A new model-based algorithm for optimizing the MPEG-AAC in MS-stereo

Olivier Derrien, Gael Richard

A new model-based algorithm for optimizing the MPEG-AAC in MS-stereo

Olivier Derrien and Gaël Richard, *Senior Member, IEEE*

Abstract— **In this paper, a new model-based algorithm for optimizing the MPEG-Advanced Audio Coder (AAC) in MS-stereo mode is presented. This algorithm is an extension to stereo signals of prior work on a statistical model of quantization noise. Traditionally, MS-stereo coding approaches replace the Left (L) and Right (R) channels by the Middle (M) and Sides (S) channels, each channel being independently processed, almost like a monophonic signal. In contrast, our method proposes a global approach for coding both channels in the same process. A model for the quantization error allows us to tune the quantizers on channels M and S with respect to a distortion constraint on the reconstructed channels L and R as they will appear in the decoder. This approach leads to a more efficient perceptual noise-shaping and avoids using complex psychoacoustic models built on the M and S channels. Furthermore, it provides a straightforward scheme to choose between LR and MS modes in each subband for each frame. Subjective listening tests prove that the coding efficiency at a medium bitrate (96 kbits/s for both channels) is significantly better with our algorithm than with the standard algorithm, without increase of complexity.**

Keywords— **Perceptual audio coding, MPEG-AAC, MS-stereo, statistical model, quantization, scalefactor, bitrate constraint, distortion constraint, optimization algorithm.**

I. Introduction

The MPEG-4 Advanced Audio Coder (AAC) is the latest international standard for high-quality lossy audio coding [1], [2]. Its application field is still expanding, including consumer audio equipment and digital video broadcasting. This codec has been derived in several *profiles* i.e. variations, for different applications: Low Complexity (LC-AAC), Low Delay (LD-AAC), High Efficiency (HE-AAC/AACPlus) etc. The MPEG-AAC is a frame-based transform-coder. Its apparent complexity is due to a large variety of coding parameters, which make the optimization process difficult to engineer and recent publications show that AAC optimization is still a current issue [3].

The MPEG-AAC is a multichannel codec, designed for stereo and *surround* audio applications. An AAC audio stream can include single channels and channel pairs. A single channel corresponds to a monophonic audio scene, a channel pair to a stereophonic scene (Left and Right channels). With the basic coding scheme for a channel pair, called *LR-stereo* in MPEG-AAC, each channel is processed as a monophonic signal. However, when a stereo signal exhibits significant inter-channel redundancy, the LR mode is quite ineffective. Improving the coding efficiency by removing the redundancy is possible with stereo coding modes.

A popular method for inter-channel decorrelation is the sum-difference transformation [4]. This technique, also referred to as *MS joint channel coding*, consists of a linear combination of the *Left* (L) and *Right* (R) channels to get *Middle* (M) and *Sides* (S) channels. M and S are coded instead of L and R and the reverse transformation is performed at the decoder side. In MPEG-AAC, LR and MS mode can be used alternatively for each frequency subband and each frame. The moderate coding gain is compensated by a small amount of side-information (one bit per subband). Other linear transformations have been proposed in the literature: Inter-channel decorrelation with a Karhunen-Loeve transform [5], inter-channel prediction [6], [7], and more recently a time-aligned version of the MS transformation [8]. With all these techniques, the coding gain is increased for some signals, but at the expense of additional side information.

Parametric stereo coding is another popular scheme for increasing the coding efficiency. A core monophonic coder is used in combination with additional parameters that describe the stereo information. The resulting auxiliary bitstream usually requires very few coding bits, but the original signals in channels L and R can not be totally recovered, even for very high bitrates. Thus, parametric stereo schemes are suitable for low bitrate applications. Originally, a simple parametric stereo mode, called *Intensity Stereo* (IS), was specified in the MPEG-AAC standard. It consists of coding only the M channel and an inter-channel intensity difference parameter for each subband. Since, many studies have been carried out on parametric stereo (see for instance [9], [10]) and in the latest extension of MPEG-AAC, called HE-AAC v2 [11], the parametric stereo mode can be considered as an improved version of the original IS mode: more parameters can be used to describe the stereo image (intensity difference, cross-correlation, phase/time difference). However, the typical bitrate for the HE-AAC v2 is quite low: 24kbps for both channels.

In this paper, we consider high-quality/high-bitrate applications and focus on the MS-stereo mode for the MPEG-AAC, especially on the implementation of the optimization algorithm which is strongly related to the coding efficiency. In a previous paper [12], we proposed a new algorithm for the single channel case, based on a statistical model of the quantization noise. In the informative annex of the MPEG-AAC standard [1], an implementation of the coding algorithm is described. In this paper, it will be referred to as the *standard* algorithm. Compared to this algorithm, our method exhibits a lower complexity and a better sound quality for the same bitrate. In this paper, we extend this model to the MS-stereo case, and propose a new efficient algorithm for coding a channel pair.

This article is divided in three parts. First, we briefly describe the MPEG-AAC codec and the MS-stereo mode.

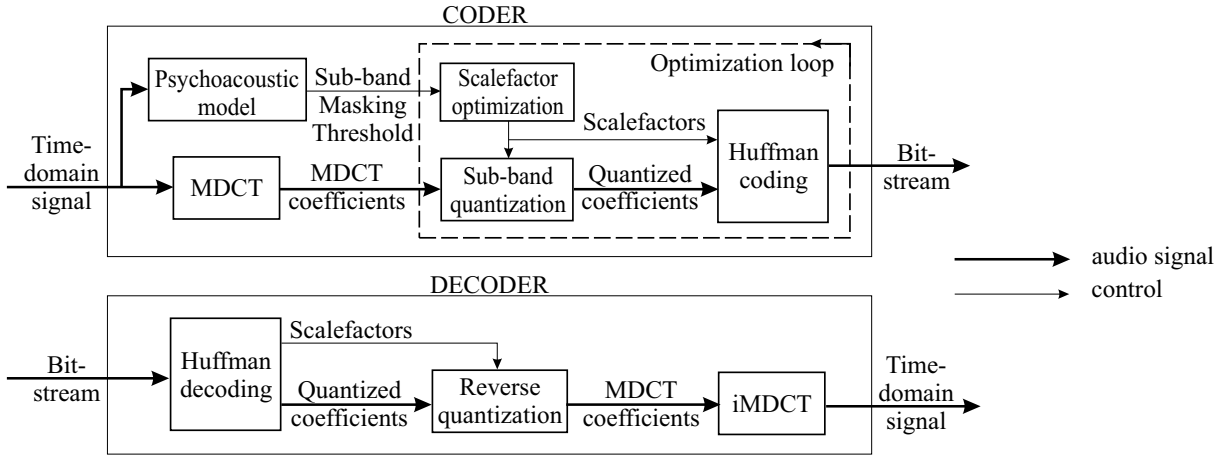

Fig. 1. Synopsis of a MPEG AAC codec.

Then, after recalling the main results of the monophonic model, we describe our stereophonic model and the new optimization algorithm. Finally, we compare our algorithm to the standard MPEG-AAC, both in terms of audio quality and computational complexity.

