Parametric audio coding with exponentially damped sinusoids

Olivier Derrien, Roland Badeau, Gaël Richard

Parametric Audio Coding with Exponentially Damped Sinusoids

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

***Abstract*—Sinusoidal modeling is one of the most popular techniques for low bitrate audio coding. Usually, the sinusoidal parameters (amplitude, pulsation and phase of each sinusoidal component) are kept constant within a time segment. An alternative model, the so-called Exponentially-Damped Sinusoidal (EDS) model, includes an additional damping parameter for each sinusoidal component to better represent the signal characteristics. It was however never shown that the EDS model could be efficient for perceptual audio coding. To that aim, we propose in this paper an efficient analysis/synthesis framework with dynamic time-segmentation on transients and psychoacoustic modeling, and an asymptotically optimal entropy-constrained quantization method for the four sinusoid parameters (e.g including damping). We then apply this coding technique to real audio excerpts for a given entropy target corresponding to a low bitrate (20 kbits/s), and compare this method with a classical sinusoidal coding scheme using a constant-amplitude sinusoidal model and the perceptually weighted Matching Pursuit algorithm. Subjective listening tests show that the EDS model is more efficient on audio samples with fast transient content, and similar to the classical model for more stationary audio samples.**

***Index Terms*—Exponentially damped sinusoids, Quantization, Entropy, Parametric audio coding.**

I. Introduction

For low bitrate audio coding applications, parametric coders are an efficient alternative to transform coders. Classical transform coders (e.g. MPEG-1 Layer 3 [1] or MPEG-2/4 AAC [2]) use a time-frequency transform with a perfect reconstruction capability. The transform coefficients are quantized in order to perform a spectral-shaping of the quantization noise according to psychoacoustic considerations. Finally, a lossless entropy-coding stage transforms the set of quantization indexes (and some extra parameters, called side-information) into a bit-stream. Such coders are efficient for high-bitrate/high-quality coding, as the reconstructed signal tends to equal the original signal as the bitrate increases.

At low bitrate, the efficiency of transform coders collapses because of an excessive amount of side-information. In contrast, parametric coders model the signal with very few meaningful parameters. This method usually does not allow perfect reconstruction, but leads to a significantly sparser representation than classical time-frequency transforms. The remaining of the coder is similar to a transform coder: parameters are quantized and quantization indexes are entropy-coded. Many parametric signal models have been proposed in the literature, but the sinusoidal model [3] remains the most popular, because most real-world audio signals are dominated by tonal components. Traditionally, in sinusoidal models used for parametric coding, the amplitude of each component is kept constant within each analysis/synthesis time segment. Both parametric codecs included in the MPEG-4 Audio standard, HILN [4] and SSC [5], use a sinusoidal model combined with a residual signal model, often called noise model. In SSC, other features are added, namely a transient model using Meixner waveforms, and alignments of analysis/synthesis time-segments with onsets. In both HILN and SSC, sinusoidal parameters are quantized independently: frequency is quantized at just noticeable distortion [6], amplitude uses a log-uniform scalar quantizer, and phase uses a uniform scalar quantizer. This scheme requires a low computation time but is obviously less efficient than a vector quantizer since it does not take advantage of the statistical dependency between parameters. Recently, more efficient joint-scalar quantizers for sinusoidal parameters have been proposed [7], [8], which almost equal a vector quantizer in terms of entropy-distortion tradeoff but with a much lower complexity.

Some studies have shown that an evolution of the traditional sinusoidal model, the Exponentially Damped Sinusoidal (EDS) model, is an efficient alternative for modeling some audio signals [9], especially when the signal exhibits many onsets (i.e. sharp amplitude variations). The time-envelope of each sinusoidal component is an increasing or decreasing exponential, controlled by an additional damping parameter. A nice feature with the EDS model is that efficient estimation algorithms have been proposed. They can be roughly divided in two categories: iterative analysis-by-synthesis methods (Matching Pursuit) [10] and subspace-based methods (MUSIC, ESPRIT) [11], [12]. However, it was never shown that the EDS model could be efficient for perceptual audio coding. In a previous paper [13], we presented a new entropy-constrained joint-quantization scheme for EDS parameters. This method was restricted to the quantization of amplitude, phase and damping. We showed that this scheme was almost as efficient as a vector quantizer in terms of entropy / SNR trade-off, and its low complexity was compatible with audio coding applications.

In this paper, we propose an extended version of this study. Our contributions concern the analysis-synthesis framework and the quantization stage: we describe, under some reasonable simplifying hypothesis, an analytic solution for optimal entropy-constrained joint-quantization of the whole parameter set. Furthermore, we now take in consideration the auditory perception in both the EDS analysis stage and the quantization stage. We also demonstrate the merit of our solution compared

O. Derrien is with Université de Toulon / CNRS LMA, Marseille, France, email: derrien@lma.cnrs-mrs.fr.

R. Badeau and G. Richard are with Institut Mines-Télécom / Télécom ParisTech / CNRS LTCI, Paris, France, email: firstname.lastname@telecom-paristech.fr.

to the more traditional sinusoidal model.

