A new quantization optimization algorithm for the MPEG advanced audio coder using a statistical sub-band model of the quantization noise

Olivier Derrien, Pierre Duhamel, Maurice Charbit, Gael Richard

A new quantization optimization algorithm for the MPEG Advanced Audio Coder using a statistical sub-band model of the quantization noise

Olivier Derrien*, Pierre Duhamel, *Fellow member, IEEE*,
Maurice Charbit and Gaël Richard, *Member, IEEE*
EDICS: 2-ACOD

Abstract—
In this paper, an improvement of the quantization optimization algorithm for the MPEG Advanced Audio Coder (AAC) is presented. This algorithm, given a bit-rate constraint, minimizes the perceived distortion generated by the signal compression. The distortion can be related to the quantization error level over frequency sub-bands through an auditory model. Thus, optimizing the quantification requires knowledge of the rate-distortion function for each sub-band. When this function can be modeled in a simple way, the algorithm can take a one-loop recursive structure. However, in the MPEG AAC, the rate-distortion function is hard to characterize, since AAC makes use of non-linear quantizers and variable length entropy coders. As a result, the standard algorithm makes use of two nested loops with a local decoder, in order to measure the error level rather than predicting its value.

We first describe a partial sub-band modeling of the rate-distortion function of interest in the MPEG AAC. Then, using a statistical approach, we find a relationship between the error level and the so-called quantization "scale-factor" and propose a new algorithm that is basically similar to a classical one loop "bit allocation" process. Finally, we describe the complete algorithm and show that it is more efficient than the standard one.

Keywords— **Perceptual audio coding, sub-band quantization, scale-factor, bit-rate constraint, distortion constraint, optimization algorithm, statistical model.**

I. Introduction

A *perceptual audio coder* is a frequency domain coder which aims, under a bit-rate constraint, to minimize a measure of distortion significantly related to auditory perception [1]. The quantization error (or quantization noise) introduced by the coding process is properly shaped along frequency sub-bands in such a way that the error is totally or partially masked by the signal itself. Thus, coding the audio signal on each time-window requires: (i) an estimation of the error shaping that is compatible with the required bit rate, (ii) a tuning of the quantization stage in such a way that this error shaping is met as precisely as possible.

O. Derrien is with the Signal, Information and System Laboratory - ISITV/SIS, Université du Sud Toulon-Var, BP 132 - 83957 LA GARDE Cedex, France - Email: olivier.derrien@univ-tln.fr - Phone: 33 (0)4 94 14 25 03 - Fax: 33 (0)4 94 14 24 48; P. Duhamel is with the Signals and Systems Laboratory - CNRS/LSS, Ecole Supélec, Plateau du Moulon, 91192 GIF-SUR-YVETTE Cedex, France - Email: pierre.duhamel@lss.supelec.fr - Phone: 33 (0)1 69 85 17 16 - Fax: 33 (0)1 69 85 17 65; G. Richard and M. Charbit are with the Signal and Image Processing Department - ENST/TSI, 46 rue Barrault 75634 PARIS Cedex 13, France - Email: gael.richard@tsi.enst.fr and maurice.charbit@tsi.enst.fr - Phone: 33 (0)1 45 81 73 65 and 33 (0)1 45 81 71 78 - Fax: 33 (0)1 45 88 79 35. This article was written when O.Derrien was with the Signal and Image Processing Department - ENST/TSI.

A. The error shaping

According to advanced hearing models for audio coding [2], [3], the perceived distortion is directly related to the spectrum of the coding error. More precisely, one usually considers the error level over specific frequency sub-bands, called *perceptual* sub-bands. The definition of these sub-bands is based on psychoacoustic measurements. Furthermore, no audible distortion is detected provided that, in each sub-band, the error level remains below a so-called *masking threshold*, which is strongly signal-dependent. For these reasons, the ratio between the noise level and the masking threshold, or Noise-to-Mask Ratio (NMR), is generally considered a relevant sub-band distortion measure in the context of audio coding. To evaluate the quality of a wide-band signal, a combination of NMR per sub-band can be used [4], although it is not totally significant.

A noise-shaping which would generate an error level lower than the masking threshold would result in a *transparent coding* and would require a minimum number of coding bits. This critical number of bits is generally referred to as the *perceptual entropy* (expressed in bits per sample) [5], noted here as E_p. The corresponding bit rate $r_p = \frac{E_p}{F_s}$, where F_s is the sample rate, can be considered the optimal working point for a perceptual audio coder. Its mean value for a 16 kHz bandwidth monophonic signal seems to be about 96 kbits/s [6] which is often too much for many audio applications. However, the transparency is not always the ultimate goal of audio coding: the ITU-R [7] specifies that, for diffusion, degradations may be "perceptible, but not annoying". Then, a satisfying rate-distortion trade-off can be reached with an optimization algorithm. Now, the MPEG-2/4 Advanced Audio Coder (AAC), considered as the most efficient state-of-the-art audio coder [8], meets the ITU-R quality specifications at 64 kbits/s per channel [9], [10].

B. Tuning the encoder

Audio coders of the previous generation (MPEG-1 Layer I and II [11]) make use of uniform scalar quantizers. In this case, a simple approximation of the sub-band *rate-distortion function*, that relates the signal to noise ratio

(SNR) to the required number of coding bits, is available. In the optimization process, setting a noise level in one sub-band is then equivalent to a *bit allocation*. In what follows, choosing a quantizer in a pre-defined set for a particular sub-band is denoted as a bit-allocation procedure. Coders of the new generation (MPEG-1 Layer III [11], MPEG-2/4 AAC [12], [13]) use non-uniform scalar quantizers associated with a noiseless coding module (Huffman). Thus, characterizing the sub-band rate-distortion function is a much more difficult task. Even though global variations are obvious (a large amount of coding bits generates a low SNR), small variations seem unpredictable. In practice, the problem is bypassed with the use of full iterative algorithms, including a local decoder.

Some studies have shown that the computational complexity of the optimization algorithm is critical for an MPEG encoder: in an MPEG-1 Layer III, which has the same quantization stage and optimization algorithm as the MPEG AAC, the predicted complexity for the quantization optimization is 70 MIPS, while the predicted complexity for the total coding process is 190 MIPS [14]. In other words, the optimization algorithm takes approximately half of the total encoder complexity. Then, in the context of real-time systems, a full-iterative optimization algorithm is a serious drawback. Recent solutions to this problem propose advanced techniques in order to accelerate the optimization process [14], [15], [16], but this generally require complex recursive structures.

In this paper, we propose a novel way to improve the efficiency of the optimization algorithm, both in terms of signal quality and complexity: we characterize the quantization process in a simple and reliable way, using a statistical model. We show that one can take advantage of these results to build a new optimization algorithm based on classical bit-allocation techniques. Compared to the standard algorithm proposed by MPEG [12], a noticeable performance improvement is observed.

II. Formulation of the coding problem

A. Notations

We assume an audio transform coder and note the block of spectral coefficients over the current time-window as $X(k)$, where $k \in \{0 \cdots N-1\}$ is a frequency index and N is the transform length. We also assume that each coefficient-block is split into variable-width frequency sub-bands. We note the limits of sub-band s as $k_{\min}(s)$ and $k_{\max}(s)$. The level of the audio signal (i.e., the estimation of the signal power) over sub-band s is

$$P_X(s) = \sum_{k=k_{\min}(s)}^{k_{\max}(s)} X^2(k). \tag{1}$$

Spectral coefficients $X(k), \quad k \in \{k_{\min}(s) \cdots k_{\max}(s)\}$ are coded with a quantizer Q_s, using $b(s)$ bits. We note the decoded coefficients as $\hat{X}(k)$. The quantization noise is defined by

$$\varepsilon_Q(k) = X(k) - \hat{X}(k) \tag{2}$$

and the noise level by

$$P_Q(s) = \sum_{k=k_{\min}(s)}^{k_{\max}(s)} \varepsilon_Q^2(k). \tag{3}$$

B. Optimal coding with a fixed bit rate

With a fixed output bit rate, the bit-rate constraint is

$$\sum_s b(s) \leq B_{\max}. \tag{4}$$

Recalling the definition of perceptual entropy, if $B_{\max}$ is greater than NE_p, transparent coding can be performed theoretically and the coding error can be maintained below the *masking threshold* in each sub-band:

$$\forall s, \quad P_Q(s) \leq T_M(s) \tag{5}$$

where $T_M(s)$ is the masking threshold, computed by the psychoacoustic model on the current time-window. However, as noted in section I, the typical working point of perceptual coders in practical situations corresponds to bit rates smaller than the perceptual entropy, which means $B_{\max} < NE_p$. Thus, there is a need for an optimization algorithm which would distribute the available binary resources among sub-bands in a way that would disturb the listener the least. Classically, the NMR is used as a sub-band distortion measure:

$$\mathrm{NMR}(s) = \frac{P_Q(s)}{T_M(s)}. \tag{6}$$

Thus, the perceived distortion is directly related to $P_Q(s)$.

