

# INFORMED SPECTRAL ANALYSIS FOR ISOLATED AUDIO SOURCE PARAMETERS ESTIMATION

*Dominique Fourer and Sylvain Marchand*

LaBRI CNRS

University of Bordeaux 1, 33405 Talence, France

firstname.lastname@labri.fr

## ABSTRACT

In this paper, we propose a new watermark-assisted method for the analysis of audio source signals present in a mixture. This work is motivated by the issue of quality-constrained source parameters estimation in under-determined mixtures where the blind approach is not efficient. Our method uses a specific coder-decoder configuration where the separated source signals are assumed to be known at the coder. The necessary information required by a classic blind estimator to reach a target quality is imperceptibly embedded into the mixture signal. At the decoder, where the isolated source signals are unknown, the analysis process is assisted by the extra information embedded into the mixture signal. Thus, this technique aims at opening new perspectives for quality-based audio applications like active listening, sound feature extraction, or high quality audio effects with a minimal amount of side information.

**Index Terms**— Informed spectral analysis, source separation, sinusoidal model

## 1. INTRODUCTION

Sound source separation is full of interest in audio processing and can allow a listener to manipulate each sound entity present in the auditory scene. The under-determined case (e.g. the number of sound entities is greater than the number of their observed mixtures) is a particularly difficult configuration. However, it is the most frequent case for mono or stereo music mixtures. This issue is often processed using sparse representations of signals [1] but the quality is not sufficient for demanding applications. Recently, Informed Source Separation (ISS) [2] proposed to embed source indexes inaudibly into the mixture signal to assist the separation process. Thus, ISS achieves to reach a better quality of separation but does not use the full potential of available information present in the mixture. The method presented here is based on informed sinusoidal parameters estimation, and allows us to recover both the source signals and their model parameters, with a desired quality. The sinusoidal model offers a sparse representation of audio signals and is suitable for music. It has shown its efficiency for representing audio signals at low bit rates (typically lower than 24 kbits/s for MPEG-4 SSC [3]) and allows sound transformations like time stretching or pitch shifting with a high quality. Efficient estimators exist for the sinusoidal model [4] but have theoretical limitations on the best achievable estimation precision. As discussed in a previous work in [5], the unique way to break these theoretical bounds consists in injecting extra information into the analysis chain. With our approach, the information provided by a classic estimator is used to reduce the amount of this extra information. In the context where

the extra information is directly embedded into the signal mixture using a QIM-based watermarking method [6], differential coding is not applicable. In fact, the signal mixture is altered during the mixing and watermarking processes and depends directly on the extra information itself. This paper proposes a solution to this issue and is organized as follows. Section 2 describes the principles of informed analysis, while its implementation is detailed in Section 3. Experiments and results for simulated and natural sounds are presented in Section 4. Section 5 concludes with discussions and future works.

## 2. INFORMED ANALYSIS FRAMEWORK

The presented method (see Fig. 1) is designed in a specific coder-decoder configuration. At the coder stage, a classic estimator is applied to each isolated source signal  $s_k[n]$  to obtain reference parameters  $P_k$ . After the mixing process, the same estimator is applied to the mixture signal  $x[n]$  to simulate the analysis process at the decoder. The necessary information needed to recover  $P_k$  from  $x[n]$  using the selected classic estimator is estimated with the generalized informed spectral analysis technique detailed below. This section describes the signal model and the informed analysis principle applied to the parameters estimation at the coder and the decoder stages.