II. MPEG-AAC MS-stereo mode

A. Quantization and coding

Figure 1 presents the general scheme of a MPEG-AAC codec. The audio signal is segmented in variable-length analysis windows (256 or 2048 samples) with 50% overlap. Over each window, the signal is transformed in a frequency domain with a Modified Discrete Cosine Transform (MDCT) [13]. In this paper, we denote $X(k)$ the MDCT coefficients corresponding to a single channel, over the current analysis window. k is a frequency index. Variable length frequency subbands are defined as non-overlapping subsets of frequency indexes: $k \in \{k_{\min}(s) \cdots k_{\max}(s)\}$ where s is a subband index. Subband width K increases along the frequency scale.

The MDCT coefficients are quantized subband by subband according to:

$$i(k) = \mathcal{R}\left(\left[\frac{X(k)}{A(s)}\right]^{\frac{3}{4}}\right) \tag{1}$$

where $A(s)$ is a scaling parameter, $\mathcal{R}$ is a rounding function and $i(k)$ are the quantization indexes. $A(s)$ follows a logarithmic scale:

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

where ϕ is an integer parameter called *scalefactor*. The rounding function is not set by the MPEG standard. The function which minimizes the quantization error is defined in [12]. A sub-optimal function is proposed in the informative annex of the MPEG document [1].

Both quantization index $i(k)$ and scalefactor $\phi(s)$ are coded with a noiseless Huffman coding module. Coded audio data are segmented in frames. One frame corresponds either to a single 2048-samples window or to a sequence of 8 256-samples window. Optimizing the coding process consists of finding the scaling parameters $A(s)$ which maximize the audio quality under a bitrate constraint. This is generally implemented with an iterative algorithm including quantization and Huffman coding modules, and a psycho-acoustic model. Both psychoacoustic model and optimization algorithm are not specified in the standard, in order to allow for future advances in technology that will improve the coding efficiency.

B. The MS-stereo mode

When the audio signal is a channel pair, we denote $X_L(k)$ and $X_R(k)$ the MDCT coefficients corresponding respectively to channels *Left* and *Right*. The MS transformation is defined by:

$$\begin{cases} X_M(k) &= \frac{1}{2}\left[X_L(k) + X_R(k)\right] \\ X_S(k) &= \frac{1}{2}\left[X_L(k) - X_R(k)\right] \end{cases} \tag{3}$$

It can be used independently for each subband. A one-bit flag per subband indicates whether the MS transformation is used. In MS mode, X_M and X_S are quantized and coded instead of X_L and X_R. On the decoder side, the reconstructed MDCT coefficients $\hat{X}_M$ and $\hat{X}_S$ are obtained after the reverse quantization process. Finally the reverse MS transformation is performed:

$$\begin{cases} \hat{X}_L(k) &= \hat{X}_M(k) + \hat{X}_S(k) \\ \hat{X}_R(k) &= \hat{X}_M(k) - \hat{X}_S(k) \end{cases} \tag{4}$$

Compared to the single channel case, the MS-stereo mode raises three additional problems: 1) the LR/MS decision, 2) the hearing model and 3) the inter-channel bit-allocation. The classical approach was originally proposed by J.D. Johnston *et al.* [4], [14]: The MS mode is enabled when the energy difference between channels M and S exceeds a given threshold. Masking thresholds for M and S are computed by extending the monophonic psychoacoustic model. The inter-channel bit-allocation is performed according to a Perceptual Entropy criterion (see also [15]).

Then, a single channel noise-shaping algorithm is applied twice to M and S.

In the standard algorithm, some improvements have been made concerning problems 1) and 3): LR and MS channels are optimized in an iterative process which uses two nested-loops and a local decoder inside the outer-loop. The outer-loop (distortion loop) changes the scalefactor values independently for each channel (L,R,M,S) according to a Noise-To-Mask criterion. The inner-loop (bitrate loop) performs a global translation of the scalefactors for all channels in parallel, in order to meet the global bitrate constraint (solves problem 3). On each iteration, quantization and Huffman coding are performed for channels LR and MS, and the mode which minimizes the number of coding bits in each subband is selected (solves problem 1).

An improvement to this framework has been proposed by C.M. Liu *et al.* [16], applied to the MPEG-1 Layer III codec, which is very close to the MPEG-AAC. The main advances are: a new method for computing the masking threshold for M and S (solves problem 2), an inter-channel bit-allocation based on a new criterion called *Allocation Entropy* (solves problem 3), and a new intra-channel noise-shaping process.

Our method is radically different on three major issues: First, it relies on a specific MS distortion model which allows us to tune the quantization for both channels at the same time. Second, with our method, the inter-channel bit-allocation problem (problem 3) is solved jointly with the noise-shaping process. Third, we consider only the distortion constraint on channels L and R, even in the MS-mode. Thus, problem 2) is no more an issue, as psychoacoustics only involve channels L and R.

III. Description of the new coding algorithm

A. Single channel error model

In this section, we briefly recall the main results of our prior work on the single-channel case (see [12] for more details). The quantization error in the transform domain is defined by:

$$\varepsilon(k) = \hat{X}(k) - X(k) \tag{5}$$

and the error energy in subband s is:

$$E_\varepsilon(s) = \sum_{k=k_{\min}(s)}^{k_{\max}(s)} \varepsilon^2(k) \tag{6}$$

The usual criterion for evaluating the perceived distortion in one subband is the distance between $E_\varepsilon(s)$ and a so-called *masking threshold* $T_m(s)$, computed by the psychoacoustic model. If $E_\varepsilon(s) \leq T_m(s)$, the masking constraint is verified, and no distortion will be perceived in this frequency subband. As the MPEG-AAC usually operates in a fixed-bitrate mode, the available bitrate is not sufficient for the masking constraint to be verified in each subband.

The fixed-bitrate problem can be efficiently tackled by solving successive variable-bitrate problems: At each step of an iterative process, a distortion level per subband $T(s)$ is defined and a fast method is used to solve a variable-bitrate problem, i.e. to determine the set of scaling parameters which minimizes the bitrate under a distortion constraint:

$$E_\varepsilon(s) \leq T(s) \tag{7}$$

At the first step of the iterative process, $T(s)$ is initialized to $T_m(s)$. If the resulting bitrate matches the bitrate constraint, the coding problem is solved. Else, the distortion levels $T(s)$ are raised until the bitrate constraint is verified.

Finding the exact solution to the variable-bitrate problem is practically too much time-consuming and therefore, it is often preferred to find a near-optimal solution with a fast method. For that purpose, a new error model was developed. This method is *statistically optimal*, and thus practically near-optimal.

We assume that the variable-bitrate problem can be solved independently in each subband, which is almost true. In the remaining of this paper, we omit the subband index s, but *subband dependant variables are noted with a bold font.* Three different quantization modes have to be considered:

1) *High resolution*: When the scaling parameter is small enough for the quantizer to work in *high resolution* mode (which means that the energy of the quantization error is small compared to the energy of the input signal), all quantization indexes $i(k)$ are greater than zero. We consider the error coefficients $\varepsilon(k)$ as random variables. The energy $\mathbf{E}_\varepsilon$ is also a random variable, and the strict distortion constraint (7) is replaced by a statistical version:

$$\text{Prob}\left\{\mathbf{E}_\varepsilon \leq \mathbf{T}\right\} \geq \alpha \tag{8}$$

where $\alpha \in [0.5, 1]$ is a confidence parameter. It reflects the confidence we have on the masking threshold. α close to 1 means that the threshold is judged reliable, α close to 0.5 means that the threshold is not reliable. We showed in [12] that $\alpha = 1$ results in a high bitrate, whereas $\alpha = 0.9$ significantly reduces the bitrate for approximately the same error level.