In the case of simple quantizers, $P_Q(s)$ and $b(s)$ are related by some simple relationship. For example, with a uniform scalar quantizer working in high resolution, we get

$$P_Q(s) = c\, P_X(s)\, 2^{-2b(s)} \tag{7}$$

where c is a constant (overload factor). In this case, a simple bit-allocation procedure can easily control the noise level by setting the number of coding bits $b(s)$. N.S. Jayant *et al.* [17] have shown that, when quantizers work in high resolution mode (i.e., $P_Q(s) << P_X(s)$), the optimal solution is obtained when the spectrum of the coding error is parallel to the masking threshold. This principle has been implemented in real coding systems, with satisfying results at a medium bit rate [5], [18]. However this approach does not apply to low bit rates. In this case, popular strategies for efficient iterative bit allocation can be summarized as follows:

- Give bits first to the sub-band with the highest NMR [19], [20]. This solution tends to give the same value of NMR to all sub-bands.
- Retrieve bits first to the sub-band with the lowest signal level (called "water-filling technique") [8]. This solution reduces the distortion on high-energy sub-bands.
- Give bits first to the sub-band where the potential gain in NMR is the most important [16]. This solution gives the lowest global NMR over all sub-bands.

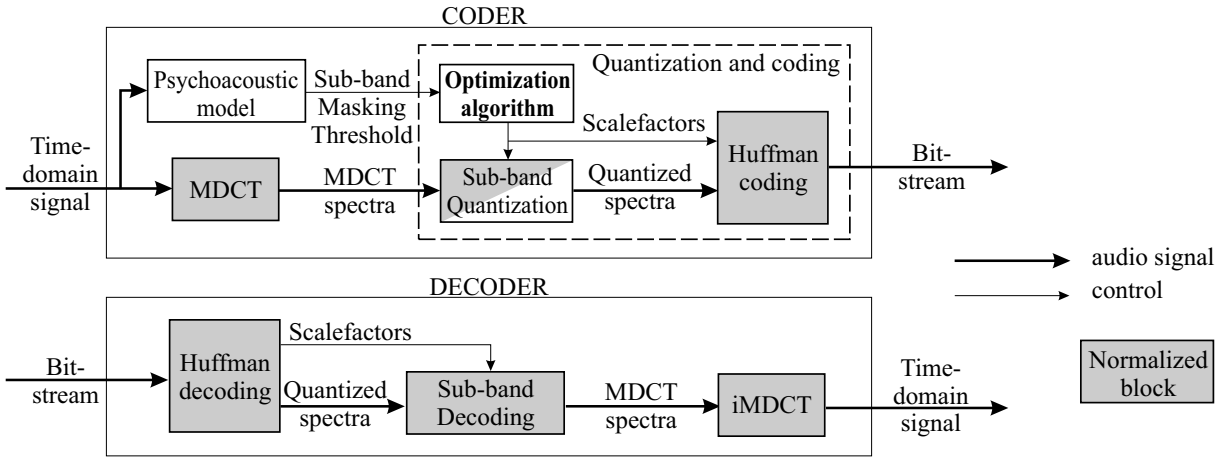


Fig. 1. Synopsis of an MPEG AAC codec.

C. Quantization and coding in the MPEG-AAC

A simplified synopsis of an MPEG AAC codec is presented in figure 1. The audio signal is transformed to a frequency domain by a 50% overlap *modified discrete cosine transform* (MDCT) [21]. The effective signal compression is realized in the *quantization module*. The quantization parameter, called *scale-factor*, can be set independently for each sub-band. The final bit-stream is obtained through a *lossless coding module* (Huffman coding). The decoder has a dual structure. The decoder modules are defined in the MPEG standard to provide full-compatibility between coders and decoders. The coder also requires control modules that are not defined in the MPEG standard in order to allow for future advances in technology that will improve the coding efficiency while remaining compatible. These control modules are the *psychoacoustic model* and the *optimization algorithm*. Based on the psychoacoustic module, the optimization algorithm tunes the quantization parameters by setting scale-factors, the value of which determines the quantization error and the bit rate.

In the coder, the quantization module generates the quantization indexes $i(k)$, corresponding to the spectral coefficients $X(k)$. The MPEG standard defines the decoding function as

$$\forall k \in \{k_{\min}(s) \cdots k_{\max}(s)\}, \quad \hat{X}(k) = A(s)\, i^{\frac{4}{3}}(k). \tag{8}$$

To simplify notations, we note $x^p = \mathrm{Sign}(x)\,|x|^p$ when p is not an integer. $A(s)$ is a scaling parameter, depending on the integer scale-factor $\phi(s)$:

$$A(s) = 2^{\frac{1}{4}\phi(s)}. \tag{9}$$

The decoding function (8) can be split into a sub-band dependent compression function:

$$f_s(x) = A(s)\, x^{\frac{4}{3}} \tag{10}$$

and a very simple sub-band independent decoding function whose reconstruction values are signed integers. The corresponding quantizer R is called the *rounding function*. The quantization process can thus be written as

$$i(k) = R\left(\left(\frac{X(k)}{A(s)}\right)^{\frac{3}{4}}\right) \tag{11}$$

where R is not explicitly defined in the standard. The choice of this function is discussed below in section III-A.

Scale-factors $\phi(s)$ are coded through a single differential Huffman codebook, while quantization indexes $i(k)$ are coded with a set of 12 Huffman codebooks. For a given dynamic range of quantization indexes, either one or two codebooks are possible. The choice is not normalized, and can be made independently for each sub-band. b_ϕ is the number of bits used for coding the set of scale-factors. The total number of bits required for coding the current MDCT spectra is

$$B = b_\phi + \sum_s b(s). \tag{12}$$

We can see that this expression is not separable along sub-bands. However, b_ϕ does not vary much with the scale-factor values $\phi(s)$. From now on, to simplify the coding problem, we consider that b_ϕ is a constant in the optimization.

D. Standard optimization algorithm for the MPEG-AAC

In an MPEG-AAC coder, no direct form is available *a priori* for the relation between the error level $P_Q(s)$ and the number of coding bits $b(s)$ in each sub-band. Only a parametric expression is available:

- The distortion function relates the scale-factor $\phi(s)$ to the error level $P_Q(s)$, through the quantization stage.
- The rate function relates the scaling parameter $\phi(s)$ to the number of coding bits $b(s)$, through the lossless coding module.

Then, all the classical bit-allocation strategies previously described do not strictly apply in this case. The standard algorithm seeks a sub-optimal solution with a two-nested-loop iterative procedure and a local decoder. The inner-loop changes the scale-factor value, independently over frequency sub-bands, in order to meet the masking constraint

(5). The outer-loop performs a global translation of the scale-factor values to meet the total bit-rate constraint. To guarantee the convergence, the scale-factor step is decreased at each iteration.

E. Basics for a new algorithm

We propose a new way to solve the coding problem in the MPEG AAC coder, the motivation for which is as follows: if it were possible to invert the distortion function, this would result in a direct relationship between $P_Q(s)$ and $b(s)$ (as with a uniform scalar quantizer). Thus, an optimization technique, similar to a single-loop iterative bit-allocation process, could be used.

Inverting the distortion function does not seem to be feasible. Therefore, we use the following procedure: given an error threshold $T(s)$, we search for the scale-factor value $\phi(s)$ that minimizes $b(s)$ under the distortion constraint:

$$\forall s, \quad P_Q(s) \leq T(s). \tag{13}$$

This so-called *secondary optimization problem* can be quickly solved thanks to an accurate quantization noise model applied in each sub-band.

Thus, a solution to the main coding problem can be reached with the following algorithm: $T(s)$ is initialized at the masking threshold $T_M(s)$. The secondary problem is solved independently over each sub-band s. If the resulting bit rate B (see equation (12)) is greater than $B_{\max}$, the thresholds $T(s)$ are increased and so on until $B \leq B_{\max}$.

The global optimization is now separated into two distinct steps: (i) a perceptual model provides the set of iso-quality error thresholds among sub-bands, (ii) given these thresholds, a quantization error model provides the scale-factors resulting in the smallest bit rate.

This procedure relies more on the auditory model than the standard one: given an arbitrary noise level (similar to a "masking" constraint), the quantization error model provides the scale-factors which meet the masking constraint with the lowest bit rate. Thus, the task of finding the adequate thresholds so that the perceptual quality is maximized is left to a perceptual model.

III. Sub-band model of the quantization noise

In this section, we look for simple solutions to the secondary coding problem: can we find scale-factor values $\phi(s)$, (or equivalently scaling parameter values $A(s)$), that minimize B under the distortion constraint (13)? Assuming that b_ϕ is a constant in equation (12), this problem can be solved sub-band by sub-band by minimizing $b(s)$. We omit the sub-band index s in the remainder of this section.

A. Setting the rounding function

A rounding function R has to be defined to achieve the exact expression of $P_Q(s)$. The MPEG standard [12] proposes

$$R(x) = \text{Sign}(x)\ \text{Int}\left(|x| + 0.4054\right). \tag{14}$$

Inside each sub-band, the optimal quantizer should minimize the NMR, i.e., minimize $P_Q(s)$. This criterion is equivalent to the Minimum Mean Square Error (MMSE) criterion, and the problem can be solved by a Lloyd-Max procedure [22].