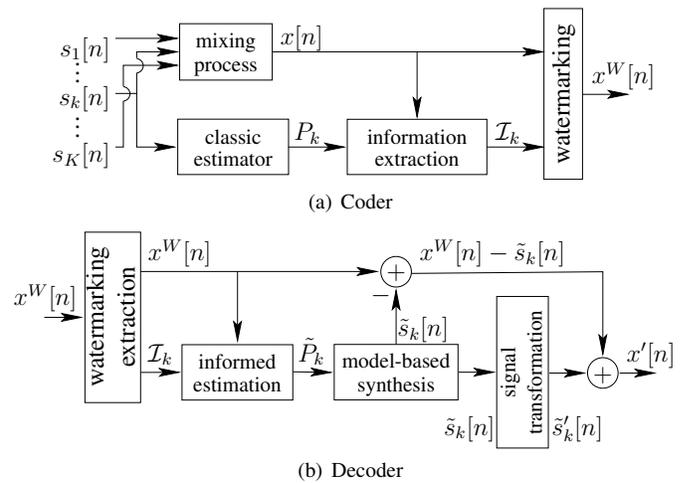


Figure 1: Structure of the proposed system for informed single-source manipulation in under-determined sound mixtures.

## 2.1. Sound source model and parameters estimation

Consider a discrete instantaneous single-channel mixture signal composed of  $K$  sources which can be expressed as follows:

$$x[n] = \sum_{k=1}^K s_k[n] + r[n] \quad (1)$$

where  $r[n]$  is the residual signal. Each source signal  $s_k[n]$  is assumed to fit the stationary sinusoidal model [7] for each local analysis frame written as:

$$s_k[n] = \sum_{l=1}^L a_l \cos(\omega_l n + \phi_l) \quad (2)$$

where  $a$ ,  $\omega$ , and  $\phi$  correspond respectively to the amplitude, frequency, and phase parameters. constant. For the analysis process, the instantaneous parameters are estimated using a classic frame-based estimator. As discussed in [5], efficient estimators like the spectral reassignment or the derivative method [4] are suitable for informed spectral analysis. In fact, these estimators achieve to reach the theoretical bounds and tend to minimize the bit rate required by the informed spectral analysis.

## 2.2. Generalization of informed spectral analysis

Informed spectral analysis first introduced in [5] for single parameter estimation (of the frequency) consists in a two-step analysis where the information is first extracted using the knowledge about the distribution of the estimation error resulting from a classic (not informed) analysis. Second, the same estimator is applied on the altered signal and the errors are systematically corrected using the previously extracted information. This approach assumes that the reference parameters are exactly known and available for the information extraction at the first step.

### 2.2.1. Single parameter informed analysis

Suppose we have to estimate a real parameter  $p \in [0, 1)$  related to the signal  $s_k[n]$ . The mixture signal  $x[n]$  is created according to (1) including the signal  $s_k[n]$  combined with the others sources plus noise. The information needed to recover  $p$  from a classic estimation  $\hat{p}$  obtained from  $x[n]$  is extracted as follows.

First, we define  $C_d : [0, 1) \rightarrow \{0, 1\}^d$  the  $d$ -bit precision fixed-point binary coding application and  $\mathcal{D}$  the decoding application.  $C = (C_1, C_2, \dots, C_d)$  denotes the representation of  $p$  and  $\tilde{p} = \mathcal{D}(C)$  is the  $d$ -bit precision value of  $p$  (see [5] for details).

Second,  $I_\sigma$  is estimated as the Most Significant Bit (msb) of the upper bound of the absolute value of the error  $|p - \hat{p}|$ . In practice,  $I_\sigma$  is estimated with a significant number of occurrences over  $\hat{p}$  using the same classic estimator for a given noise variance (resulting from mixing and watermarking processes).

Thus,  $C_d(p)$  is separated respectively in a reliable part (the useful information provided by the classic estimator) and an unreliable part as we have:

$$C_d(p) = \underbrace{C_1, C_2, \dots, C_{I_\sigma-1}}_{\text{reliable part}}, \underbrace{C_{I_\sigma}, \dots, C_d}_{\text{unreliable part}}. \quad (3)$$

Using the error correction algorithm detailed in [5],  $\hat{p}$  can be recovered from any estimated value  $\tilde{p}$  if  $I_\sigma$  verifies  $I_\sigma \leq \text{msb}(C(|p - \tilde{p}|))$ , using the extra information:

$$\mathcal{I} = C_{I_\sigma-1}, C_{I_\sigma}, \dots, C_d. \quad (4)$$

This algorithm substitutes the unreliable part of  $C(\hat{p})$  with  $\mathcal{I}$  and requires the bit value at position  $I_\sigma - 1$  to solve eventual carry or exception problems. As explained in the introduction, the exact value of  $\hat{p}$  estimated at the decoder is unknown at the coder due to the watermarking process dependent on the extra information embedded. Thus, a classic closed-loop differential predictive coder [8] is not applicable in this particular configuration.