The probability density function (pdf) of $\mathbf{E}_\varepsilon$ must be known to solve (8). Its exact expression would be far too complex, so we chose a simple model. Equation (6) shows that, if $\varepsilon(k)$ are independent and equally distributed, and if $\mathbf{K}$ (number of MDCT coefficients) is large enough, $\mathbf{E}_\varepsilon$ will follow a Gaussian law according to the central-limit theorem. Its mean and variance are:

$$\boldsymbol{\mu} \simeq \mathbf{K}\,\mathbb{E}\left[\varepsilon^2\right] \tag{9}$$

$$\boldsymbol{\sigma}^2 \simeq \mathbf{K}\left(\mathbb{E}\left[\varepsilon^4\right] - \mathbb{E}\left[\varepsilon^2\right]^2\right) \tag{10}$$

We have also considered a nonasymptotic model using a Gamma-law. With this finer model, there is no assumption made on $\mathbf{K}$. Both models are equivalent on large subbands. We expected similar performances on large subbands and an improvement on narrow subbands. However, as we observed no significant improvement, we finally chose the simple Gaussian model.

In high-resolution, approximations for the second- and fourth-order moments of the quantization error can be obtained, assuming that the rounding error is a white and uniformly distributed random variable:

$$\mathbb{E}\left[\varepsilon^2\right] \simeq a_2 \mathbf{m}_{\frac{1}{2}} \mathbf{A}^{\frac{3}{2}} \quad (11)$$

$$\mathbb{E}\left[\varepsilon^4\right] \simeq a_4 \mathbf{m}_1 \mathbf{A}^3 \quad (12)$$

a_2 and a_4 are multiplying factors which depend on the rounding function $\mathcal{R}$, $\mathbf{m}_p$ is defined in the current subband as:

$$\mathbf{m}_p = \frac{1}{\mathbf{K}} \sum_k |X(k)|^p \quad (13)$$

and $\mathbf{A}$ is the scaling parameter. For the sub-optimal rounding function proposed in the MPEG document, the analytic expression for the multiplying factors is:

$$a_p = \frac{4^p}{(p+1)3^p}\left[(1-N_m)^{p+1} + N_m^{p+1}\right]$$

where $N_m = 0.4054$, referred to as the *magic number* in [1]. For the optimal rounding function, the analytic expression for the multiplying factors is:

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

As $\mathbf{E}_\varepsilon$ is modelled with a Gaussian law, the distortion constraint (8) is equivalent to:

$$\boldsymbol{\mu} + \beta\boldsymbol{\sigma} \leq \mathbf{T} \quad (14)$$

where β is a secondary parameter depending on α:

$$\beta = \sqrt{2}\,\mathrm{Erf}^{-1}(2\alpha - 1) \quad (15)$$

Erf is the standard error function [17]. We assume that the bitrate is a decreasing function of $\mathbf{A}$. Then, the near-optimal value of the scaling parameter $\mathbf{A}_{opt}$ is obtained when (14) is an equality. We combine equations (9) and (10) with equations (11) and (12). We get:

$$\mathbf{A}_{opt} \simeq \left(\frac{\mathbf{T}}{\mathbf{K}a_2\mathbf{m}_{\frac{1}{2}} + \beta\sqrt{2\mathbf{K}(a_4\mathbf{m}_1 - a_2^2\mathbf{m}_{\frac{1}{2}}^2)}}\right)^{\frac{2}{3}} \quad (16)$$

2) *Dead zone*: When the scaling parameter is large enough for the quantization indexes $i(k)$ to be all zero, the quantizer is in the *dead zone*. The bitrate is zero for this subband, the output of the quantizer is also zero. The quantization error is the input signal itself, and the error energy is given by:

$$\mathbf{E}_\varepsilon = \mathbf{E}_X = \sum_k X(k)^2 \quad (17)$$

3) *Transition mode*: Between the high-resolution mode and the dead-zone, we do not propose a specific model, we simply extend the other two models. We consider that when $\mathbf{A} \leq \mathbf{A}_0$, the *high-resolution* expression (16) is valid, and when $\mathbf{A} \geq \mathbf{A}_0$, the *dead zone* expression (17) is valid. $\mathbf{A}_0$ is chosen at the junction of both modes:

$$\mathbf{A}_0 = \left(\frac{\mathbf{E}_X}{\mathbf{K}a_2\mathbf{m}_{\frac{1}{2}} + \beta\sqrt{2\mathbf{K}(a_4\mathbf{m}_1 - a_2^2\mathbf{m}_{\frac{1}{2}}^2)}}\right)^{\frac{2}{3}} \quad (18)$$

Finally, the solution of the variable-bitrate problem is:

a) If $\mathbf{T} < \mathbf{E}_X$, the nearly-optimal scaling parameter value $\mathbf{A}_{opt}$ is given by equation (16).

b) If $\mathbf{T} \geq \mathbf{E}_X$, we can choose any scaling parameter value greater than $\mathbf{A}_0$, given by equation (18).

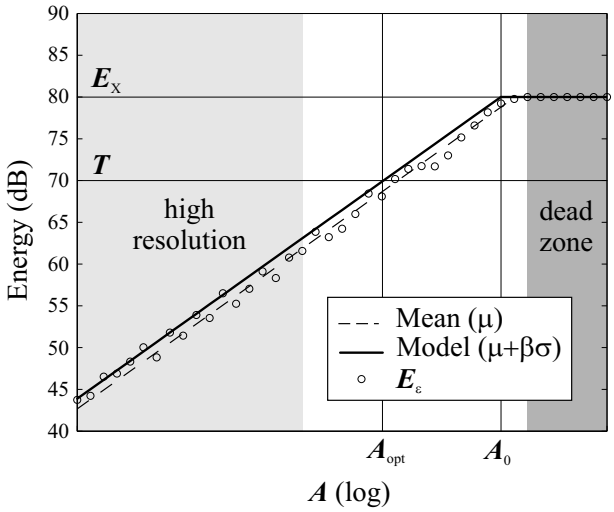

Fig. 2. Example of the distortion function and quantization model.

This model is illustrated on Figure 2, where the exact values of the error energy for a 16-coefficients subband and the model estimation as functions of the scaling parameter are drawn. We chose $\alpha = 0.9$. The *model* curve corresponds to $\boldsymbol{\mu} + \beta\boldsymbol{\sigma}$ when $\mathbf{A} < \mathbf{A}_0$ and to $\mathbf{E}_X$ when $\mathbf{A} \geq \mathbf{A}_0$.

The typical SNR for a fixed-bitrate coder is 10 dB. We can see on figure 2 that this corresponds to the transition mode. But one can also see that extending the high-resolution and dead-zone models is accurate.