We note $[r_{j-1}, r_j[$ and $[q_{j-1}, q_j[$ respectively as the j-th quantization intervals of quantizers R and Q. The corresponding reconstruction points are respectively j and $\hat{X}_j$. According to equation (8), we have

$$\hat{X}_j = A\, j^{\frac{4}{3}}. \tag{15}$$

The limits of quantization intervals are related through the compression function defined by equation (10):

$$r_j = f_s^{-1}(q_j). \tag{16}$$

If the reconstruction values are set, the MMSE of quantizer Q is obtained when the nearest neighbour condition is verified [23]:

$$q_j = \frac{1}{2}\left(\hat{X}_j + \hat{X}_{j+1}\right) = \frac{A}{2}\left(j^{\frac{4}{3}} + (j+1)^{\frac{4}{3}}\right). \tag{17}$$

The optimum quantization intervals for R are then

$$r_j = \left[\frac{1}{2}\left(j^{\frac{4}{3}} + (j+1)^{\frac{4}{3}}\right)\right]^{\frac{3}{4}}. \tag{18}$$

We can note that

- $r_0 \approx 1 - 0.4054$ and $r_{-1} \approx 0.4054 - 1$. The optimal quantizer and the one proposed in the MPEG document have the same central interval.
- $r_j \to j + 0.5$ when $j \to +\infty$. In high resolution, the optimal quantizer behaves like the "Round" function.

We now assume that the basic quantizer R is the one defined by equation (18).

B. Deterministic approach

A first approach consists of choosing the quantization parameters in such a way that the error level never exceeds the given threshold, for any sub-band and any time-window, on any audio signal. When the error threshold equals the masking threshold, if the masking threshold were an absolute measure, this constraint would be the transparency limit.

The exact expression of the quantization error is obtained by combining equations (8), (11) and (2):

$$\varepsilon_Q(k) = X(k) - A\, R^{\frac{4}{3}}\left(\left(\frac{X(k)}{A}\right)^{\frac{3}{4}}\right) \tag{19}$$

and the distortion is obtained with equation (3). Solving inequality (13) in a formal way given only equation (19) is almost impossible. However, under a high resolution hypothesis, a simplification can be found. We note $\varepsilon_R(k)$ as the error introduced by quantizer R. We have

$$\varepsilon_Q(k) = X(k)\left(1 - \left[1 - \varepsilon_R(k)\left(\frac{A}{X(k)}\right)^{\frac{3}{4}}\right]^{\frac{4}{3}}\right). \tag{20}$$

When R works in high resolution mode, it can be reasonably assumed that

$$|\varepsilon_R(k)| << \left|\frac{X(k)}{A}\right|^{\frac{3}{4}}.$$

With a first-order development around zero, we obtain the asymptotic expression of $\varepsilon_Q(k)$:

$$\varepsilon_Q(k) \sim \frac{4}{3}\,\varepsilon_R(k)\,A^{\frac{3}{4}}\,X^{\frac{1}{4}}(k) \tag{21}$$

and the asymptotic expression of the distortion:

$$P_Q \sim \frac{16}{9}\,A^{\frac{3}{2}} \sum_{k=k_{\min}}^{k_{\max}} \varepsilon_R^2(k)\;|X(k)|^{\frac{1}{2}}. \tag{22}$$

In high resolution, the optimal quantizer is equivalent to the "Round" function, which means $|\varepsilon_R(k)| \leq 1/2$. This leads to an over-estimation of the distortion:

$$P_Q \leq \frac{4}{9}\,A^{\frac{3}{2}} \left[\sum_{k=k_{\min}}^{k_{\max}} |X(k)|^{\frac{1}{2}}\right]. \tag{23}$$

Then, a sufficient condition for the distortion constraint to be true is

$$\frac{4}{9}\,A^{\frac{3}{2}} \left[\sum_{k=k_{\min}}^{k_{\max}} |X(k)|^{\frac{1}{2}}\right] \leq T. \tag{24}$$

Figure 2 represents the exact value of P_Q and the over-estimation function for different values of A, with a real signal over an 8-coefficient sub-band. Figure 3 represents the corresponding values of b. P_Q is a globally increasing function of A, and b a decreasing function. Thus, a sub-optimal solution to the problem is the highest value of A which verifies condition (24), i.e.

$$A = \left(\frac{9}{4}\,\frac{T}{\sum_{k=k_{\min}}^{k_{\max}} |X(k)|^{\frac{1}{2}}}\right)^{\frac{2}{3}}. \tag{25}$$

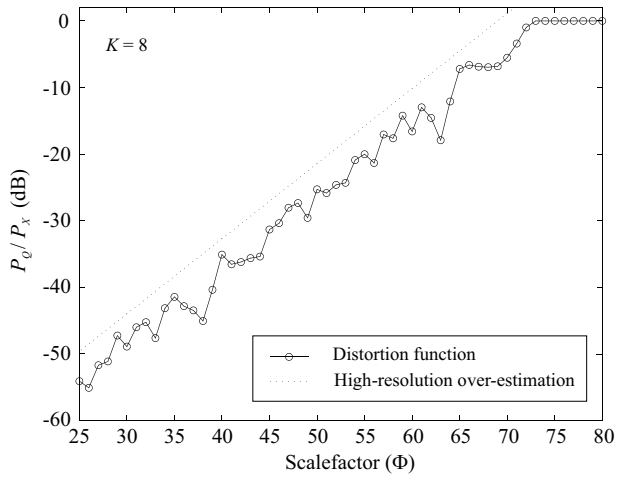


Fig. 2. Example of the distortion function.

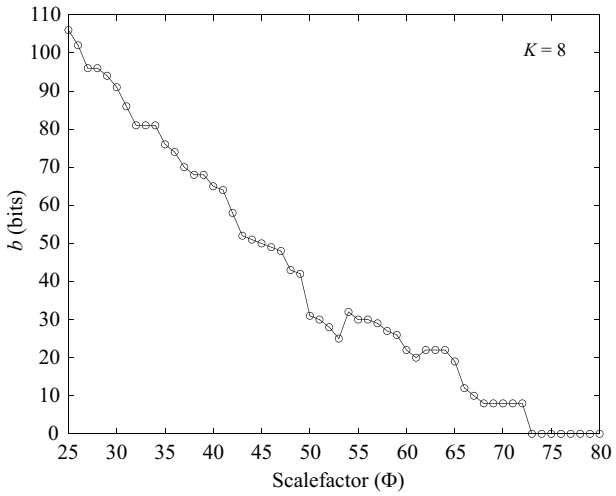


Fig. 3. Example of the rate function, corresponding to the distortion function shown in figure 2.

We compare this solution to the optimum (reached with an exhaustive search) in terms of bit rate. We take 300 long windows of audio signal "d" (see table I), for $K = 8$ and $K = 32$, and for different values of the error threshold T, set by the Signal to Noise Ratio defined by

$$\text{SNR} = \frac{P_X}{T}. \tag{26}$$

This solution meets the distortion constraint but, as we can see in figure 7, it requires a much higher number of coding bits than the optimum.

This is due to the fact that we set the scale-factors in such a way that the upper bound on the quantization noise, for any signal, is smaller than the masking threshold. Even if this upper bound is attainable, such requirements seem unrealistic in practical situations.

C. Statistical approach

In what follows, we propose another solution, which solves a more realistic problem by relaxing the distortion constraint: we now allow the quantization noise level to exceed the threshold for a given percentage of the time.

C.1 Statistical distortion constraint

In this new situation, P_Q is a random variable. A confidence interval criteria has previously been introduced by L. Karray *et al.* for image coding [24]. We adapt this criteria to our problem and replace constraint (13) by

$$\text{Prob}\,\{P_Q \leq T\} \geq \alpha \tag{27}$$

where $\alpha \in [0, 1]$ is a confidence parameter. It means that we allow the distortion to exceed the threshold, but we control the probability of such occurrences.

C.2 The Gaussian model

$X(k), \varepsilon_R(k), \varepsilon_Q(k)$ are now random variables. The probability density function (PDF) of P_Q must be known to solve inequality (27). Its exact expression would be far too

complex, so we chose a simple model. Equation (3) shows that, if $\varepsilon_Q(k)$ are independent and equally distributed and if $K = k_{\max} - k_{\min} + 1$ is large enough, according to the Central-Limit theorem [25], P_Q will follow a Gaussian law:

$$\sqrt{K}\left(\frac{1}{K}P_Q - \mathbb{E}\left[\varepsilon_Q^2\right]\right) \overset{K\to\infty}{\longrightarrow} \mathcal{N}(0,\,\sigma) \tag{28}$$

with

$$\sigma^2 = \mathbb{E}\left[\varepsilon_Q^4\right] - \mathbb{E}\left[\varepsilon_Q^2\right]^2. \tag{29}$$

We note $\sigma^2_{P_Q}$ as the variance of P_Q. The distortion constraint (27) is equivalent to

$$\mathbb{E}[P_Q] + \beta\,\sigma_{P_Q} \leq T \tag{30}$$

with

$$\beta = \sqrt{2}\,\mathrm{Erf}^{-1}(2\alpha - 1). \tag{31}$$

Erf^{-1} is the inverse standard error function (see [26] section 26.2. for details). Equation (28) leads to

$$\begin{cases} \mathbb{E}[P_Q] & = & K\,\mathbb{E}\left[\varepsilon_Q^2\right] \\ \sigma^2_{P_Q} & \overset{K\to\infty}{\sim} & K\sigma^2. \end{cases} \tag{32}$$

We have also considered a non-asymptotic model using a Gamma-law. This finer model is equivalent to the Gaussian one on large sub-bands. We expected similar performances on large sub-bands and an improvement on narrow sub-bands. However, we observed no significant improvement, and we finally chose to present only the Gaussian model.

