the post-masking curve will not be audible. A low amplitude sound occurring until 15ms before t , if located under the pre-masking curve will also be not heard.

We also use the frequency masking effect: if a sound $S1$, perfectly audible if emitted alone, occurs at the exact same time than a sound $S2$ of close frequency and of greater amplitude than $S1$ then the sound $S1$ will be inaudible.

The global threshold of hearing

To determine the analytic expression of the masking spreading due to both of these masking effects we use an existing model [4]. But first we have to determine if a frequency of the signal is a tone or a noise. We use the following model:

A frequency f is considered as a tone (we consider the k^{th} element of its FFT) if its power $P[k]$ is

- greater than $P[k-1]$ and $P[k+1]$
- 7 dB greater than the other frequencies in its neighborhood, this neighborhood depending on the frequency f itself:

f	Neighbourhood
$0.17 \text{ Hz} < f < 5.5 \text{ kHz}$	$[k-2\dots k+2]$
$5.5 \text{ kHz} \leq f < 11 \text{ kHz}$	$[k-3\dots k+3]$
$11 \text{ kHz} \leq f < 20 \text{ kHz}$	$[k-6\dots k+6]$

If a frequency is not a tone then it is considered as a noise. The general idea to determine the noise maskers is to take all frequency components within a critical band that do not fit within tone neighborhoods, add them together, and place them at the geometric mean location within the critical band. Repeat this for all critical bands from 20Hz to 20KHz.

The masking effects can be described as a function of the location i of the masker, the one j of the masked frequency, of the power spectrum P in j , and of the location difference $\Delta = z(i) - z(j)$ between the masker and the masked frequencies. SF is the time masking effect depending on Δ . The tone threshold is :

$$T_{\text{tone}}(i, j) = P(j) - 0.275 \times z(j) + SF(i, j) - 6.025 \text{ dB} \quad (1)$$

The noise threshold is :

$$T_{\text{noise}}(i, j) = P(j) - 0.175 \times z(j) + SF(i, j) - 2.025 \text{ dB} \quad (2)$$

Where:

$$SF(i, j) = \begin{array}{ll} 17 \times \Delta - 0.4 \times P(j) + 11 & -3 \leq \Delta < -1 \\ (0.4 \times P(j) + 6) \times \Delta & -1 \leq \Delta < 0 \\ -17 \times \Delta & 0 \leq \Delta < 1 \\ (0.15 \times P(j) - 17) \times \Delta - 0.15 \times P(j) & 1 \leq \Delta < 8 \end{array}$$

Once this global threshold of hearing has been computed for the entire signal we keep only MDCT and DWT coefficients located under this threshold as the host coefficients where the watermark will be embedded. Sparsely modifying those coefficients will, indeed, not be audible.

C. The Embedding Process

We chose to embed as a message a 1-bit quantized chirp (its samples are either +1 or -1) whose length permits to embed information such as the identity of the file owner. This watermark is inserted in a binary form. If the code used to encode the copyright information is a 6-bit code then a name of 10 alphanumeric characters needs 60 bits to be embedded [6],[7].

To insure more security and inaudibility to the embedded watermark, different methods can be used. We use two classical ones: the spread spectrum and an amplitude modulation. The spread spectrum consists in multiplying each sample of the watermark with a secret key sequence only known by the person embedding the watermark. We use in this paper a 20 bits random sequence to spread the watermark on the spectrum. The mean of this sequence is positive (it will be used in the recovery process).

Then we make an amplitude modulation on the watermark to ensure that it varies in the same proportions than the signal and thus can remain under the global threshold of hearing. To compute the modulation coefficient, for each MDCT host coefficient we compute the mean of the MDCT coefficients of the initial signal in the time window he belongs to. If we call y the watermarked signal, b the message to embed, p the secret key and λ the modulation factor for this specific time window, then the k^{th} sample of the watermark is:

$$y_k = x_k + \lambda \cdot b_k \cdot p \quad (3)$$

D. Recovering the Watermark

We use to recover the watermark a non-blind algorithm, which means that we need the original non watermarked signal to recover the message inserted. Basically the process to recover the watermark is the reverse of the embedding process:

1. Decompose the watermarked and the initial signals in MDCT/DWT coefficients using the MMP.
2. Determine the host coefficients of the initial signal using the psychoacoustic model previously presented and keep the corresponding coefficients among the watermarked signal's ones.
3. Subtract term to term the non watermarked host coefficients to the watermarked ones. We obtain the term $\lambda \cdot b_k \cdot p$ of the expression (3).

4. Window the signal obtained at step 3. in 20 samples parts and compute the mean of each part. If the mean is positive then the corresponding sample of the embedded message is +1 and if the mean is negative then the corresponding sample of the message is -1.