### 2.2.2. Entire sinusoidal model informed analysis

Consider now we have to estimate  $P = (a, \omega, \phi)$  a vector of  $\mathbb{R}^3$ . As  $a$ ,  $\omega$ , and  $\phi$  have different physical meaning, they require a different relative accuracy in order to minimize a defined distortion measure.

First,  $P$  is optimally vector quantized using Entropy Constrained Unrestricted Spherical Quantization (ECUSQ) [9] which minimizes the expected value of the squared error between synthesized signals according to (2). The overall bit budget  $d$  affected to  $P = (a, \omega, \phi)$  depends on the value of each vector component and results in a variable bit rate for a fixed entropy  $H_t$  (e.g. if  $a \approx 0$ , we need to allocate bits neither to phase nor to frequency). The function  $b_{a,\omega,\phi}(d)$  which returns the number of bits allocated to each vector component of  $P$  for a maximal overall bit budget  $d$  is  $\lceil \log_2 \gamma \rceil$  where  $\gamma$  is the point density (given by ECUSQ) and  $\lceil \cdot \rceil$  denotes the ceiling function.

Second, informed spectral analysis is applied separately on each vector component of  $P$  which can be processed as for the single parameter case. The coding application  $C_d : [0, 1)^3 \rightarrow \{0, 1\}^d$  uses a simple concatenation and is written as:

$$C_d(P) = C_e(a), C_f(\omega), C_g(\phi) \text{ with } e + f + g = d \quad (5)$$

where  $e = b_a(d)$ ,  $f = b_\omega(d)$ , and  $g = b_\phi(d)$ . Thus, the extra information is  $\mathcal{I} = \mathcal{I}_a, \mathcal{I}_\omega, \mathcal{I}_\phi$ .

For the decoding process,  $I_\sigma$  and  $d$  are necessary and sufficient to extract the information corresponding to each vector component. In fact, [9] shows that the optimal quantizer of  $\phi$  and  $\omega$  depends on the value of the amplitude.  $\tilde{a}$  is computed first using  $b_a(d)$ , then  $b_\omega(d)$  and  $b_\phi(d)$  can be calculated by ECUSQ. This point is detailed in Section 3.

## 3. IMPLEMENTATION

### 3.1. Overall algorithm

The complete method presented in Fig. 1 is implemented for the coder and decoder according to Algorithms 1 and 2. The target quality is fixed and supposed known at the coder and the decoder.

Each  $I_{\sigma,k,l}$  can be vector quantized while the unquantized value verifies the inequality described in Section 2.2.1 for  $I_\sigma$  on each vector component.

The binary mask denoted  $\text{mask}[n]$  and used by the classic frame-based estimator can often be omitted or at least transmitted with a negligible bit rate. In fact, classic peak picking can give a reliable approximation of the binary mask. Furthermore, as sinusoidal model is a sparse representation, the mask can be efficiently compressed.