C.3 High-resolution solution

Under the Gaussian assumption, we only need to estimate the first and second moments of quantization error ε_Q. Under a high resolution hypothesis, we assume that ε_R and X are independent variables [27]. The asymptotic expression (21) leads to

$$\mathbb{E}\left[\varepsilon_Q^p\right] \sim \left(\frac{4}{3}\right)^p A^{\frac{3p}{4}}\,\mathbb{E}[\varepsilon_R^p]\,\mathbb{E}\left[|X|^{\frac{p}{4}}\right]. \tag{33}$$

When the quantizer R works in high resolution mode, ε_R can be modeled by a uniform random variable on $\left[-\frac{1}{2},\,\frac{1}{2}\right[$ [27]. Then, we have

$$\mathbb{E}[\varepsilon_R^p] = \frac{1}{(p+1)\,2^p} \tag{34}$$

and

$$\mathbb{E}\left[\varepsilon_Q^p\right] = a_p\,A^{\frac{3p}{4}}\,\mu_{\frac{p}{4}} \tag{35}$$

with

$$a_p = \frac{2^p}{(p+1)\,3^p} \tag{36}$$

$$\mu_p = \mathbb{E}[|X|^p]. \tag{37}$$

Equation (32) can now be written as

$$\begin{cases} \mathbb{E}[P_Q] & \sim & K a_2 \mu_{\frac{1}{2}}\,A^{\frac{3}{2}} \\ \sigma^2_{P_Q} & \sim & K(a_4\mu_1 - a_2^2\mu^2_{\frac{1}{2}})\,A^3. \end{cases} \tag{38}$$

Figure 4 shows histograms of P_Q (for $K = 8$ and $K = 32$) over 300 long windows of signal "d". P_X was normalized to 90 dB, and the scale-factor value is 52, which corresponds to a 18 dB SNR. Our model seems to fit the data accurately, even on the narrow sub-band ($K = 8$).

Finally, we obtain an explicit expression of the distortion constraint (30) for large values of K in high resolution mode:

$$\left(K a_1 \mu_{\frac{1}{2}} + \beta\sqrt{K\left(a_2\mu_1 - a_1^2\mu^2_{\frac{1}{2}}\right)}\right)\,A^{\frac{3}{2}} \leq T. \tag{39}$$

As the rate function is globally decreasing (see section III-B), a sub-optimal solution of the secondary coding problem is

$$A = \left(\frac{T}{K a_2 \mu_{\frac{1}{2}} + \beta\sqrt{2K(a_4\mu_1 - a_2^2\mu^2_{\frac{1}{2}})}}\right)^{\frac{2}{3}}. \tag{40}$$

To evaluate this solution, we quantize each audio signals from the material provided in table I, with a scaling parameter determined according to equation (40) (for details of implementation, see section IV-B) and estimate $\mathrm{Prob}\{P_Q \leq T\}$ for different values of K. This is called the *threshold verification level* (TVL). Figure 5 represents the TVL for $K = 8$ and $K = 32$. The error threshold for each sub-band is still set by the specification of the SNR. First, we can observe that the distortion constraint (27) is always met, which confirms that our solution is valid. Second, the TVL increases as the SNR decreases, which means that this solution resembles the deterministic over-estimation in low resolution conditions. Unfortunately, this procedure would also result in a bit-rate waste for low SNR values (the percentage of time when the error level exceeds the given threshold is overestimated). This is attributed to the use of a high resolution model.

C.4 Improved solution

The previous solution was based on high resolution approximations to obtain analytic expressions of $\mathbb{E}\left[\varepsilon_Q^p\right]$. Now, we reject this hypothesis, but still keep the Gaussian model. The exact expression of ε_Q is given by equation (19). *A priori*, $\mathbb{E}\left[\varepsilon_Q^p\right]$ depends on A and on the PDF of MDCT coefficients X. We assume that X follows a centered law (not necessarily Gaussian) of variance σ_X^2. The corresponding normalized variable $\tilde{X}$ (of variance 1) verifies $X = \sigma_X\,\tilde{X}$. Thus, if we note

$$A = \sigma_X\,\tilde{A} \tag{41}$$

we have

$$\varepsilon_Q = \sigma_X\,\tilde{\varepsilon}_Q \tag{42}$$

with

$$\tilde{\varepsilon}_Q = \tilde{X} - \tilde{A}\,R^{\frac{4}{3}}\left(\left(\frac{\tilde{X}}{\tilde{A}}\right)^{\frac{3}{4}}\right). \tag{43}$$

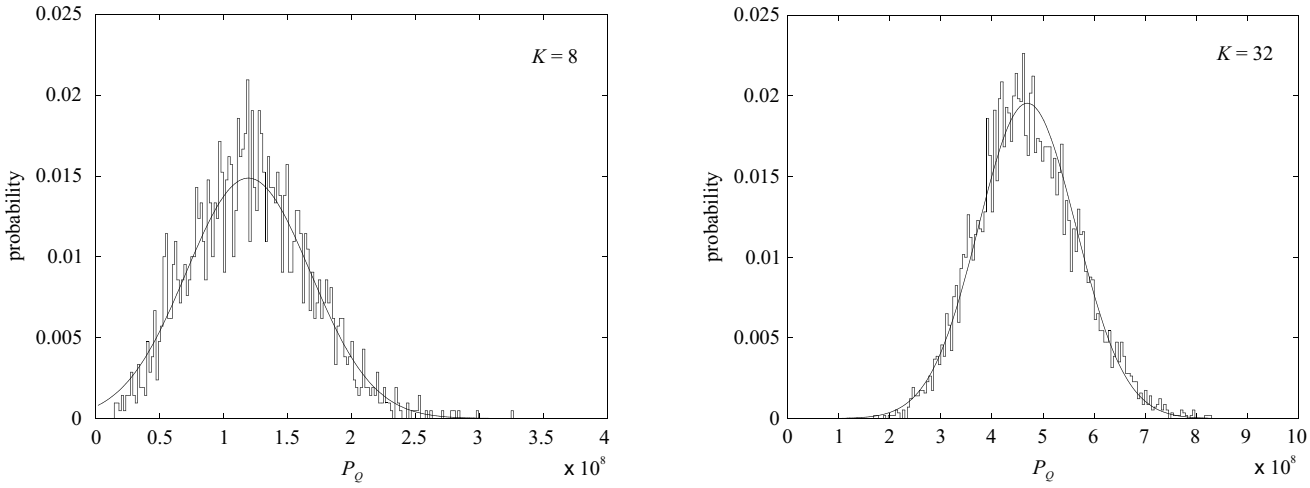


Fig. 4. Gaussian model and histograms of the error level over 300 long windows of signal "d". P_X is normalized to 90 dB. SNR = 18 dB.

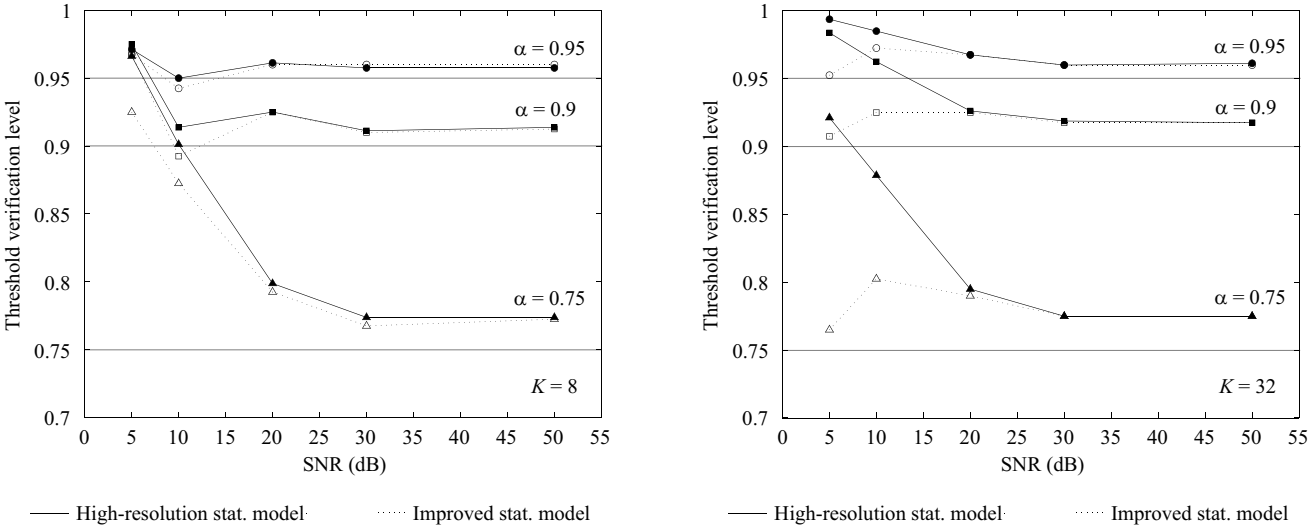


Fig. 5. Threshold verification level for different values of SNR, measured on 2400 long windows (300 windows for each signal).