Concerning the analysis-synthesis framework, three main difficulties must be taken into consideration: first, as explained in [12], the efficiency of the EDS model is conditioned to the accuracy of time-segmentation. Thus ideally, the analysis window should be rectangular and time-segments boundaries should be aligned with onsets, but the practical determination of boundaries location is still an issue. Second, the selection of sinusoidal components should be made according to psychoacoustic considerations. In other words, for a given number of components, only the most perceptually significant ones should be selected. Two different approaches have been proposed [14], [15], but they are not fully compatible with parametric audio coding requirements. Third, as explained in [16], the time-envelopes resulting for the synthesis with an EDS model are not smooth, and smoothness is desirable in the context of audio coding in order to avoid discontinuities at time-segment boundaries. In this paper, we propose an efficient analysis/synthesis framework for the EDS model which meets these three requirements.

Concerning quantization, the inclusion of an additional damping parameter in the signal model calls for a new quantization scheme. We propose an extension of the joint-scalar quantization method by Korten *et al.* [8], which allows jointly quantizing the four parameters of each sinusoid (amplitude, phase, damping and pulsation) under an entropy constraint according to psychoacoustic considerations. When the quantization stage is followed by a noiseless coder, the entropy of quantization indexes is an estimation of the output bitrate in the ideal case. In this paper, we do not consider noiseless coding: we evaluate our coding scheme in terms of distortion for a given (target) entropy. We do not consider either the quantization and coding of perceptual and segmentation data. We focus on the quantization of EDS parameters, which is the critical part in terms of entropy/distortion trade-off. A global description of the coding/decoding system that we propose is illustrated on figure 1.

This paper is organized as follows. In the first part, we present the signal model, the parametric estimation method and the analysis-synthesis framework. In the second part, we consider the optimal quantization of parameters under an entropy constraint, first in the single-sinusoid case, then in the multiple sinusoid-case with psychoacoustic considerations, which is the real situation in audio coding. In the last part, we evaluate our coding scheme in terms of perceived distortion, and compare it with a similar coding scheme using a constant-amplitude sinusoidal model, using state-of-the art methods for parametric estimation and quantization.

II. The analysis/synthesis framework

A. Signal model

The EDS model for a signal $x(t), t \in [0, T]$ can be written as

$$x(t) = \sum_{k=0}^{K-1} s_k(t) + \varepsilon(t), \tag{1}$$

where T is the length of the analysis/synthesis time-segment, K is the model order (i.e. the number of sinusoidal components), $s_k(t)$ is an exponentially damped sinusoid and $\varepsilon(t)$ is the residual signal. t can be a continuous-time variable (then T represents seconds) or a discrete-time index (then T represents a number of samples). Each EDS is defined as

$$s_k(t) = \alpha_k \, (z_k)^t. \tag{2}$$

α_k are the *complex amplitudes*, z_k the *poles*. Formulation (2) is mathematically convenient, but it is often clearer to use more explicit expressions. We write the poles as

$$z_k = e^{\frac{1}{T}(\delta_k + i\omega_k)}, \tag{3}$$

where δ_k are the *dampings* and ω_k the *pulsations*. Pulsation is directly related to the frequency of the oscillating part of each component (frequency is defined as $\omega_k / (2\pi)$, but we rather use pulsation for convenience sake), and damping is related to the sharpness of the exponential time-envelope. A positive damping corresponds to an increasing envelope, a negative damping to a decreasing envelope and a null damping to a constant-amplitude sinusoid. Note that we choose to normalize dampings and pulsations with respect to T. As we will use this model with variable-length time-segments, this ensures that the statistical distribution of dampings and pulsations is consistent over all segments. Furthermore, this expression will appear to be more convenient while optimizing the quantization.

We write the amplitudes as

$$\alpha_k = \begin{cases} a_k \, e^{-\delta_k + i\psi_k} & \text{if} \quad \delta_k \geq 0, \\ a_k \, e^{i\psi_k} & \text{if} \quad \delta_k < 0, \end{cases} \tag{4}$$

where a_k are the *real amplitudes* and ψ_k the *phases*. Practically, a_k corresponds to the maximum value of the exponential time-envelope, at $t = 0$ for a negative damping, and at $t = T$ for a positive damping. Using different expressions for α_k with positive and negative dampings avoids numerical errors while estimating a_k in the case of high (absolute) dampings.

It appears that quantizing directly ψ_k is not efficient. Indeed, in sinusoidal modeling, the phase origin is usually not at the beginning of the analysis segment: It was proved in [8] that, for constant-amplitude sinusoids and symmetric analysis windows, the energy of the quantization error is minimal when the phase origin is located in the middle of the segment. In our case, we show in section III-A1 that the optimal phase origin depends on damping, and must be adjusted independently for each component. For this purpose, we define a new phase parameter $\phi_k = \psi_k + \omega_k \tau_k$, where τ_k sets the location of the phase origin. The constant-amplitude case corresponds to $\tau_k = \frac{1}{2}$. The phase parameter to be quantized is ϕ_k. We show that τ_k only depends on the damping parameter δ_k, and thus does not need to be transmitted to the decoder.

In this study, we focus on the sinusoidal part, and do not consider the coding of the residual signal $\varepsilon(t)$.