B. Extending the error model to MS stereo

In this section, we present the new MS-stereo model for the quantization noise. When MS-stereo is enabled for a particular subband, channels M and S are quantized instead of L and R. The variable-bitrate coding problem consists of minimizing the total bitrate for both channels M and S, under a distortion constraint. The distortion constraint is evaluated after the reverse MS-transformation:

$$\begin{cases} \mathbf{E}_{\varepsilon_L} \leq \mathbf{T}_L \\ \mathbf{E}_{\varepsilon_R} \leq \mathbf{T}_R \end{cases} \quad (19)$$

The quantization error samples in the transform domain are $\varepsilon_M(k)$ and $\varepsilon_S(k)$. After the reverse MS-transformation, the quantization error samples are:

$$\begin{cases} \varepsilon_L(k) = \varepsilon_M(k) + \varepsilon_S(k) \\ \varepsilon_R(k) = \varepsilon_M(k) - \varepsilon_S(k) \end{cases} \quad (20)$$

and the error energy in each channel is:

$$\begin{cases} \mathbf{E}_{\varepsilon_L} &= \sum_k \varepsilon_L^2(k) \\ \mathbf{E}_{\varepsilon_R} &= \sum_k \varepsilon_R^2(k) \end{cases} \tag{21}$$

We now have to consider four different situations:

1) *Full high-resolution*: When the quantizers on channels M and S are in high-resolution mode, $\mathbf{E}_{\varepsilon_L}$ and $\mathbf{E}_{\varepsilon_R}$ are considered as random variables, and the statistical distortion constraints are:

$$\begin{cases} \text{Prob}\left\{\mathbf{E}_{\varepsilon_L} \leq \mathbf{T}_L\right\} \geq \alpha \\ \text{Prob}\left\{\mathbf{E}_{\varepsilon_R} \leq \mathbf{T}_R\right\} \geq \alpha \end{cases} \tag{22}$$

Using the same hypothesis as with the single channel model, we assume that $\mathbf{E}_{\varepsilon_L}$ (resp. $\mathbf{E}_{\varepsilon_R}$) follows a Gaussian law. Thus, (22) can be written as :

$$\begin{cases} \boldsymbol{\mu}_L + \beta\,\boldsymbol{\sigma}_L \leq \mathbf{T}_L \\ \boldsymbol{\mu}_R + \beta\,\boldsymbol{\sigma}_R \leq \mathbf{T}_R \end{cases} \tag{23}$$

The parameters $\boldsymbol{\mu}_L$ and $\boldsymbol{\sigma}_L^2$ (resp. $\boldsymbol{\mu}_R$ and $\boldsymbol{\sigma}_R^2$), given by equations (9) and (10), depend on the moments $\mathbb{E}\left[\varepsilon_L^2\right]$ and $\mathbb{E}\left[\varepsilon_L^4\right]$ (resp. $\mathbb{E}\left[\varepsilon_R^2\right]$ and $\mathbb{E}\left[\varepsilon_R^4\right]$). But the moments of the quantization error as functions of the scaling parameter, given by equations (11) and (12), involve channels M and S. Thus, $\boldsymbol{\mu}_L$, $\boldsymbol{\mu}_R$, $\boldsymbol{\sigma}_L^2$ and $\boldsymbol{\sigma}_R^2$ must be re-written as functions of the error moments on M and S. Here, we need a new hypothesis on the correlation between the error samples $\varepsilon_M(k)$ and $\varepsilon_S(k)$. When a quantizer is in high-resolution mode, the quantization error is approximated to be statistically independent from the input signal [18]. Then, assuming that $\varepsilon_M(k)$ and $\varepsilon_S(k)$ are independent variables is reasonable. From equation (20), we get:

$$\mathbb{E}\left[\varepsilon_L^2\right] \simeq \mathbb{E}\left[\varepsilon_R^2\right] \simeq \mathbb{E}\left[\varepsilon_M^2\right] + \mathbb{E}\left[\varepsilon_S^2\right] \tag{24}$$

$$\mathbb{E}\left[\varepsilon_L^4\right] \simeq \mathbb{E}\left[\varepsilon_R^4\right] \simeq \mathbb{E}\left[\varepsilon_M^4\right] + \mathbb{E}\left[\varepsilon_S^4\right] + 6\mathbb{E}\left[\varepsilon_M^2\right]\mathbb{E}\left[\varepsilon_S^2\right] \tag{25}$$

Combining these equations with (9) and (10) applied to channels L-R and with (11) and (12) applied to channels M-S, we get:

$$\boldsymbol{\mu}_L \simeq \boldsymbol{\mu}_R \simeq \boldsymbol{K} a_2 \left(\mathbf{m}_{\frac{1}{2}M}\mathbf{A}_M^{\frac{3}{2}} + \mathbf{m}_{\frac{1}{2}S}\mathbf{A}_S^{\frac{3}{2}}\right) \tag{26}$$

$$\boldsymbol{\sigma}_L^2 \simeq \boldsymbol{\sigma}_R^2 \simeq \boldsymbol{K}\left[\boldsymbol{\delta}_M\mathbf{A}_M^3 + \boldsymbol{\delta}_S\mathbf{A}_S^3 + 4\boldsymbol{\delta}_{MS}\mathbf{A}_M^{\frac{3}{2}}\mathbf{A}_S^{\frac{3}{2}}\right] \tag{27}$$

the parameters $\boldsymbol{\delta}_M$, $\boldsymbol{\delta}_S$ and $\boldsymbol{\delta}_{MS}$ are :

$$\begin{aligned} \boldsymbol{\delta}_M &= a_4\mathbf{m}_{1M} - a_2^2\mathbf{m}_{\frac{1}{2}M}^2 \\ \boldsymbol{\delta}_S &= a_4\mathbf{m}_{1S} - a_2^2\mathbf{m}_{\frac{1}{2}S}^2 \\ \boldsymbol{\delta}_{MS} &= a_2^2\mathbf{m}_{\frac{1}{2}M}\mathbf{m}_{\frac{1}{2}S} \end{aligned} \tag{28}$$

As one can see, $\mathbf{E}_{\varepsilon_L}$ and $\mathbf{E}_{\varepsilon_R}$ have approximately the same mean and variance. Thus, the distortion constraints (23) are equivalent to a single equation :

$$\boldsymbol{\mu}_L + \beta\,\boldsymbol{\sigma}_L \simeq \boldsymbol{\mu}_R + \beta\,\boldsymbol{\sigma}_R \leq \min\left(\mathbf{T}_L, \mathbf{T}_R\right) \tag{29}$$

Combining this equation with equations (26) and (27) leads to a new equation which cannot be easily simplified. To circumvent this difficulty, we propose an approximation of $\boldsymbol{\sigma}_L$ and $\boldsymbol{\sigma}_R$ with Taylor series. We denote:

$$\xi = \left(\frac{\boldsymbol{A}_S}{\boldsymbol{A}_M}\right)^{\frac{3}{2}} - 1 \tag{30}$$

From equation (27), we get:

$$\boldsymbol{\sigma}_L \simeq \boldsymbol{\sigma}_R \simeq \sqrt{\boldsymbol{K}\boldsymbol{\Delta}}\,\mathbf{A}_M^{\frac{3}{2}}\left[1 + \frac{2\boldsymbol{\delta}_S + 4\boldsymbol{\delta}_{MS}}{\boldsymbol{\Delta}}\,\xi + \frac{\boldsymbol{\delta}_S}{\boldsymbol{\Delta}}\,\xi^2\right]^{\frac{1}{2}} \tag{31}$$

where:

$$\boldsymbol{\Delta} = \boldsymbol{\delta}_M + \boldsymbol{\delta}_S + 4\boldsymbol{\delta}_{MS} \tag{32}$$