This expression is similar to equation (19). It means that only the quantization error $\tilde{\varepsilon}_Q$, obtained with a normalized signal $\tilde{X}$, has to be studied. According to equations (29), (32) and (42), the moments of P_Q are

$$\begin{cases} \mathbb{E}[P_Q] & = & K\,\sigma_X^2\,\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right] \\ \sigma_{P_Q}^2 & \overset{K\to\infty}{\sim} & K\,\sigma_X^4\,\left(\mathbb{E}\left[\tilde{\varepsilon}_Q^4\right]-\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right]^2\right). \end{cases} \tag{44}$$

Then, the distortion constraint (30) is equivalent to

$$\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right]+\beta\sqrt{\frac{1}{K}\left(\mathbb{E}\left[\tilde{\varepsilon}_Q^4\right]-\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right]^2\right)}\leq\frac{T}{K\,\sigma_X^2}. \tag{45}$$

A sub-optimal solution is the highest value of $\tilde{A}$ which verifies inequality (45). Finding this solution requires that $\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right]$ and $\mathbb{E}\left[\tilde{\varepsilon}_Q^4\right]$ can be evaluated as functions of $\tilde{A}$. Since a general analytic expression is difficult to find, we measure these moments on a corpus made with real audio signals. We split the audio material described in table I in two parts: the first one, composed of audio signals from "a" to "d", is used for measures (see figure 6). An iterative process, describing the measurement curves, is used to seek the sub-optimal solution (see section IV-B).

The second part, composed of audio signals from "e" to "h", is used for verifications. The protocol is similar to the one used for the high-resolution solution. We can see in figure 5 that the TVL is significantly more precise at low SNR on the large sub-band. This result is consistent with our hypothesis: we rejected the high-resolution approximation, but we kept the Gaussian asymptotic model ($K \to \infty$).

D. Bit-rate evaluation

In previous sections, we have proposed three simple sub-optimal solutions to the second coding problem. Now, we evaluate how far these solutions are from the optimal one, in terms of bit rate.

We still consider a single sub-band s, and measure the required number of coding bits $b(s)$. The optimal solution is obtained with an exhaustive search. This technique gives

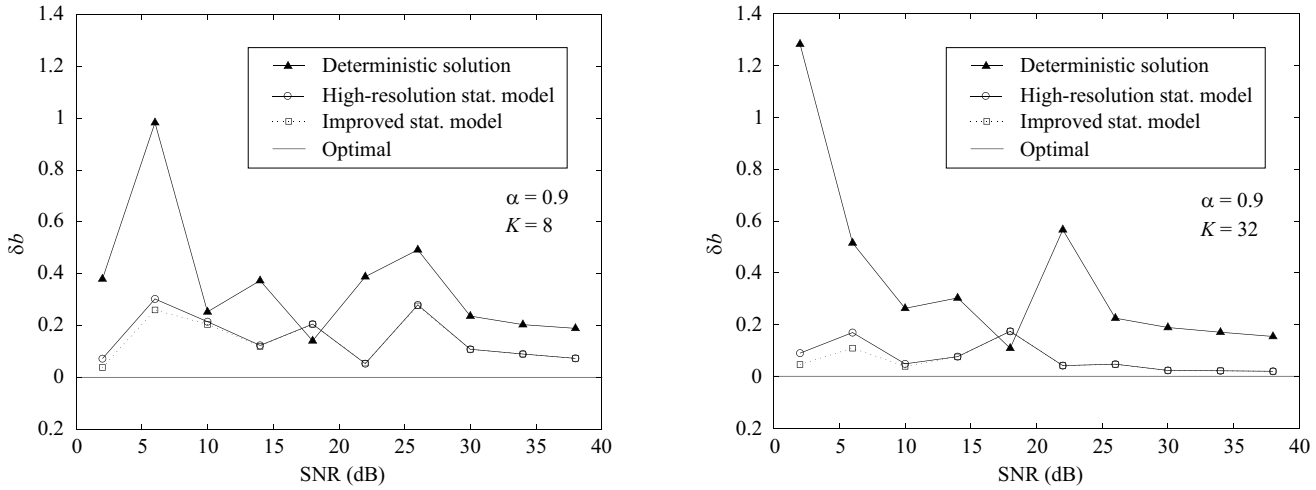


Fig. 7. Bit rate as a function of the SNR, measured on 2400 long windows (300 windows for each signal). $\delta b = \frac{b - b_{\text{opt}}}{b_{\text{opt}}}$. b_{opt} is the absolute optimal bit rate, i.e., the minimum bit rate under a strict distortion constraint.

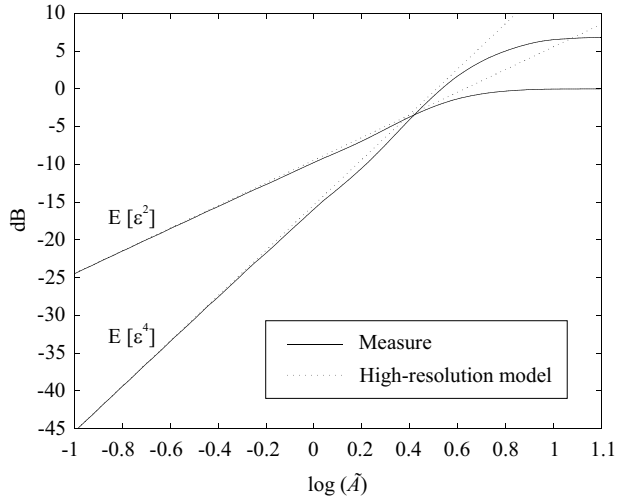


Fig. 6. $\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right]$ and $\mathbb{E}\left[\tilde{\varepsilon}_Q^4\right]$ as a function of $\tilde{A}$, measured over 1200 long windows (300 windows for each signal from "a" to "d").

a bit-rate reference, but cannot be used for coding as it is extremely slow. Figure 7 shows the results for $K = 8$ and $K = 32$, and $\alpha = 0.9$.

We can conclude that:

1. The deterministic solution generates a greater bit rate than the others, especially at low SNR.
2. The high-resolution statistical model reduces the bit rate significantly. This effect is greater on large sub-bands.
3. The improved statistical model reduces the bit rate at low SNR.

It appears that both statistical models are better than the deterministic solution.

Setting the parameter α is a trade-off between the TVL and the bit rate. It also denotes the confidence we have in the auditory model. The better trade-off for solving the second coding problem, as defined in section II-E, seems to be reached with a high confidence parameter, typically $\alpha = 0.9$. This value will be used in section IV. If the algorithmic complexity is critical, the high-resolution approximation is better. If not, the improved solution can be used.

IV. Description of the allocation algorithm

We have described a sub-band model for the quantization noise. Given an error threshold $T(s)$, we can find the sub-band parameter $A(s)$, hence the scale-factor value $\phi(s)$, that minimizes the number of coding bits $b(s)$ under a distortion constraint (13).

This section now considers the main coding problem, and aims at minimizing the perceived distortion under a bit-rate constraint. Our model performs a spectral noise shaping by setting the error threshold $T(s)$ for each sub-band. Then, the optimum solution to the main coding problem can be reached with a single-loop iterative process as described in section II-E. The progressive degradation of the perceived distortion level (i.e., the calculation of the error thresholds $T(s)$) will be discussed in the next section.

A. Progressive degradation of the perceived distortion

For a given wide-band MDCT spectrum $X(k)$ and a set of masking thresholds $T_M(s)$, computed by the psychoacoustic model, we look for a set of iso-quality distortion thresholds $T_i(s)$ for which the perceived distortion increases with index i. This problem is similar to the one treated in many bit-allocation algorithms and the same techniques should apply here. The solution we propose is based on a combination of several popular techniques: constant NMR for high SNR (1st phase), water-filling for medium to low SNR (2nd phase), with a protection factor to avoid large distortion levels at low frequencies. And finally a constant SNR degradation for very low SNR (3rd phase).