B. Setting the model order

For a given signal and a given time-segment, finding the optimal value of parameter K, i.e. the number of sinusoids, is a challenging problem. Many methods have been proposed to automatically determine the best model order with respect to

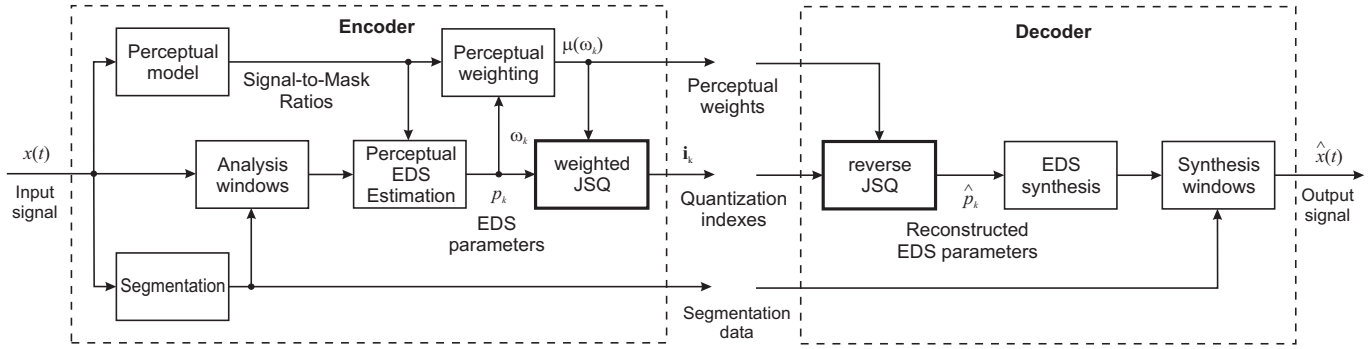

Fig. 1. Block-diagram of the coding/decoding scheme. JSQ means Joint-Scalar Quantization.

various criteria, e.g. Maximum Likelihood, Information Theoretic (ITC), Akaike Information (AIC), Maximum Description Length (MDL), Efficient Detection (EDC) or ESTimated ERror (ESTER). For a review, see [17]. However, all these methods rely on purely algebraic considerations and the results are not always perceptually optimal. In this paper, we do not directly address this problem. We consider a fixed mean-bitrate approach: we set the average entropy of quantization indexes for one component, i.e. the average amount of coding bits per component, which determines the model order.

C. Estimation of the parameters

1) Classical EDS model: Actually, the most efficient schemes for EDS parameter estimation are subspace methods [12], [14], [15], [18]. These methods are well known for their good spectral properties, and as explained in section IV-A, fast implementations have a lower complexity than fast Matching Pursuit. In this paper, we use the estimation scheme proposed in [18], which is an extension of the ESPRIT algorithm [19]. The basic algorithm can be briefly described as follows, using the discrete-time signal model.

We define the signal vector:

$$\boldsymbol{x} = \left[\begin{array}{cccc} x[0] & x[1] & \dots & x[T-1] \end{array} \right]^t, \tag{5}$$

where $(.)^t$ stands for the transposition operator. The Hankel signal matrix is defined as:

$$\boldsymbol{X} = \left[\begin{array}{cccc} x[0] & x[1] & \dots & x[Q-1] \\ x[1] & x[2] & \dots & x[Q] \\ \vdots & \vdots & & \vdots \\ x[R-1] & x[R] & \dots & x[T-1] \end{array} \right], \tag{6}$$

where $Q > K$, $R > K$, and $Q + R - 1 = T$. $Q \approx R$ was proved to be an efficient solution [20], thus we choose $Q = \lfloor T/2 \rfloor$ where $\lfloor . \rfloor$ stands for the lowest integer. We also define the complex-amplitude vector:

$$\boldsymbol{\alpha} = \left[\begin{array}{cccc} \alpha_0 & \alpha_1 & \dots & \alpha_{K-1} \end{array} \right]^t, \tag{7}$$

and the Vandermonde matrix of the poles:

$$\boldsymbol{Z}^T = \left[\begin{array}{cccc} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_{K-1} \\ \vdots & \vdots & \vdots & \vdots \\ z_0^{T-1} & z_1^{T-1} & \dots & z_{K-1}^{T-1} \end{array} \right]. \tag{8}$$

Thus, equation (1) can be written as:

$$\boldsymbol{x} = \boldsymbol{Z}^T \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{9}$$

Performing a singular value decomposition (SVD) on $\boldsymbol{X}$ leads to:

$$\boldsymbol{X} = [\boldsymbol{U_1 U_2}] \left[\begin{array}{cc} \boldsymbol{\Sigma_1} & 0 \\ 0 & \boldsymbol{\Sigma_2} \end{array} \right] \left[\begin{array}{c} \boldsymbol{V_1} \\ \boldsymbol{V_2} \end{array} \right], \tag{10}$$

where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal matrices respectively containing the K largest singular values, and the smallest singular values; $[\boldsymbol{U_1 U_2}]$ and $[\boldsymbol{V_1 V_2}]$ respectively are the corresponding left and right singular vector matrices. The shift-invariance property of the signal space spanned by $\boldsymbol{V}_1$ yields:

$$\boldsymbol{V}_1^{\downarrow} \boldsymbol{\Phi} = \boldsymbol{V}_1^{\uparrow}, \tag{11}$$

where the poles are the eigenvalues of matrix $\boldsymbol{\Phi}$. $(.)^{\uparrow}$ and $(.)^{\downarrow}$ respectively stand for the operators discarding the first line and the last line of a matrix. The estimation of $\boldsymbol{\Phi}$ that minimizes the mean square error is:

$$\boldsymbol{\Phi} = \left(\boldsymbol{V}_1^{\downarrow} \right)^{\dagger} \boldsymbol{V}_1^{\uparrow}, \tag{12}$$

where $(.)^{\dagger}$ denotes the pseudo-inverse operator. Then, z_k can be obtained by diagonalizing $\boldsymbol{\Phi}$. The associated Vandermonde matrix $\boldsymbol{Z}^T$ is computed. Finally, the optimal amplitudes with respect to the least square criterion are obtained by:

$$\boldsymbol{\alpha} = (\boldsymbol{Z}^T)^{\dagger} \boldsymbol{x}. \tag{13}$$

2) Perceptual EDS model: The previous estimation scheme is optimal according to the MSE criterion when the signal follows the EDS model and when the residual signal is a stationary white noise. However, real-life audio signals do not exactly follow the EDS model and the residual is not white, which degrades the performance of the estimation. Furthermore, this method minimizes the MSE, which is not a perceptual criterion. Some estimation methods use a whitening pre-filter in order to improve the performance of the estimation [21], and some other focus on the perceptual aspects: in [15], a subband-analysis approach is used. This method seems efficient, but is hardly compatible with audio coding, because the unavoidable frequency-domain overlap between subbands leads to multiple representation of some components, which degrades the coding efficiency. In [22], Chen *et al.* modified the Total Least Squares (TLS) algorithm, which is very similar to ESPRIT, in order to get approximately the result of the

original method applied to a pre-filtered version of the input signal, but without the drawback of deleting the first samples of the filtered signal corresponding to the length of filter impulse response. In [14], a modified version of this method is applied to audio signals with a perceptual pre-filter. In this study, we propose another modified version of this method. We define the perceptual pre-filtering matrix of size $R \times R$ as

$$\boldsymbol{M} = \begin{bmatrix} \mu[\tau] & \dots & \mu[0] & 0 & \dots & 0 \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ \mu[N-1] & & \ddots & & \ddots & 0 \\ 0 & \ddots & & \ddots & & \mu[0] \\ \vdots & \ddots & \ddots & & \ddots & \vdots \\ 0 & \dots & 0 & \mu[N-1] & \dots & \mu[\tau] \end{bmatrix}, \tag{14}$$

where $\mu[n], n \in \{0 \dots N-1\}$ is the impulse response of the perceptual filter, and $\tau = \frac{N+1}{2}$, N being an odd number. In the estimation algorithm, the SVD is actually performed on $\boldsymbol{MX}$ instead of $\boldsymbol{X}$ (equation (10)), which is similar to pre-filtering the audio signal x by μ. Due to the Toeplitz structure of matrix $\boldsymbol{M}$, the shift-invariance of the signal space is preserved, and the eigenvalues of $\boldsymbol{\Phi}$ are the most important poles from a perceptual point of view. With this method, psychoacoustics is taken into consideration in the selection of sinusoidal components, but the residual signal is still generally not white. However, this does not seem to significantly impact the precision of pole estimation. The estimation of complex amplitudes is still given by (13).

Different versions of the pre-filtering matrix have been proposed in the literature. In [14], a circular Toeplitz matrix is used, which is similar to performing a circular convolution on the input signal. This is known to introduce an edge-effect distortion, and thus does not seem suitable for audio coding. In contrast, we introduce a zero-padding in the pre-filtering matrix to avoid this edge-effect.

To compute the perceptual filter coefficient $\mu[n]$, we use an earing model. We appply the MPEG #2 psychoacoustic model (described in [2]) to the input signal. We get signal-to-mask ratios (SMR) over constant-length time segments (2048 samples, with 50% overlap). As explained in the following section, our analysis segments are aligned with onsets and thus are not necessarily aligned with the psychoacoustic model segments. A reasonable way to compute the SMR over each analysis segment is to average the SMR over psychoacoustic model segments that overlap this analysis segment.

In the literature, the perceptual filter in the frequency domain is usually identified to the inverse of the masking threshold. From our experiments, we found out that better results were obtained while directly using the SMR. Minimizing the MSE weighted by the inverse of the masking threshold is equivalent to minimizing the mean SMR. This makes sense since some studies have shown that the SMR is an accurate measure of perceptual distortion in a single perceptual subband. The mean SMR over all frequency bands is usually considered as a global measure of perceptual distortion, but it does not take the absolute energy of each spectral component into account. It is well known in audio coding engineering that a medium degradation on a high energy component will probably be much more annoying than a severe degradation on a low energy component, since the first one will more likely affect the global energy of the audio signal. This is especially true for low frequencies. Using the SMR as perceptual weight allows minimizing a criterion that takes both the relative energy (with respect to the masking threshold) and the absolute energy of each spectral component into account.

Finally, we compute the impulse response of the perceptual filter by applying an inverse Discrete Fourier Transform to the SMR.