3. Tests results and discussion

A. Imperceptibility of the Watermark.

Once this embedding process has been realized we checked the first property a watermark has to verify, its inaudibility, by listening tests. Five testers have been asked to give a mark a five levels scale to 6 watermarked signals to compare their quality to the original audio quality. The five levels are 5: imperceptible, 4: perceptible but not annoying, 3: slightly annoying, 2: annoying, 1: very annoying.

	Signals					
	Class	Country	Folk	Jazz	Pop	Rock
Mean of 5 marks	4.8	5	4.8	5	5	5

Fig. 3. Results of listening tests on watermarked signals.

B. Watermark Recovery

The different attacks we used are:

- Low Pass Filtering: the signal is filtered using a FIR lowpass filter with normalized cutoff frequency 0,025.
- Addition of Noise: we add a white noise to the audio signal to obtain a 30 dB SNR.
- Inversion: inversion of audio samples
- MP3 compression: we encode the .wav watermarked files into .mp3 format with an MPEG1 layer3 encoder and then decode it.

Figure 4 shows the results of the recovering process without any attack and after attacks such as Low Pass Filtering, noise addition, inversion, and mp3 compression for different bit rates.

Files / Attacks	No attack	LPF	Noise	Inver-sion	Mp3 Compression	
					96Kbits/s	128Kbits/s
Class	0.0	37.5	20.0	0.0	0.0	0.0
Count	0.0	25.0	0.0	0.0	0.0	0.0
Folk	0.0	29.8	12.2	0.0	0.0	0.0
Jazz	0.0	40.2	8.8	1.96	1.96	1.96
Pop	0.0	37.1	19.0	0.0	0.0	0.0
Rock	0.0	38.1	0.0	0.0	0.0	0.0

Fig. 4. BER of watermark recovering without and with attacks.

We can notice that without any attack the message embedded is fully recovered. Concerning the different

attacks, we obtain really good results with a bit error rate between the recovered and the inserted message inferior to 2% for inversion and mp3 compression. The low pass filtering gives the less good results, with a BER varying depending on the type of audio signal.

We can improve the detection results by a multi watermarking scheme. We embedded the same message 4 times in a test signal and after computing the median of the 4 extracted messages we have been able to fully recover the inserted message. Of course such a technique is more greedy in terms of host coefficients (4 same messages of 15 bits inserted a 5 seconds audio signal instead of a 60 bits one) but can easily be applied on a full audio signal of a few minutes.

4. Conclusion

In this study we proposed a new watermarking scheme combining a novel decomposition algorithm, the MMP, and a psychoacoustic model to increase the efficiency of the watermark insertion. A message is inserted in the time frequency and time scale domains on Modified DCT and DWT coefficients.

The experimental results show a perfect imperceptibility of the inserted watermark after listening tests explained by the reconstruction of the signal mainly due to the MMP and the choice of embedding spots regarding to the defaults of the auditory system. Good results to different kinds of attacks can also be noticed, especially mp3 compression and noise addition. A multi-watermarking scheme permits to improve the robustness of the inserted message.

REFERENCES

- [1] Laurent Daudet, "Sparse and Structured Decompositions of Signals With the Molecular Matching Pursuit", IEEE Transactions On Audio, Speech, And Language Processing , Vol. 14, NO. 5, Sept. 2006
- [2] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," IEEE Trans. Signal Processing, Vol. 41, NO. 12, pp. 3397-3415, Dec. 1993.
- [3] Yun-Hui Fan; Madisetti, V.K.; Mersereau, R.M., "On fast algorithms for computing the inverse modified discrete cosine transform", Signal Processing Letters, IEEE, Vol. 6, Issue 3, pp.61-64, Mar. 1999
- [4] Rémi Gribonval and Emmanuel Bacry, "Harmonic Decomposition of Audio Signals With Matching Pursuit", IEEE Transactions On Signal Processing, Vol. 51, NO. 1, Jan. 2003
- [5] Davis Pan, "A Tutorial on MPEG/Audio Compression", IEEE Transactions on Multimedia, Vol. 2, Issue 2, pp.60 - 74, Jun. 1995
- [6] Serhat Erkuçut, Sridhar Krishnan, and Mehmet Zeytinoglu, "A Robust Audio Watermark Representation Based on Linear Chirps", IEEE Transactions on Multimedia, Vol. 8, Issue 5, pp. 925 - 936, Oct. 2006
- [7] Foo Say Wei, Xue Feng and Li Mengyuan, "A Blind Audio watermarking Scheme Using Peak Point Extraction", Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium, pp. 4409 - 4412, Vol. 5, May 2005