When the audio signal has a strong stereo effect, $\mathbf{A}_M$ and $\mathbf{A}_S$ should have close values, which would lead to $\xi \simeq 0$. It appears that this is true even with a weak stereo effect. Using first order Taylor series, we get:

$$\left[1 + \frac{2\boldsymbol{\delta}_S + 4\boldsymbol{\delta}_{MS}}{\boldsymbol{\Delta}}\,\xi + \frac{\boldsymbol{\delta}_S}{\boldsymbol{\Delta}}\,\xi^2\right]^{\frac{1}{2}} \simeq 1 + \frac{\boldsymbol{\delta}_S + 2\boldsymbol{\delta}_{MS}}{\boldsymbol{\Delta}}\,\xi \tag{33}$$

and:

$$\boldsymbol{\sigma}_L \simeq \boldsymbol{\sigma}_R \simeq \boldsymbol{\gamma}_M\mathbf{A}_M^{\frac{3}{2}} + \boldsymbol{\gamma}_S\mathbf{A}_S^{\frac{3}{2}} \tag{34}$$

where parameters $\boldsymbol{\gamma}_M$ and $\boldsymbol{\gamma}_S$ are:

$$\begin{aligned} \boldsymbol{\gamma}_M &= \sqrt{\tfrac{\boldsymbol{K}}{\boldsymbol{\Delta}}}\left(\boldsymbol{\delta}_M + 2\boldsymbol{\delta}_{MS}\right) \\ \boldsymbol{\gamma}_S &= \sqrt{\tfrac{\boldsymbol{K}}{\boldsymbol{\Delta}}}\left(\boldsymbol{\delta}_S + 2\boldsymbol{\delta}_{MS}\right) \end{aligned} \tag{35}$$

On audio excerpt #8, which has the weakest stereo effect in our selection, coded at 48 kbits/s, we measured $\xi = -0.074 \pm 0.002$ (95 % confidence interval), and the maximum error measured for the linear approximation of $\boldsymbol{\sigma}_L$ and $\boldsymbol{\sigma}_R$ is 0.5%.

Finally, the distortion constraint (29) is equivalent to:

$$\left(\mathbf{K}a_2\mathbf{m}_{\frac{1}{2}M} + \beta\boldsymbol{\gamma}_M\right)\mathbf{A}_M^{\frac{3}{2}} + \left(\mathbf{K}a_2\mathbf{m}_{\frac{1}{2}S} + \beta\boldsymbol{\gamma}_S\right)\mathbf{A}_S^{\frac{3}{2}} \leq \mathbf{T} \tag{36}$$

with $\mathbf{T} = \min\left(\mathbf{T}_L, \mathbf{T}_R\right)$.

Solving the variable-bitrate problem in full high-resolution mode requires a model for the bitrate function, i.e. the amount of coding bits for a single channel, in one particular subband, as a function of the scaling parameter.

As the coding module uses 11 Huffman codebooks, codewords from codebook #11 being interleaved with escape sequences, building an analytical model for the bitrate function seems very difficult. So, we chose an empirical approach: We measured the bitrate function for each value of the subband width K, on a database of 8 audio excerpts, actually excerpts #1, 2, 3, 5, 6 in table I, and 3 other ones that were not retained for the listening tests. The conclusion is that the number of coding bits per subband can be reasonably modelled by a decreasing linear function of $\log\left(\mathbf{A}\right)$. On Figure 3, we plot the mean bitrate function

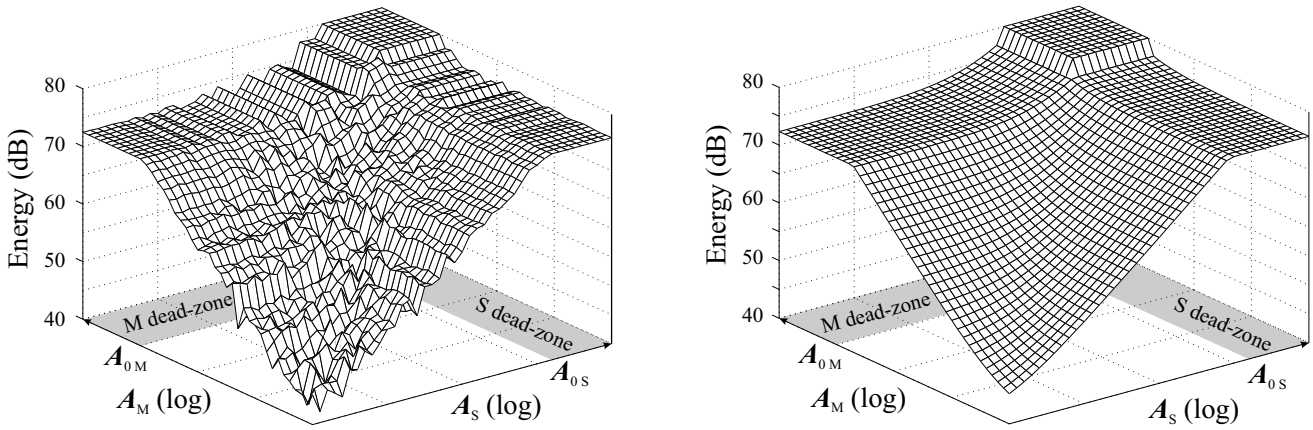

Fig. 4. Example of the distortion function (on the left) and model (on the right) for one output channel (Left).

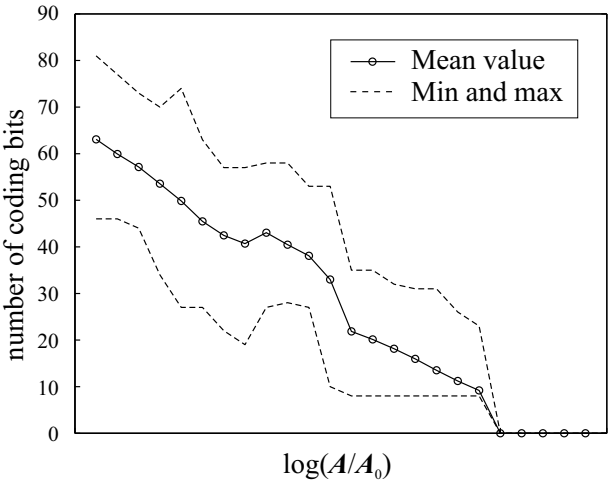

Fig. 3. Mean bitrate function and extreme values. Subband width: 16 coefficients. Audio excerpt #8.

and the extreme values for $K = 16$, for audio excerpt #8. One can see that the linear model is a reasonable approximation. However, the high variability justifies the use of an iterative process where the model is only used for optimizing the inter-channel bit-allocation.

Assuming a log-linear model for the bitrate as a function of the scaling parameter, it appears that minimizing the total bitrate is equivalent to maximizing $\log(\mathbf{A}_M \mathbf{A}_S)$ under the distortion constraint (36). The solution can be easily obtained with a Lagrange-multiplier maximization technique:

$$\mathbf{A}_{Mopt} \simeq \left(\frac{\mathbf{T}}{2\left(\mathbf{K} a_2 \mathbf{m}_{\frac{1}{2}M} + \beta \boldsymbol{\gamma}_M\right)} \right)^{\frac{2}{3}} \tag{37}$$

$$\mathbf{A}_{Sopt} \simeq \left(\frac{\mathbf{T}}{2\left(\mathbf{K} a_2 \mathbf{m}_{\frac{1}{2}S} + \beta \boldsymbol{\gamma}_S\right)} \right)^{\frac{2}{3}} \tag{38}$$

2) *Channel M high-resolution*: When the quantizer on channel M is in high-resolution mode, and the quantizer on channel S is in the dead-zone, $\mathbf{E}_{\varepsilon_M}$ is considered as a random variable and $\mathbf{E}_{\varepsilon_S}$ is a constant, equal to $\mathbf{E}_{X_S}$. The distortion constraints are:

$$\begin{cases} \mathbf{E}_{\varepsilon_L} = \mathbf{E}_{\varepsilon_M} + \mathbf{E}_{X_S} \leq \mathbf{T}_L \\ \mathbf{E}_{\varepsilon_R} = \mathbf{E}_{\varepsilon_M} + \mathbf{E}_{X_S} \leq \mathbf{T}_R \end{cases} \tag{39}$$