D. Setting the analysis/synthesis segments and windows

The EDS model is efficient assuming that the boundaries of analysis/synthesis segments are aligned with onsets. Thus, we first perform an onset detection on the input signal. We use the method proposed by C. Duxbury *et al.* [23], which is based on a subband decomposition and an energy-variation criterion in each subband. This method determines the onset positions with a precision of 256 samples. For a more efficient detection, we moved *a posteriori* the onset location to the nearest local maximum of instantaneous energy. Once the onset locations are estimated, the boundaries of analysis/synthesis segments are computed. As a certain amount of overlap is unavoidable, we choose to locate the beginning of a segment 32 samples before the onset, and the end of the previous segment 32 samples after. A maximum segment length is set for implementation convenience to 2048 samples. This means that we consider two overlap categories: short overlaps (64 samples, i.e. 1.5 ms at 44100 samples per second) around onsets, and long overlaps (1024 samples, i.e. 23 ms at 44100 samples per second) in stationary regions.

Considering windows, the only choice for analysis is the rectangular window, because any other would degrade the estimation of dampings. At synthesis, a smooth window is necessary to avoid discontinuities, e.g. the square-sine (Hann) window. The onset locations and synthesis windows are illustrated on a portion of a glockenspiel signal in figure 2.

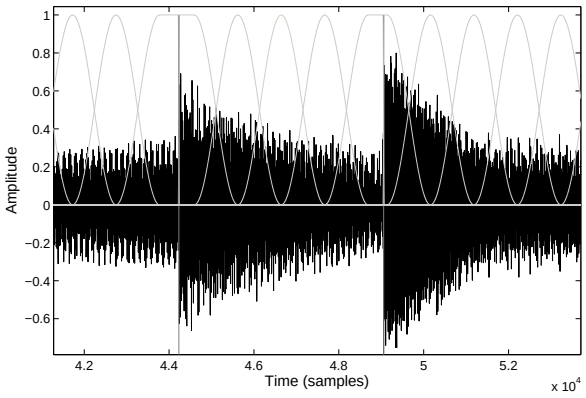

Fig. 2. Glockenspiel signal, onset locations and synthesis windows.

III. JOINT-SCALAR QUANTIZATION OF PARAMETERS

In this section, we consider the quantization of sinusoid parameters. For a single sinusoid $s_k(t)$, the set of parameters to be quantized is written $p_k = \{a_k, \delta_k, \omega_k, \phi_k\}$. Note that the phase parameter is obtained according to $\phi_k = \psi_k + \omega_k \tau_k$ where ψ_k is the original phase given by the estimation algorithm, and τ_k is the phase-origin parameter which will be set as a function of δ_k in section III-A1. After quantization and decoding, the reconstructed parameters are $\hat{p}_k = \{\hat{a}_k, \hat{\delta}_k, \hat{\omega}_k, \hat{\phi}_k\}$, and the reconstructed sinusoid is denoted $\hat{s}_k(t)$. The theoretically optimal solution would be to apply a 4D trained vector quantizer to p_k, but complexity quickly becomes a serious issue when the dimension increases. For amplitude-constant sinusoids, a more efficient joint-scalar quantization scheme was proposed by Korten *et al.* [8]. We propose below an extension of this approach to the case of exponentially damped sinusoids.

A. Single sinusoid case

1) Defining the quantizers: In a first step, we describe a quantization method that minimizes the mean quadratic distortion between the original and the decoded signals, under a constraint on the entropy of quantization indexes. With the continuous-time model, the distortion is defined as

$$d(p_k, \hat{p}_k) = \frac{1}{T} \int_0^T \left| s_k(t) - \hat{s}_k(t) \right|^2 dt. \tag{15}$$

In practice, $x(t)$ is a discrete-time signal, but the continuous-time expression is still a good approximation as long as $1/T$ is much smaller than the sampling frequency.

In this section, as we consider the quantization of a single sinusoid, we omit index k. In appendix A, we show that $d(p_k, \hat{p}_k)$ can be approximated under a high resolution hypothesis, i.e. when $\hat{p} \approx p$. We also show that the distortion can be minimized with respect to the phase origin parameter τ. We get

$$\begin{aligned} d(p, \hat{p}) \approx\ & a^2 h_1(2\delta) + \hat{a}^2 h_1(2\hat{\delta}) - 2a\hat{a}\, h_1(\delta + \hat{\delta}) \\ & + a\hat{a}(\omega - \hat{\omega})^2 h_2(\delta + \hat{\delta}) + a\hat{a}(\phi - \hat{\phi})^2 h_1(\delta + \hat{\delta}), \end{aligned} \tag{16}$$

where functions h_1 and h_2 are defined in the appendix (equation (29)). This approximation is valid when the phase origin parameter is set according to

$$\tau = \frac{(\delta + \hat{\delta}) - 1 + e^{-(\delta + \hat{\delta})}}{(\delta + \hat{\delta})\left(1 - e^{-(\delta + \hat{\delta})}\right)} \approx \frac{2\delta - 1 + e^{-2\delta}}{2\delta\left(1 - e^{-2\delta}\right)}. \tag{17}$$

It can be easily proved that $\tau \to 0$ when $\delta \to -\infty$, and $\tau \to 1$ when $\delta \to \infty$, which means that when the damping is highly positive (resp. negative), the phase origin is located in $t = T$ (resp. in $t = 0$), i.e. where the instantaneous energy is maximum. Furthermore, one can see that $\tau = \frac{1}{2}$ when $\delta = 0$, which is consistent with the case of a constant-amplitude sinusoidal model.