On masked sub-bands, the signal is irrelevant because it is imperceptible to the listener and therefore does not have

Number	Author	Identification	Style	Duration
a	J.J. Cale	"Cocaine"	Rock (instrumental)	8.1 s
b	A. Soler	"Fandango"	Classical (harpsichord)	7.8 s
c	J. Copeland	"Hold On"	Blues (instrumental)	8.6 s
d	J. F. Haendel	"Messiah"	Classical (choir)	7.8 s
e	T. Chapman	"Talkin' About Revolution"	Folk (singing voice)	8.7 s
f	St Germain	"Rose Rouge"	Electronic (instrumental)	8.4 s
g	S. Rollins	"In a Sentimental Mood"	Jazz (traditional)	8.9 s
h	L. v. Beethoven	6th Symphony	Classical (orchestra)	9.0 s

TABLE I
AUDIO MATERIAL FOR THE VALIDATION OF THE MODEL.

to be coded ($\hat{X}(k) = 0$). We set

$$\forall i,\ T_i(s) = P_X(s) \tag{46}$$

where $P_X(s)$ is defined in equation (1). Over unmasked sub-bands (i.e., when $T_M(s) \leq P_X(s)$), $T_i(s)$ should satisfy

$$T_M(s) \leq T_i(s) \leq P_X(s). \tag{47}$$

From now on, we assume that all variables are in dB. For unmasked sub-bands, we first determine a protection threshold: $G(s) = P_X(s) - \tau(s)$. $\tau(s)$ depends on the window size and on the sampling frequency (see table II). The initialization is made with: $T_0(s) = T_M(s)$.

Each threshold $T_i(s)$ is obtained from $T_{i-1}(s)$, with three different rules, depending on i:

- **1st phase**, until $T_i(s) - T_M(s) \leq 6$dB

$$\tilde{T}_i(s) = T_{i-1}(s) + r_1$$

$$T_i(s) = \min\left(\tilde{T}_i(s), G(s)\right)$$

- **2nd phase**, until $T_i(s) < G(s)$ for at least one sub-band

$$\tilde{T}_i(s) = \max\left(\tilde{T}_{i-1}(s), m_{i-1} + r_1\right)$$

$$T_i(s) = \min\left(\tilde{T}_i(s), G(s)\right)$$

- **3rd phase**

$$T_i(s) = T_{i-1}(s) + r_2$$

with $m_i = \min_s\left(\tilde{T}_i(s)\right)$. r_1 and r_2 are step constants, set respectively to 1 dB and 0.25 dB.

B. Implementation of the sub-band model

The moments of MDCT coefficients μ_p are measured with the following classical estimator

$$\hat{\mu}_p = \frac{1}{K} \sum_{k=k_{\min}}^{k_{\max}} |X(k)|^p \tag{48}$$

and the nearest integer scale-factor value is obtained from the scaling parameter value with

$$\phi = \text{Round}\left(4 \log_2(A)\right). \tag{49}$$

Long window		Short window	
s	$\tau(s)$ (dB)	s	$\tau(s)(dB)$
1 - 3	10	1	8
4 - 5	9	2	5
6	8	3 - 11	2
7 - 8	7	12 - 14	0
9 - 10	6		
11 - 12	5		
13	4		
14	3		
15 - 40	2		
41 - 49	0		

TABLE II
PROTECTION FACTOR FOR A 48 KHZ SAMPLE RATE.

To implement the improved statistical solution, $\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right]$ and $\mathbb{E}\left[\tilde{\varepsilon}_Q^4\right]$ have to be measured as a function of $\tilde{A}$ on test signals of unity variance, and stored. As we can see in figure 6, these functions are regular so only a small number of points have to be measured (we took 40 points). The intermediate values are obtained with a log-linear interpolation. We can also notice that the exact values of $\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right]$ and $\mathbb{E}\left[\tilde{\varepsilon}_Q^4\right]$ are always lower than their high-resolution approximations. This means that the sub-optimal value of the scaling parameter A is greater than the one obtained with the high-resolution model, and quite close to it.

As a result, our iterative algorithm is summarized as follows:

1. Initialize A at the high-resolution value defined by equation (40) and obtain the normalized scaling parameter:

$$\tilde{A} = \frac{A}{\sqrt{\tilde{\mu}_2}}$$

2. Interpolate $\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right](\tilde{A})$ and $\mathbb{E}\left[\tilde{\varepsilon}_Q^4\right](\tilde{A})$.
3. Estimate the left part of the distortion constraint (45). Increase $\tilde{A}$ and iterate steps 2 and 3 until

$$\mathbb{E}\left[\tilde{\varepsilon}_Q^2\right] + \beta\sqrt{\frac{1}{K}\left(\mathbb{E}\left[\tilde{\varepsilon}_Q^4\right] - \mathbb{E}\left[\tilde{\varepsilon}_Q^2\right]^2\right)} > \frac{T}{K\,\mu_2}$$

If the step size for $\tilde{A}$ is small enough (we chose 0.5 dB), the previous value is close to the sub-optimal solution. We finally get the nearest scale-factor value with

$$\phi = \text{Round}\left(2\ \log_2\left(\mu_2\right) + 4\log_2\left(\tilde{A}\right)\right). \qquad (50)$$

As the initialization value is close to the optimal, this search requires very few iterations.

V. Performance evaluation

The four main performance criteria for an audio coder, according to N. Jayant [1], are: signal quality, efficiency (bit rate), complexity (computation time) and delay. The delay is fixed by the MPEG standard and the bit rate depends on the application. For a monophonic signal, we consider a 64 kbits/s bit rate, which should generate near perfect quality or perceptible, but not annoying, degradations for some signals, and a 48 kbits/s bit rate for subjective evaluations, which should generate slightly annoying degradations for some signals.

To simplify the evaluation procedures, we evaluated signal quality and complexity for the standard algorithm and only one of our two model-based algorithms. We chose to implement the one based on the high-resolution statistical model, as it seems to provide a good trade-off between complexity and signal quality.

Both optimization algorithms (standard and model-based) are used in the same AAC main profile codec. The sample rate is 48 kHz. The psychoacoustic model is the one proposed in the MPEG standard. The MDCT window is derived from the Kaiser-Bessel function. The switch between long and short windows is enabled.

A. Signal Quality

The signal quality can be assessed using objective quality measures (see [28] for a selection of 6 different methods). However, as mentioned in [29], the ultimate test of any audio product is the human listener. A number of subjective test methods have been proposed, amongst which a few have led to ITU recommendations [30], [31], [32]. In this work, we refer largely to the ITU recommendation BS.1116 [31], which is especially designed for subjective assessment of small impairments in audio systems. The subjective evaluation was carried out at a bit rate of 48 kbits/s since near transparent quality is obtained for both codecs at 64 kbits/s or higher bit rates.

A.1 Test procedure

The test followed the common "triple stimulus / hidden reference / double blind" approach. This method consists of presenting three versions of an audio signal: *"Reference"*, *"A"* and *"B"*, where *"Reference"* is the reference signal (unprocessed), and where one of the other two versions is a hidden reference (unprocessed) - for example *"A"* - and the other is the coded version - for example *"B"*. For each trial, the hidden reference (*"A"* or *"B"*) is randomly chosen. The subject is free to listen to each signal as many times as necessary. Then, the quality of the signals *"A"* and *"B"* are assessed using a nearly continuous grading scale (steps of 0.1) between 1.0 (very annoying impairment) and 5.0 (imperceptible impairment), see table III. Since the listener knows in advance that either *"A"* or *"B"* is a hidden reference, at least one grade of 5.0 must be given. Each subject carried out the test individually over a single session. The average duration of a session was 25 minutes. As advised in [29], listeners are strongly encouraged to guess which signal is the hidden reference even if the impairment is imperceptible (a typical grade of 4.8 or 4.9 is then given in this case). The tests were conducted in a quiet environment using high quality headphones (*Sennheiser eh2270*).

Impairment	Grade
Imperceptible	5.0
Perceptible but not annoying	4.0
Slightly annoying	3.0
Annoying	2.0
Very annoying	1.0

TABLE III
The ITU-R five point impairment scale

A.2 Test material

It is widely acknowledged that critical audio test items should be chosen in order to reveal differences among systems. Critical audio material refers to audio excerpts that stress the systems under test. In our case, the selection was done by choosing a subset of excerpts where audio impairments of both coding schemes were the most audible and by favoring the widest variety of musical content and style. All excerpts are monophonic and were played at a sample rate of 48 kHz. Table IV gives the list of the selected test material.

A.3 Listeners

A total of 16 subjects participated to the listening test. All subjects were familiar with audio systems and two of them were familiar with audio coding evaluation. It is important to note that the authors directly involved in the coder optimization were not included in the test. All subjects underwent a training phase which allowed them to become more experienced listeners in identifying coding artefacts. This training phase was always guided by a test supervisor. A post-screening of all listeners was carried out to only keep "reliable" listeners. More precisely, this post-screening meant excluding all listeners who failed to recognize the hidden reference in a significant way, i.e., those listeners who gave a grade below 4.5 to at least one hidden reference. After the post-screening stage, 10 listeners were judged reliable.