### 3.2. Quantization and coding

ECUSQ gives the analytic expression of the sinusoidal parameters optimal quantizer using the high-resolution assumption (the error is uniformly distributed on each quantization cell). Thus, for a target

**Algorithm 1** Coder**input:**  $s_k[n]$ : isolated source signals**output:**  $x^W[n]$ : watermarked mixture

- Estimate  $P_{k,l}$  from  $s_k[n]$  using the reassignment method [4].
- Compute  $b_{a,\omega,\phi}$  from  $P_{k,l}$  using the ECUSQ method [9].
- Compute binary mask $[n]$  containing the support of detected peaks in the discrete amplitude spectrum.
- Estimate  $I_{\sigma,k,l}$  and  $\mathcal{I}_{k,l}$  from  $\hat{P}_{k,l}$  using the informed spectral analysis method (see Section 2) with simulated mixing process according to (1) combined with watermark [10].
- Compute  $x^W[n]$  using [10] containing mask $[n]$ ,  $I_{\sigma,k,l}$  and  $\mathcal{I}_{k,l}$ .

**Algorithm 2** Decoder**input:**  $x^W[n]$ : watermarked mixture**output:**  $\tilde{s}_k[n]$ ,  $\tilde{P}_{k,l}$ : isolated source signals and parameters

- Recover mask $[n]$ ,  $I_{\sigma,k,l}$  and  $\mathcal{I}_{k,l}$  from watermark extraction from  $x^W[n]$  using [10] and [9] for  $b_{a,\omega,\phi}$  computation.
- Estimate  $\hat{P}_{k,l}$  using mask $[n]$  and the reassignment method [4].
- Compute  $\tilde{P}_{k,l}$  with  $I_{\sigma,k,l}$  and  $\mathcal{I}_{k,l}$  using the informed spectral analysis (see Section 2).
- Synthesize  $\tilde{s}_k[n]$  from  $\tilde{P}_{k,l}$  according to (2).

entropy (average information with variable bit rate) related to a fixed resolution (constant bit rate), the average distortion measure is minimized. The unrestricted term refers to the fact that the parameters are dependently quantized, which outperforms the restricted spherical quantizers. The ECUSQ point density functions are expressed as follows (see [9] for details):

$$g_a(a, \omega, \phi) = 2^{(1/3)\tilde{H}_t - 2\beta(A) - \log_2(\sigma_w)}, \quad (6)$$

$$g_\phi(a, \omega, \phi) = ag_a(a, \omega, \phi), \quad (7)$$

$$g_\omega(a, \omega, \phi) = \sigma_w ag_a(a, \omega, \phi), \quad (8)$$

with  $\tilde{H}_t = H_t - h(A)h(\Omega)h(\Phi)$  where  $H_t$  is the target entropy and  $h(\cdot)$  denotes the entropy of each parameter probability distribution respectively denoted  $f_A(a)$ ,  $f_\Omega(\omega)$ , and  $f_\Phi(\phi)$ . We define  $\beta(A) = \int f_A(a) \log_2(a) da$  and  $\sigma_w^2 = \frac{1}{\|w\|^2} \sum_{n=n_0}^{n_0+N-1} w(n)n^2$ . Here,  $w(n)$  denotes the analysis window of size  $N$ . The quantization step sizes are given by the reciprocal values of the point densities  $\Delta = g^{-1}$ . Each quantization point is located in the middle of the corresponding quantization cell. The first quantization point is chosen to be 0.  $g_\omega$  and  $g_\phi$  depend linearly on the amplitude and are computed using  $\tilde{a}$ . The number of bits allocated to each parameter for fixed-point binary coding is 0 if  $g \leq 1$  and  $\lceil \log_2(g) \rceil$  elsewhere.

**3.3. Watermarking process**

The extra information is inaudibly embedded into the mixture using the method described in [10]. This technique inspired from the Quantization Index Modulation (QIM) [6] is based on Modified Discrete Cosine Transform (MDCT) coefficients quantization. We selected this method for the large capacity provided, higher than 200kbits/s for 16-bit PCM signals, and for its high resulting quality.

The embedded extra information is designed for the frame-based spectral analysis at the decoder stage and thus is well suited for real-time processing.