Optimization of entropy-constrained quantizers under a high resolution hypothesis was originally introduced by A. Gersho [24]. In this approach, the quantizers are defined by their quantization cell density (QCD), which can be seen as the inverse of the quantization step-size. In order to derive an analytic expression for the optimal QCD, we make a simplifying assumption: amplitude, phase, damping and pulsation are quantized with scalar quantizers, but depending on one another. This is called *joint-scalar quantization* (JSQ). Then, the global QCD can be written as the product of four scalar functions: g_A, g_Δ, g_Ω and g_Φ respectively for amplitude, damping, pulsation and phase. We note $i = \{i_a, i_\delta, i_\omega, i_\phi\}$ the set of quantization indexes associated with $\hat{p}$. We note P, $\hat{P}$ and I the random variables associated respectively with p, $\hat{p}$ and i. The optimal quantizers minimize the mean distortion $D = \mathbb{E}[d(P, \hat{P})]$ under the constraint $H(I) \leq \mathcal{H}$, where $H(I)$ denotes the entropy of quantization indexes and $\mathcal{H}$ the target entropy for one component.

In appendix A, we show that the mean distortion can be approximated as a function of scalar QCDs:

$$\begin{aligned} D \approx \tfrac{1}{12} \int \rho_P(p) \Big[& \tfrac{h_1(2\delta)}{g_A^2(p)} + \tfrac{a^2 h_1"(2\delta)}{g_\Delta^2(p)} \\ & + \tfrac{a^2 h_2(2\delta)}{g_\Omega^2(p)} + \tfrac{a^2 h_1(2\delta)}{g_\Phi^2(p)} \Big]\, dp. \end{aligned} \tag{18}$$

where $\rho_P(p)$ is the probability density function of EDS parameters and $h_1"$ stands for the second order derivative of function h_1.

The joint entropy of quantization indexes can be approximated by [24]

$$H(I) \approx h(P) + \int \rho_P(p) \log_2 \left[g_A(p) g_\Delta(p) g_\Omega(p) g_\Phi(p) \right] dp, \tag{19}$$

where $h(P)$ is the joint differential entropy of EDS parameters defined as

$$h(P) = -\int \rho_P(p)\ \log_2(\rho_P(p))\, dp. \tag{20}$$

Assuming that the entropy-distortion function of any quantizer (i.e. D as a function of $H(I)$) is decreasing, the optimal solution (i.e. the minimum value for D) is reached when $H(I) = \mathcal{H}$. This constrained optimization problem can be conveniently solved with a Lagrange optimization technique. In appendix B, we show that the optimal QCDs are

$$\begin{cases} g_A(\delta) \approx h_1(2\delta)^{\frac{1}{2}}\ 2^{\frac{1}{4}(\mathcal{H} - \mathcal{H}_0)} \\ g_\Phi(a, \delta) \approx a\, h_1(2\delta)^{\frac{1}{2}}\ 2^{\frac{1}{4}(\mathcal{H} - \mathcal{H}_0)} \\ g_\Delta(a, \delta) \approx a\, h_1"(2\delta)^{\frac{1}{2}}\ 2^{\frac{1}{4}(\mathcal{H} - \mathcal{H}_0)} \\ g_\Omega(a, \delta) \approx a\, h_2(2\delta)^{\frac{1}{2}}\ 2^{\frac{1}{4}(\mathcal{H} - \mathcal{H}_0)}, \end{cases} \tag{21}$$

where $\mathcal{H}_0$ is a constant that only depends on the probability density function of the parameters.

One can see that the amplitude, pulsation and phase quantizers are uniform (the QCD does not depend on the variable to be quantized), but the quantization step-size depends on damping for amplitude, and on both damping and amplitude for pulsation and phase. The damping quantizer is non-uniform: the QCD is maximal for small dampings and decreases as the absolute value of damping increases, i.e. the quantization cells are smaller for high dampings.

Combining equations (21) and equation (18) gives the theoretical entropy-distortion function, i.e. the mean distortion D as a function of the entropy of quantization indexes $\mathcal{H}$:

$$D \approx \frac{1}{3} 2^{\frac{1}{2}(\mathcal{H}_0 - \mathcal{H})}. \tag{22}$$

2) Implementation issues: The quantizers defined by equation (21) can be implemented with compression/expansion functions and a scalar uniform quantizer, the QCD being the slope of the compression function [25]. We get the following compression functions:

$$\begin{cases} f_A(a,\delta) \approx h_1(2\delta)^{\frac{1}{2}}\ 2^{\frac{1}{4}(\mathcal{H}-\mathcal{H}_0)}\ a \\ f_\Phi(a,\delta,\phi) \approx a\ h_1(2\delta)^{\frac{1}{2}}\ 2^{\frac{1}{4}(\mathcal{H}-\mathcal{H}_0)}\ \phi \\ f_\Delta(a,\delta) \approx a\ 2^{\frac{1}{4}(\mathcal{H}-\mathcal{H}_0)} \int_0^\delta h_1"(2u)^{\frac{1}{2}} du \\ f_\Omega(a,\delta,\omega) \approx a\ h_2(2\delta)^{\frac{1}{2}}\ 2^{\frac{1}{4}(\mathcal{H}-\mathcal{H}_0)}\ \omega \end{cases} \quad (23)$$