A.4 Results

The results of the subjective test for the 10 reliable subjects are given in figure 8. Similarly to [33], the results are given as "diffgrades", which means the grades awarded to

Number	Author	Identification	Style	Duration
1	J.J. Cale	"Cocaine"	Rock (instrumental)	8.1 s
2	YOU band	"Crescent Serenade"	Modern jazz (octet)	12.1 s
3	Suzanne Vega	"Tom's Dinner"	Singing voice	9.5 s
4	J.F. Haendel	"Messiah"	Classical (choir)	7.8 s
5	J. P. Sousa	"The Stars And Stripes Forever"	Brass band	12.1 s
6	Paul Simon	"Late In The Evening"	Pop (instrumental)	7.9 s
7	M. Ravel	"Alborada Del Gracioso"	Classical (orchestra)	8.4 s
8	HH Band	"Kitchen"	Jazz (traditional)	8.5 s
9	J. Haydn	String Quartet "Kaiser"	Classical (quartet)	15.7 s
10	22 Project Band	"Tea Break"	Blues (instrumental)	9.1 s

TABLE IV
AUDIO MATERIAL FOR SUBJECTIVE EVALUATION. ITEMS 2,5 AND 7 TO 10 ARE EXTRACTED FROM RWC DATABASE [35]

the coded version minus the grades awarded to the hidden reference. For example, an impairment grade of 3.0 awarded to the coded version results in a diffgrade of -2.0. Figure 8 displays the results as mean scores with 95% confidence interval which are determined as follows [34]: first, for each codec i, the mean score for each item j, is given by

$$m_{ij} = \frac{1}{N}\sum_{k=1}^{N} s_{ijk} \tag{51}$$

where N is the number of subjects and s_{ijk} is the diffgrade scores given by subject k. The overall mean scores are then the mean of the m_{ij} values. The 95% confidence intervals are computed as

$$\left[m_{ij} - 1.96\frac{\sigma_{ij}}{\sqrt{N}}\,,\; m_{ij} + 1.96\frac{\sigma_{ij}}{\sqrt{N}}\right] \tag{52}$$

where σ_{ij} is the standard deviation of the scores s_{ijk} over all subjects.

From these results, it can be clearly seen that our proposed algorithm provides a significantly better quality for all but two test items, for which the diffgrade scores of both codecs are within the 95% confidence intervals. On average (right of figure 8), the proposed codec significantly surpasses the standard codec.

B. Complexity

To evaluate the complexity, we measured the mean computation time necessary for coding one time-window, for the material provided in table IV and for bit rates of 64 and 48 kbits/s.

We characterized both the computation time of the optimization algorithm and the total computation time. We precise that the implementation was made on a MATLAB 6 platform, and that we did not use a fast scheme (FFT based) for the implementation of the filter-bank (MDCT). Thus, the results might slightly differ with a compiled coder (for example from a source code in C), and the total computation time would be lower with a fast MDCT scheme. The results are presented in figure 9: bar lengths give the execution time of the entire coding process. The white part represents the execution time of the standard algorithm and the grey part the execution time of the high-resolution model-based algorithm. The black part represents the remaining computation time (window-switching, MDCT and psycho-acoustic model), which is common to both implementations.

From these results, we can conclude that:
1. With the standard algorithm, the optimization and quantization module takes 48% of the computation time at 48 kbits/s and 44% at 64 kbits/s, which fits the predicted results (see section I).
2. For the optimization and quantization module alone, the computation time is reduced by 38% at 48 kbits/s and 56% at 64 kbits/s with our algorithm.
3. For the whole coding process, the computation time is reduced by 20% at 48 kbits/s and 31% at 64 kbits/s.

VI. CONCLUSION

This article proposes a slight change in perspective towards high-quality audio coding: classically, the masking threshold is assumed to define transparency, and lower quality encoded signals are obtained by reference to this transparency threshold. As a result, the control of the actual quantization error level in all sub-bands is usually quite loose for these lower quality signals. In our approach, we first begin by defining as precisely as possible an error threshold providing the required quality. Then, the quantization stage is tuned in such a way that the corresponding distortion constraint is met with a specific criterion: the error threshold should not be exceeded by more than a percentage of time α. This percentage is introduced because, as expected, the threshold is not an absolute value, but is rather loosely defined. Parameter α thus represents the confidence we have in the perceptual model. Clearly, the perceptual model used in this paper for obtaining iso-quality masking profiles is very simple, and can be improved.

It appears that our new algorithm is more efficient than the standard one proposed in the MPEG standard: according to a normalized subjective listening test, our coding algorithm increases the signal quality, while the computation

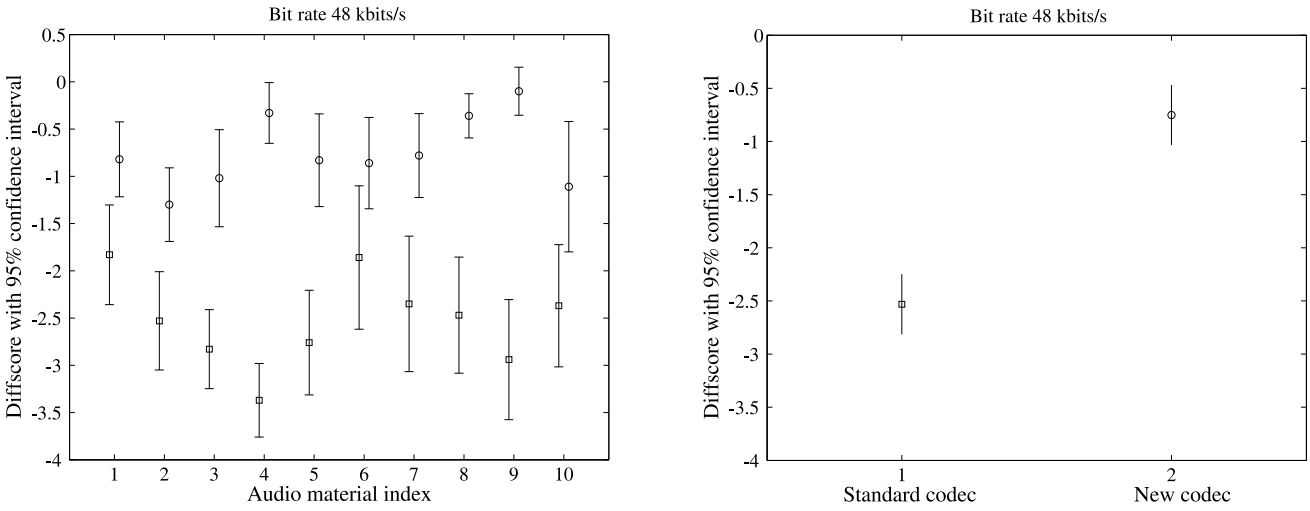


Fig. 8. (Left) Mean diffgrade results over all reliable listeners for each item; (Right) Mean diffgrade results over all listeners and all items for each codec (circles correspond to the new algorithm and squares to the standard codec).

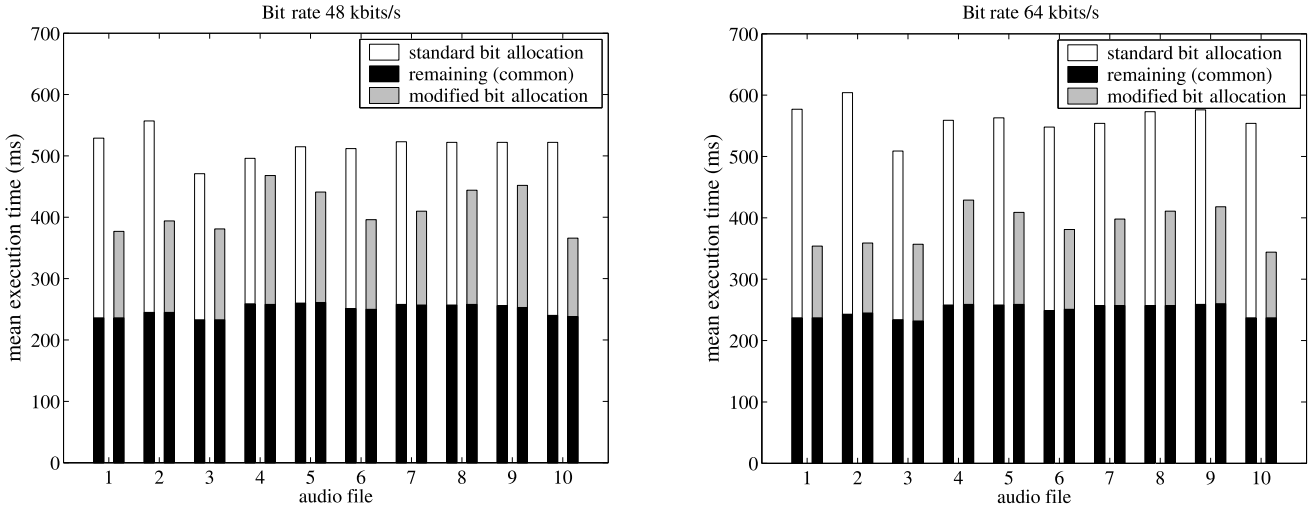


Fig. 9. Mean execution time necessary for coding one time-window, for the audio material provided in table IV and for bit rates of 64 and 48 kbits/s. The modified bit-allocation procedure uses the high-resolution model-based algorithm.

time is significantly reduced.