**4. EXPERIMENTAL RESULTS****4.1. Simulation**

Consider here a discrete-time signal  $s[n]$  sampled at  $F_s = 44.1$ kHz composed of 1 sinusoid generated according to (2). A white Gaussian noise of fixed variance is mixed with  $s[n]$  in order to result a SNR from  $-20$ dB to  $50$ dB. Phase follows the uniform density function  $U(0, 2\pi)$ . Amplitude and frequency follow a Rayleigh distribution respectively of parameter  $\sigma_a = 0.2$  for  $a \in [0, 1]$  and  $\sigma_\omega = \pi/11$  for  $\omega \in [0, \pi]$ . The analysis frame uses the Hann window of odd length  $N = 1023$  with estimation time set at the center. The target entropy  $H_t$  is calculated from the ECUSQ rate-distortion function [9] for a target SNR set at  $45$ dB. For each generated signal,  $I_\sigma$  is estimated using the knowledge about the initial SNR uniformly vector quantized with 4 bits on the  $[-20$ dB,  $\text{SNR}^{\text{target}}$ ] interval. Fig. 2(b) shows that the extra-information bit rate required to reach  $\text{SNR}^{\text{target}}$  decreases when the precision of the classic estimator increases. As shown on Fig. 2(a), the exact target SNR is not reached every time due to  $I_\sigma$  estimation errors and quantization.

**4.2. Application to real sounds**

For this experiment<sup>1</sup>, a single-channel 44.1kHz-sampled music signal is processed. This musical piece is composed of a male singer voice and a rhythmic guitar. During the preliminary analysis step, the sinusoidal parameters of the target source signal  $s_k[n]$  are first estimated [4] using a Hann analysis window of length  $N = 1023$  with 50% overlap. The instantaneous mixture signal is computed according to (1) with a version of  $s_k[n]$  synthesized from the reference parameters. Fig. 3(a) and 3(b) show the bit rate necessary to reach a given target SNR respectively for the voice and the guitar source signals. The informed approach combined with the reassignment estimator achieves to reduce the bit rate of about 50% for the guitar signal and about 60% for the voice signal in comparison to uniformly quantized parameters.

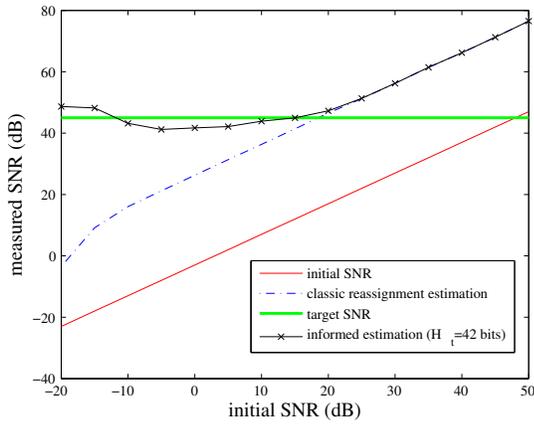
**5. CONCLUSION AND FUTURE WORK**

We have proposed a complete application for informed spectral analysis based on sinusoidal modeling combined with a watermarking technique. In [5] the informed spectral analysis principle was introduced and its theoretical study was completed. This paper generalizes the informed approach to the entire sinusoidal model for application to real sounds signals. This method based on parameters coding is flexible enough to be adapted to others estimators and watermarking techniques to reach any fixed target quality. This method has still to be optimized using entropy coding to reduce the bit rate of the extra information.

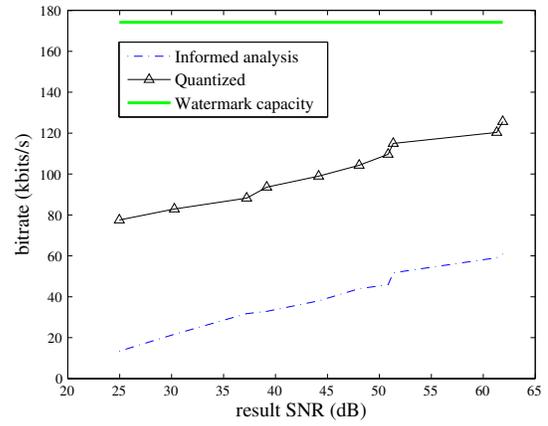
**6. ACKNOWLEDGMENTS**

This research was partly supported by the French ANR DReaM project (ANR-09-CORD-006).