For amplitude and phase, the compression functions are linear with respect to the main variable. For damping, computing the compression function is not straightforward: The integral cannot be analytically computed, but it can be approximated with numerical integration techniques. So, we pre-compute and store a sampled version of the integral (i.e. for a finite set of values of δ). The compression function value for any δ is obtained by linear interpolation between sampled values. We use a pseudo-logarithmic sampling for δ (i.e. a logarithmic sampling, except around zero where a uniform sampling is used), which is much more efficient than a uniform sampling. For amplitude, phase and pulsation, we chose zero as the central reconstruction value and for damping, we found that the best results are obtained when zero is a boundary between two quantization cells. This implementation satisfies the conditions of symmetry around zero given in appendix A. For phase, the step-size of the quantizer is slightly modified in order to cover $[0, 2\pi]$ with an integer number of quantization cells. In the encoder, amplitude quantization and damping quantization are related, so they must be jointly performed. For quantizing phase and frequency, and for computing the phase origin τ, using the decoded values instead of the original ones minimizes the final MSE. This requires a local decoder inside the encoder, as illustrated on figure 3.

3) Real entropy-distortion function: First, we evaluate the performance of our quantization scheme on synthetic data. Like in [7] and [8], we assume that amplitude, phase and damping are statistically independent. In the literature, the amplitude is usually Rayleigh distributed and the phase is uniformly distributed over $[0, 2\pi]$. With the EDS model, we found out that the amplitude is more likely Gamma distributed ($p = 1$ and $\theta = 0.21$). For damping, only the distribution of $|\delta|$ is significant. Experiments showed that $\log(|\delta|)$ approximately follows a centered Gaussian distribution ($\sigma = 1.2$), and $\log(\omega)$ approximately follows a non-centered Gaussian distribution ($\mu = 5.5$, $\sigma = 1$).

We evaluated the entropy-distortion curve on $N = 10^8$ sets of parameters where amplitude, phase, damping and pulsation are independently generated using the distributions described above. The constant $\mathcal{H}_0$, defined in appendix B (equation (44)), cannot be computed analytically, but it can be numerically estimated. We obtained $\mathcal{H}_0 \approx -5.2$ bits. Both measured and theoretical entropy-distortion curves are plotted in figure 4. One can observe that theoretical and practical curves diverge in low resolution but converge under a high resolution hypothesis. We also compared our method with an entropy-constrained vector quantizer (VQ) as described by Chou *et al.* [26], for different sizes of the training database made of independent realizations of the same distributions. One can see that the VQ is always more efficient than the joint-scalar quantizers. However, as explained in [26], the minimum achievable distortion is directly related to the size of the database, and thus to the complexity. The largest tractable database on a high-performance workstation was 10^6 set of parameters, which corresponds to a minimum distortion of -51 dB. In terms of complexity, the joint scalar quantizer clearly outperforms the VQ, and there is no limit to the minimum achievable distortion. Thus, the JSQ is more suitable to high-resolution applications.

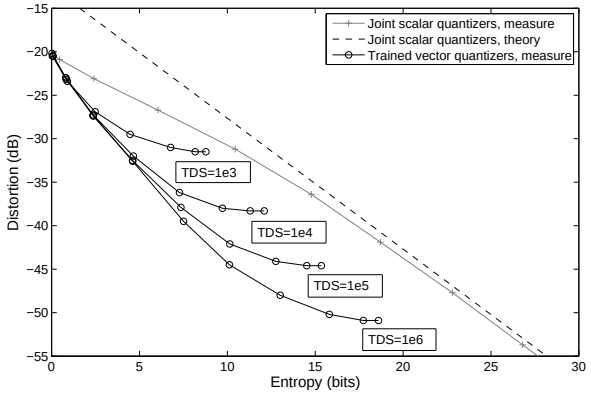

Fig. 4. Entropy-Distortion functions, evaluated on $N = 10^8$ set of parameters (synthetic data), of the real JSQ, theoretical JSQ and real trained VQ, for several training database sizes (TDS).

4) Distribution of entropy between parameters: We also considered the distribution of entropy between the quantization indexes associated to the 4 parameters. For different values of the target entropy, we computed the ratio $H(I_x|I_y, I_z, I_u)/H(I_x, I_y, I_z, I_u)$, x being amplitude, phase, pulsation or damping, and $\{y, z, u\}$ the other 3 parameters. The results are plotted on figure 5. One can notice that pulsation always requires the greatest part of the entropy (which is consistent with the results reported in [8]), and the damping always requires the lowest part, especially in low resolution. Asymptotically, it seems that all four parameters contribute equally, but this is out of the range of audio coding applications. Usually, sinusoidal coding requires between 15 and 20 bits per components. At these values, about 50% is devoted to pulsation, and about 12% is devoted to damping.