This coding problem is similar to a single-channel optimization on channel M, with the following distortion constraint:

$$\mathbf{E}_{\varepsilon_M} \leq \min\left(\mathbf{T}_L, \mathbf{T}_R\right) - \mathbf{E}_{X_S} \tag{40}$$

According to the single-channel model, the nearly-optimal solution is:

$$\mathbf{A}_{Mopt} \simeq \left(\frac{\mathbf{T} - \mathbf{E}_{X_S}}{\mathbf{K} a_2 \mathbf{m}_{\frac{1}{2}M} + \beta \sqrt{2\mathbf{K}(a_4 \mathbf{m}_{1M} - a_2^2 \mathbf{m}^2_{\frac{1}{2}M})}} \right)^{\frac{2}{3}} \tag{41}$$

3) *Channel S high-resolution*: When the quantizer on channel S is in high-resolution mode, and the quantizer on channel M is in the dead-zone, the coding problem is exactly similar to the previous one, and the nearly-optimal solution is:

$$\mathbf{A}_{Sopt} \simeq \left(\frac{\mathbf{T} - \mathbf{E}_{X_M}}{\mathbf{K} a_2 \mathbf{m}_{\frac{1}{2}S} + \beta \sqrt{2\mathbf{K}(a_4 \mathbf{m}_{1S} - a_2^2 \mathbf{m}^2_{\frac{1}{2}S})}} \right)^{\frac{2}{3}} \tag{42}$$

4) *Full dead-zone*: When both quantizers are in the dead-zone, any scaling parameters $\mathbf{A}_M \geq \mathbf{A}_{M_0}$ and $\mathbf{A}_S \geq \mathbf{A}_{S_0}$ are optimal.

This model is illustrated on Figure 4. On a 16-coefficients subband, for output channel L and for one long analysis window from audio excerpt #8, we draw the exact values of the error energy and the model estimation as functions of scaling parameters on channels M and S. One can see that the model is an accurate approximation of the actual distortion function.

C. Stereophonic optimization process for fixed bitrates

In the previous section, we have described an error model for the MS-stereo mode. Given distortion levels on chan-

nels L and R, it allows us to compute the values of the scaling parameters $A_M(s)$ and $A_S(s)$ which solve the variable-bitrate problem, i.e. minimize the number of coding bits under a distortion constraint. In this section, we revisit the fixed-bitrate problem: i.e. minimizing the perceived distortion under a bitrate constraint. As for the single-channel algorithm, this method is merely a single-loop process. The block-diagram of the optimization algorithm is presented on Figure 5. The notations are :

- $T_{mC}(s)$: Masking threshold for channel $C \in \{L, R\}$ and subband s, computed by the psychoacoustic model.
- $T_C^i(s)$: Distortion level for channel $C \in \{L, R\}$ and subband s at iteration i.
- $A_C^i(s)$: Scaling parameter (related to the scalefactor) for channel $C \in \{L, R, M, S\}$ and subband s at iteration i.
- $b_C^i(s)$: Required number of coding bits for channel $C \in \{L, R, M, S\}$ and subband s at iteration i, after quantization and Huffman coding.
- $B_{\max}$: Maximum number of coding bits per frame, depending on the output bitrate.

The computation of the distortion levels $T_C^{i+1}(s)$, for channels $C \in \{L, R\}$, from previous values $T_C^i(s)$, uses a process similar to that of the single-channel algorithm: For high SNR (1st phase), it is a *water-filling* technique [19] with a protection factor. For low SNR (2nd phase), a constant SNR degradation is performed. During the 1st phase, the water-filling technique retrieves bits from the sub-bands with the lowest signal energy in order to minimize the distortion on high-energy sub-bands. The protection factor is used to avoid large distortion levels at low frequencies. During the second phase, a uniform bit retrieval along subbands is performed in the case of very low bitrate constraint, but some noticeable distortions will then be clearly perceived.

The complete description of this process, applied both to channels L and R, is as follows, where $\tau(s)$ is the protection factor for each subband (see [12] for numerical values and implementation details). The protection threshold is defined by:

$$G(s) = \frac{E_X(s)}{\tau(s)}$$

The protection threshold can be interpreted as the maximum error energy required in each subband to preserve a perceptually-acceptable level of distortion.

- **1st phase**, until $T^i(s) < G(s)$ for at least one sub-band

$$T^i_{\min} = \min_s \left(T^i(s)\right)$$

$$T^{i+1}(s) = \min\left(\max\left(T^i(s), r_1\, T^i_{\min}\right), G(s)\right)$$

- **2nd phase**

$$T^{i+1}(s) = r_2\, T^i(s)$$

step-constants r_1 and r_2 have been set respectively to 1 dB and 0.25 dB.

IV. Performance evaluation

In this section, our coding algorithm is compared to the algorithm described in the informative annex of the MPEG

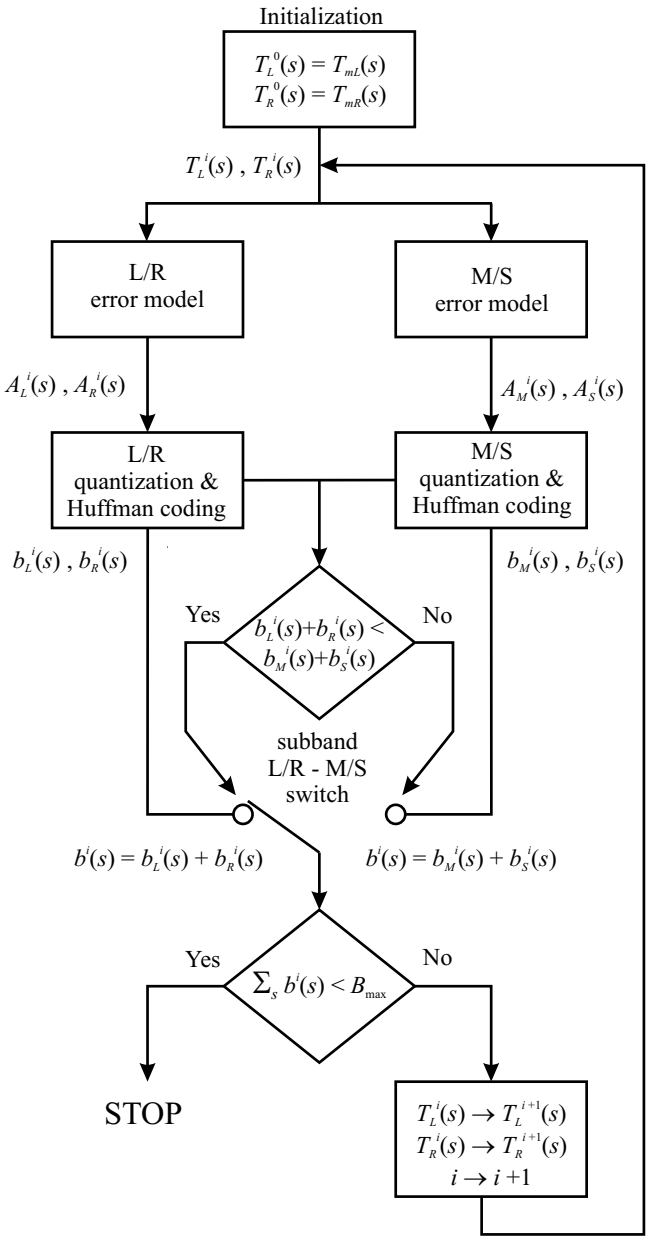

Fig. 5. Block-diagram of the optimization algorithm.

standard [1], referred to as the standard coder in this paper. We used the Low Complexity profile. The standard coder was chosen preferably to an embedded AAC coder, because it is a public implementation which allows a fair comparison: The only difference between both coders under test is the optimization algorithm. All other components are the same.