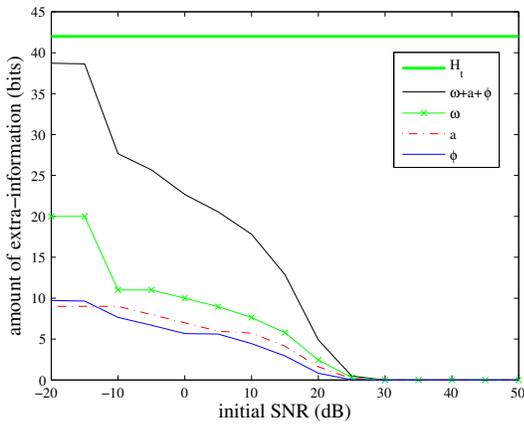
<sup>1</sup>Sounds examples are available on-line at URL: <http://www.labri.fr/~fourer/publi/WASPA11/>



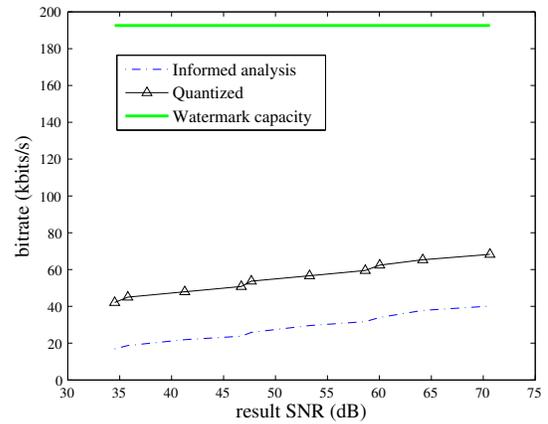
(a)



(a) Voice



(b)



(b) Guitar

Figure 2: Average SNR 2(a) and corresponding average bit rate 2(b) for a target quality set at 45dB for 10000 random generated signals for each SNR.

Figure 3: Bit rate comparison between the full informed approach (quantization without estimator) and informed analysis with  $I_{\sigma,k,l}$  vector quantized with 5 bits.

7. REFERENCES

[1] P. Bofill and M. Zibulevski, "Underdetermined blind source separation," in *Signal Processing*, vol. 81, no. 11, 2001, pp. 2353–2362.

[2] M. Parvaix and L. Girin, "Informed source separation of underdetermined instantaneous stereo mixtures using source index embedding," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 245–248.

[3] E. Schuijers, W. Oomen, B. Brinker, and J. Breebaart, "Advances in parametric coding for high-quality audio," in *Proc. 114th Conv. Audio Eng. Soc.*, Mar. 2003, pp. 201–204.

[4] S. Marchand and P. Depalle, "Generalization of the derivative analysis method to non-stationary sinusoidal modeling," in *Proc. DAFX Conf.*, Sep. 2008, pp. 281–288.

[5] S. Marchand and D. Fourer, "Breaking the bounds: Introducing informed spectral analysis," in *Proc. DAFX Conf.*, Sep. 2010, pp. 359–366.

[6] B. Chen and G. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Trans. on Information Theory*, vol. 47, no. 4, pp. 1423–1443, May 2001.

[7] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acous., Speech, and Sig. Proc.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[8] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. USA: Kluwer Academic Publishers, 1991, ISBN:0-7923-9181-0.

[9] P. Korten, J. Jensen, and R. Heusdens, "High-resolution spherical quantization of sinusoidal parameters," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 15, no. 3, pp. 966–981, Mar. 2007.

[10] J. Pinel, L. Girin, C. Baras, and M. Parvaix, "A high-capacity watermarking technique for audio signals based on MDCT-domain quantization," in *Int. Congress on Acoustics*, Oct. 2010.