In the long term, the main advantage of our optimization scheme is its flexibility towards psychoacoustics: our models for the quantization noise can be used with many models of perceived quality degradation. Then, improved perceptual models should directly result in improved coding efficiency.

References

[1] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385–1422, October 1993.

[2] C. Colomes, M.Lever, Y.F. Dehery J.B. Rault, and G. Faucon, "A perceptual model applied to audio bit-rate reduction," *Journal of the Audio Engeneering Society*, vol. 43, no. 4, pp. 233–239, 1995.

[3] F. Baumgarte, "Improved audio coding using a psychoacoustic model based on a cochlear filter bank," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 495–503, October 2002.

[4] K. Brandenburg and T. Sporer, "NMR and Masking Flag : Evaluation of quality using perceptual criteria," *Proceedings of the 11th International Conference of the Audio Engineering Society*, pp. 169–179, 1992.

[5] J.D. Johnston, "Transform coding of audio signals using percetual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, February 1988.

[6] R. Veldhuis, "Bit rates in audio source coding," *IEEE Journal on Selected Areas in Commmunications*, vol. 10, no. 1, pp. 86–96, 1992.

[7] ITU-R, *ITU-R Recommendation BS.1115. Low bit-rate audio coding*, 1994.

[8] M. Bosi and E. Goldberg, *Introduction to Digital Audio Coding and Standard*, Kluwer Academic Publishers, 2002.

[9] D. Kirby and K. Watanabe, "Report on the formal subjective listening tests of MPEG-2 NBC multichannel audio coding)," Tech. Rep. JTC1/SC29/WG11 N1419, ISO/CEI, November 1996.

[10] K. Watanabe D. Meares and E. Scheirer, "Report on the MPEG-2 AAC stereo verification tests," Tech. Rep. JTC1/SC29/WG11 N2006, ISO/CEI, February 1998.

[11] International Organization for Standardization, *ISO/IEC 11172-3 (Information Technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbits/s) - Part 3: Audio*, 1993.

[12] International Organization for Standardization, *ISO/IEC 13818-7 (MPEG-2 Advanced Audio Coding, AAC)*, 1997.

[13] International Organization for Standardization, *ISO/IEC 14496-3 (Information technology - Very low bitrate audio-visual coding - Part3: Audio)*, 1998.

[14] H. O. Oh, C. J. Song, Y. C. Park, and D. H. Youn, "Low power MPEG audio encoders using simplified psychoacoustic model and fast bit allocation," *IEEE Transactions on Consumer Electronics*, vol. 47, no. 3, pp. 613–621, August 2001.

[15] D. Domazet and M. Kovac, "Advanced software implementation of MPEG-4 AAC audio encoder," *Proceedings of the 4th EURASIP Conference on Video/Image Processing and Multimedia Communications*, pp. 679–684, July 2003.

[16] C. H. Yang and H. M. Hang, "Efficient bit assignment strategy for perceptual audio coding," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, vol. V, pp. 405–408, 2003.

[17] N. S. Jayant and P. Noll, *Digital Coding of Waveforms : Principles and applications to Speech and Video*, Prentice Hall, 1984.

[18] Y. Mahieux and J.P. Petit, "High-quality audio transform coding at 64 kbps," *IEEE Transactions on Communications*, vol. 42, no. 11, pp. 3010–3019, 1994.

[19] D.H. Teh, S.N. Koh, and A.P. Tan, "Efficient bit allocation algorithm for ISO MPEG audio encoder," *Electronic Letters*, vol. 34, no. 8, pp. 721–722, April 1998.

[20] K.T. Fung, Y.L. Chan, and W.C. Siu, "A fast bit allocation algorithm for MPEG audio encoder," *Proceedings of International Symposium on Intelligent Multimedia, Video and Speech Processing*, pp. 5–8, May 2001.

[21] J.P. Princen and A.B. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-34, pp. 1153–1161, 1986.

[22] J. Max, "Quantization for minimum distortion," *IRE Transactions on Information Theory*, vol. IT-6, pp. 7–12, March 1960.

[23] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kulwer Academic Publisher, Norwell, Massachusetts, 1992.

[24] L. Karray, P. Duhamel, and O. Rioul, "Image coding with an L^∞-norm and confidence interval criteria," *IEEE Transactions on Image Processing*, vol. 7, no. 5, pp. 621–631, 1998.

[25] M. Fisz, *Probability Theory and Mathematical Statistics*, Wiley, New York, 1963.

[26] M. Abramowitz and A. Stegun, *Handbook of Mathematical Functions*, Dover Publications Inc., New York, 1970.

[27] S.P. Lipshitz, R.A. Wannamaker, and J. Vanderkooy, "Quantization and dither : a theoretical survey," *Journal of the Audio Engeneering Society*, vol. 40, no. 5, pp. 355–374, May 1992.

[28] ITU-R, *ITU-R Recommendation BS.1387. Method for objective measurements of perceived audio quality*, 2001.

[29] T.Grusec, L.Thibault, and G.Soulodre, "Subjective evaluation of high quality audio coding systems: Methods and results in the two-channel case," *Proceedings of the 99th International Conference of the Audio Engineering Society*, New York City, October 1995, preprint #4065.

[30] ITU-R, *IUT-R Recommendation BS.5562. Subjective assessment of sound quality*, 1990.

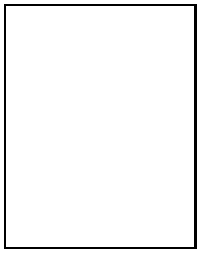
[31] ITU-R, *ITU-R Recommendation BS.1116. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, 1997.

[32] ITU-R, *ITU-R Recommendation BS.1534. Method for the subjective assessment of intermediate quality levels of coding systems*, 2003.

[33] D. Kirby and K. Watanabe, "Formal subjective testing of the MPEG-2 NBC multichannel coding algorithm," *Proceedings of the 102th International Conference of the Audio Engineering Society*, Munich, March 1997, preprint #4418.

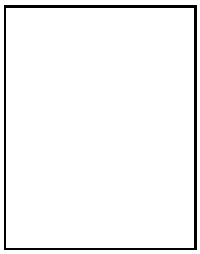
[34] F. P. Myburg, *Design of a scalable parametric audio coder*, Ph.D. thesis, Technische Universiteit Eindhoven, 2004.

[35] M. Goto, H. Hashigushi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," *Proceedings of the International Conference on Music Information Retrieval*, October 2002.



Olivier Derrien received the Eng. Degree from ENST (National School of Engineering in Telecommunications), Paris, France, in 1998, and the Ph.D. from the same school in 2002 in the area of audio processing. In 2002-03, he was teaching and research Assistant at the University of Paris-XI, Orsay, France.

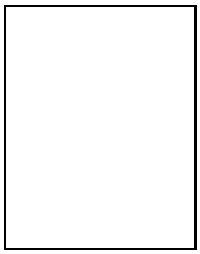
He joined the University of Toulon, France, in September 2003 as an Associate Professor in the field of telecommunications. His research interests inlcude audio and multimedia signal processing, especially audio coding and music recognition.



Pierre Duhamel (Fellow, IEEE, 1998) received the Eng. Degree in Electrical Engineering from the National Institute for Applied Sciences (INSA) Rennes, France in 1975, the Dr. Eng. Degree in 1978, and the Doctorat ès sciences degree in 1986, both from Orsay University, Orsay, France.

From 1993 to Sept. 2000, he has been professor at ENST, Paris (National School of Engineering in Telecommunications) with research activities focused on signal processing for communications. He was head of the Signal and Image processing Department from 1997 to 2000. He is now with CNRS/LSS (Laboratoire de Signaux et Systemes, Gif sur Yvette, France), where he is developing studies in signal processing for communications (including equalization, iterative decoding, multicarrier systems) and signal/image processing for multimedia applications, including source coding, joint source/channel coding, watermarking, and audio processing.

Dr. Duhamel was chairman of the DSP committee from 1996 to 1998, and a member of the SP for Com committee until 2001. He was an associate Editor of the IEEE Transactions on Signal Processing from 1989 to 1991, an associate Editor for the IEEE Signal Processing Letters, and a guest editor for the special issue of the IEEE Trans. on SP on wavelets.



Gaël Richard (Member, IEEE, 2002) received the State Engineering degree from the cole Nationale Suprieure des Tlcommunications (ENST), Paris, France, in 1990 and the Ph.D. degree from LIMSI-CNRS, University of Paris-XI, in 1994 in the area of speech synthesis. He received the Habilitation Diriger des Recherches degree from the University of Paris XI in September 2001.

After the completion of his Ph.D., he spent two years at the CAIP Center, Rutgers University, Piscataway, NJ, in the speech processing group of Prof. J. Flanagan, where he explored innovative approaches for speech production. Between 1997 and 2001, he successively worked for Matra Nortel Communications and for Philips Consumer Comunications. In particular, he was the project manager of several large-scale European projects in the field of multimodal verification and speech processing.

He joined the Department of Signal and Image Processing, ENST, as an Associate Professor in the field of audio and multimedia signals processing. He is co-author of over 50 papers and inventor in a number of patents, he is also one of the expert of the European commission in the field of man/machine interfaces.