A. Subjective evaluation

The signal quality can be assessed using objective quality tests, but as mentioned in [22], the ultimate quality test of any audio compression technique is the human listener. In this work, we refer, to a large extent, to the ITU recommendation BS.1534-1 [21] (often referred as MUSHRA test) which is especially designed for the subjective assessment of intermediate quality audio coding systems. The

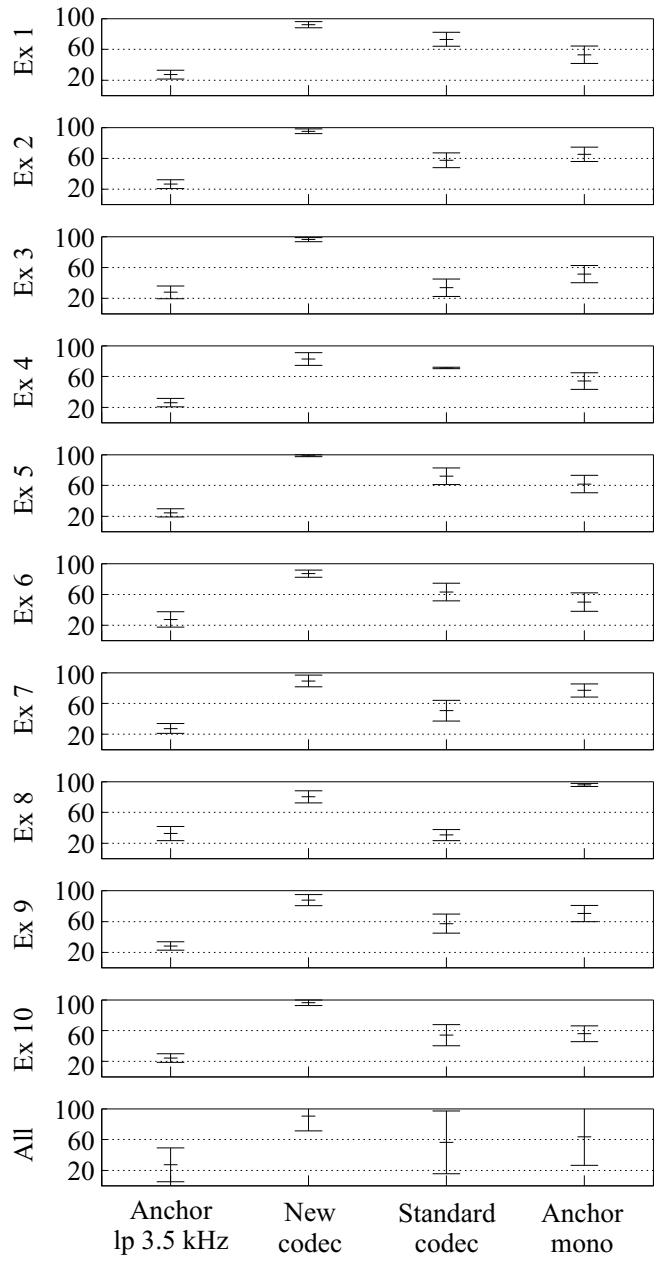

Fig. 6. MUSHRA test results with 95% confidence intervals for each audio excerpt.

subjective evaluation was carried out at a bit rate of 96 kbit/s since near transparent quality is obtained for both codecs at 128 kbit/s or higher bitrates.

Table I gives the list of the selected test material. The selection of audio test items was done by choosing a subset of excerpts where audio impairments of both coding schemes were the most audible and by favouring the widest variety of musical content. All excerpts are stereophonic and were played at a sampling rate of 48 kHz.

Our coder is subjectively evaluated and compared to the standard AAC coder and two anchor signals. The first one, required by the MUSHRA protocol, is a 3.5 kHz low-pass version of the reference signal. We chose to add a second anchor signal: a monophonic version of the reference signal, i.e. the average of Left and Right channels. The score given to this mono anchor is informative on the level of stereo of the test signal.

A total of 17 selected subjects participated to the listening test and scored the different signals according to their quality from score 0 (extremely poor quality) to 100 (transparent). Even if some of them were familiar with audio coding evaluation, all subjects underwent a training phase which allowed them to better identify the typical coding artefacts of the tested coders[1]. The participants were post-screened according to the score they have attributed to the reference: the data for listeners who gave a score under 80 (which correspond to the lowest mark for the category *excellent quality*) were discarded. As a consequence, 14 listeners were judged reliable and therefore kept for the results. Note that none of the authors have participated to the test.

The results of the subjective test are summarized in Figure 6. The results are given as mean absolute scores for each signal with 95% confidence intervals.

The main result is that the proposed coder provides a significantly better quality than the standard one for all test items. It is also interesting to notice that both codecs were judged better than the low-pass anchor, except for the standard codec on items #3 and 8. To our opinion, this illustrates the main weaknesses of the standard algorithm: On the one hand, when the input signal has a wide spectral content, for example the harpsichord accompaniment on item #3, the power spectral density (PSD) of the error can strongly vary from one frame to another, which creates *birdies-like* degradations. Since our method is temporally more stable, this inherent weakness is greatly reduced. On the other hand, when the input signal possesses a globally contrasted PSD (for example Suzanne Vega's voice on item #8), the standard algorithm produces an error with a globally smooth PSD, which creates *gaps* in the spectrogram of the coded signal. By contrast, our methods adapts the PSD of the error to the PSD of the input signal, and avoids such degradations.

The comparison with the mono anchor is more difficult to interpret as many listeners mentioned the difficulty to assess the reduction of stereophonic effect compared to *artefact-like* degradations. However, it seems that the score given to the mono anchor is actually related to the level of stereo: For item #8, which is nearly a monophonic signal, the mono anchor obtains almost 100%, and for items #1, 3 and 6, which provide a strong stereo effect, the score given to the mono anchor is the lowest. Furthermore, one can notice that the proposed coder is always judged better than the mono anchor, except for item #8 which is a very special case. This proves that our codec does not significantly degrade the stereophonic rendering in order to reduce traditional coding artefacts.

[1]In practice, the training phase is done in two phases. First, the listeners learn typical quality degradations due to bitrate reduction on typical signals (Low pass filtering, birdies and pre-echoes). Note that no specific artefacts associated with stereophonic degradations were included. Second, the subjects listen to all items included in the test, in which case stereophonic degradations are presented.

Id	Author	Identification	Style	Duration
1	The Beatles	Drive My Car	Pop-Rock	8.5 s
2	J.J. Cale	Cocaine	Pop-Rock	9.8 s
3	A. Vivaldi	Gloria	Choir	11.0 s
4	M. Marais	Le Labyrinthe	Viola da gamba	8.6 s
5	Anonymous	Saltarello	Medieval	7.6 s
6	Simon & Garfunkel	Sound of Silence	Singing voice	9.3 s
7	Supertramp	Goodbye Stranger	Pop-Rock	8.3 s
8	S. Vega	Toms dinner	Singing voice	9.5 s
9	H. Texier	Tzigane	Jazz	10.0 s
10	R. Galliano	Viaggio	Jazz	10.5 s

TABLE I
AUDIO MATERIAL FOR SUBJECTIVE EVALUATION.

Finally, it can be observed that relatively small confidence intervals are obtained and that they are in general smaller for the proposed coder than for the standard AAC coder. This may be explained by the fact that when an artefact is clearly audible (which is more often the case with the standard ACC coder than with our coder), the listeners often have a different perception on its acceptability and therefore use significantly different scores, leading to an increase of the confidence intervals for the standard AAC coder. The relatively small confidence intervals obtained is to our opinion the consequence of the specific training phase conducted beforehand by all listeners which in fact leads to a better agreement of the listeners during the test phase.

B. Complexity

In a previous paper [12], we showed that the optimization algorithm takes about 50% of the whole computation time and that our model-based algorithm for monophonic signals requires about 40% less computation time than the standard algorithm at 48 kbits/s.

In MS stereo, we possibly expect different results, because the quantization process is more complex than twice the monophonic case: The standard algorithm relies on three nested-loops (distortion loop for channel M, distortion loop for channel S, bitrate loop for both channels), with simple calculations inside each loop. Our model-based algorithm has only one bitrate loop, but the calculations inside the loop are more complex.

To evaluate the complexity, we measured the mean CPU time required for coding one analysis window, for the excerpts listed in Table I, and for a bitrate of 96 kbits/s.[2] On Figure 7, we plot the mean execution time and 95% confidence interval for each audio excerpt, and for all excerpts. The remaining computation time (window-switching, MDCT and psycho-acoustic model), which is common to both implementations, is approximately 0.5s.

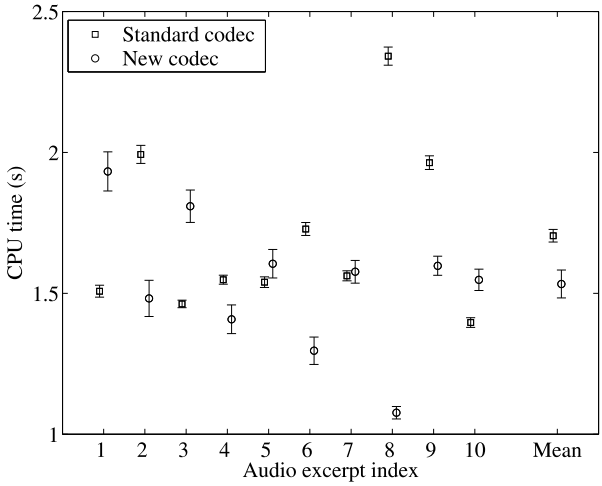

Fig. 7. Mean execution time and 95% confidence intervals for each audio excerpt.

One can see that the optimization algorithm takes about 70% of the whole computation time, which is more than in the monophonic case. For the optimization part, our algorithm performs better on excerpts #2, 4, 6, 8, 9, the standard algorithm on excerpts #1, 3, 5, 7, 10. In average, our algorithm is 10% faster. One can also notice that confidence intervals are larger with our algorithm. This point can be explained by the small value for the step constant r_2: In the second phase of the optimization, reached when psychoacoustics require a much larger amount of coding bits than available, we choose to slowly raise the distortion level in order to get a very progressive quality degradation, which increases the execution time. In contrast, in the standard algorithm, the variation applied to the scalefactors is raised at each iteration. This ensures a fast convergence, but results in a poor subjective quality when psychoacoustics require a larger amount of coding bits.

[2]Note 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 time-frequency transform (MDCT). Thus, the results might slightly differ with a compiled coder, and the total computation time would be lower with a fast MDCT scheme.

V. Conclusion

In this paper, we have described a new coding algorithm for the MPEG Advanced Audio Coding in MS-stereo mode, based on a subband model for the quantization noise. Our approach is radically different from the MPEG standard algorithm: we propose a global approach for coding both channels in the same process. First, a quantization error model allows us to tune the quantizers on channels M and S with respect to a distortion constraint on the reconstructed channels L and R as they will appear in the decoder. This approach leads to a more efficient perceptual noise-shaping and to avoid the use of complex psychoacoustic models built on the MS channels. Furthermore, it provides a straightforward scheme to choose between LR and MS modes in each subband for each frame.

Subjective listening tests performed with trained subjects prove that the coding efficiency at a medium bitrate (96 kbits/s for both channels) is significantly better with our algorithm, with no increase of complexity.

Our method is compatible with almost any psychoacoustic model. For our experimentations, we used the binaural extension of the model proposed in the MPEG standard, but further studies should focus on improving the psychoacoustic modelling of binaural effects, because this aspect is strongly related to the coding efficiency.

VI. Acknowledgements

The authors wish to thank all listeners who have accepted to participate to the listening tests. The authors are also grateful to the anonymous reviewers who greatly helped to improve the original manuscript.

References

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

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

[3] C. Bauer, M. Feller and G. Davidson, "Multidimensional Optimization of MPEG-4 AAC Encoding," *proc. IEEE ICASSP*, Toulouse, France, May 2006.

[4] J.D. Johnston, "Perceptual Transform Coding of Wideband Stereo Signals," *proc. IEEE ICASSP*, 1990.

[5] D. Yang, H. Ai, C. Kyriakakis, C. Faller and C.C.J. Kuo, "High-Fidelity Multichannel Audio Coding With Karhunen-Loève Transform" *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 365-380, July 2003.

[6] H. Fuchs, "Improving Joint Stereo Audio Coding By Adaptive Inter-Channel Prediction," *proc. IEEE WASPAA*, Mohonk, NY, October 1993.

[7] H. Fuchs, "Improving MPEG Audio Coding by Backward Adaptive Linear Stereo Prediction" *proc. 99th AES Convention*, 1995.

[8] J. Lindbom, J.H. Plasberg and R. Vafin, "Flexible Sum-Difference Stereo Coding Based on Time-Aligned Signal Components" *IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, New Paltz, New-York, October 2005.

[9] F. Baumgarte and C. Faller, "Binaural Cue Coding - Part I: Psychoacoustics Fundamentals and Design Principles" *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 509-519, November 2003.

[10] F. Baumgarte and C. Faller, "Binaural Cue Coding - Part II: Schemes and Applications" *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 520-531, November 2003.

[11] International Organization for Standardization, *ISO/IEC 14496-3:2005 (Information technology - Coding of audio-visual objects – Part 3: Audio)*, 2005.

[12] O. Derrien, P. Duhamel, M. Charbit and G. Richard, "A New Quantization Optimization Algorithm for the MPEG Advanced Audio Coder Using a Statistical Subband Model of the Quantization Noise," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 1328–1339, July 2006.

[13] 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.

[14] J.D. Johnston and A.J. Ferreira, "Sum-Difference Stereo Transform Coding, " *proc. IEEE ICASSP*, pp. 569-572, 1992.

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

[16] J.C. Liu, W.C. Lee and Y.H. Hsiao, " M/S Coding Based on Allocation Entropy, " *proc. DAFX'03*, London, UK, September 2003.

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

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

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

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

[21] ITU, *ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems,* 2003.

[22] 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, October 1995, preprint #4065.

[23] 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.

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