B. Multiple sinusoids case

In a second step, we consider the optimal quantization of the whole set of parameters $\mathbf{p} = \{p_0 \dots p_{K-1}\}$ for a given analysis time-segment. We assume that the random variables $\{P_0 \dots P_{K-1}\}$ are independent and identically distributed. We seek for the quantizers that minimize a perceptual distortion measure under an entropy constraint. We assume that we define a measure of perceptual significance (or perceptual weight) $\mu(\omega)$ as a function of pulsation. The inverse of the masking threshold over the current analysis/synthesis segment is usually chosen, but like in section II-C, we obtained better results with the SMR computed with MPEG psychoacoustic

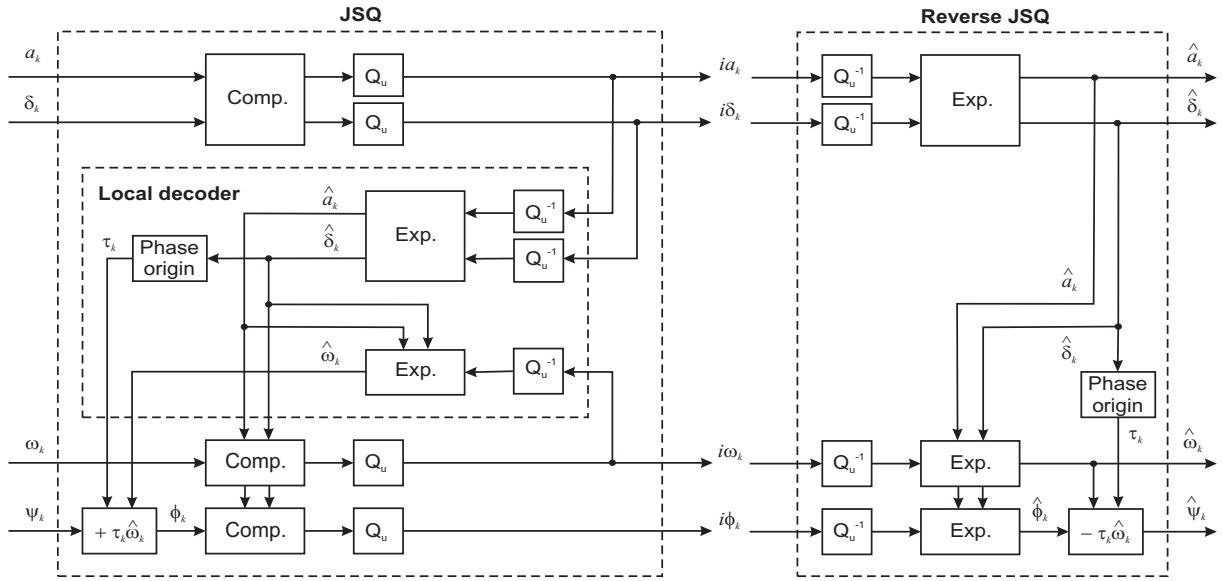

Fig. 3. Block-diagram of the quantization/reverse-quantization scheme. Comp/Exp mean compression and expansion functions, Q_u is a (single) uniform scalar quantizer depending on the perceptual weight $\mu(\omega_k)$ and on the constants $\mathcal{H}$, $\mathcal{H}_0$ and $\mathcal{H}_\mu$.

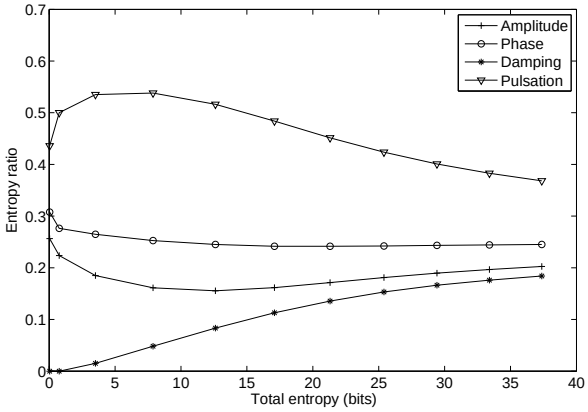

Fig. 5. Distribution of entropy between parameters in the 4-parameter JSQ, evaluated on $N = 10^8$ set of parameters (synthetic data).

model #2 [2]. In appendix C, we show that the mean distortion over the current segment can be approximated by

$$D \approx \frac{1}{12}\sum_{k=0}^{K-1}\int \rho_P(p_k)\,\mu(\omega_k)\left[\frac{h_1(2\delta_k)}{g^2_{A_k}(p_k)} + \frac{a_k^2 h_1"(2\delta_k)}{g^2_{\Delta_k}(p_k)} + \frac{a_k^2 h_2(2\delta_k)}{g^2_{\Omega_k}(p_k)} + \frac{a_k^2 h_1(2\delta_k)}{g^2_{\Phi_k}(p_k)}\right] dp_k. \tag{24}$$

This expression does not take the distortion coming from the interaction between two components into account, which is valid when the sinusoidal components in each segment are not very close in the spectral domain.

Assuming that $\mathbf{i} = \{i_0 \dots i_{K-1}\}$ stands for the whole set of quantization indexes in the current segment, and similarly to section III-A1, the entropy of quantization indexes can be written as

$$H(\mathbf{I}) \approx Kh(P) + \sum_{k=0}^{K-1}\int \rho_P(p_k)\log_2\left[g_{A_k}(p_k)g_{\Phi_k}(p_k)g_{\Delta_k}(p_k)g_{\Omega_k}(p_k)\right] dp_k. \tag{25}$$

Applying the same method as in section III-A1, we get the optimal QCDs:

$$\begin{cases} g_{A_k} \approx \mu(\omega_k)^{\frac{1}{2}}\; h_1(2\delta_k)^{\frac{1}{2}}\; 2^{\frac{1}{4}(\frac{\mathcal{H}^K}{K}-\mathcal{H}_0-\mathcal{H}_\mu)} \\ g_{\Phi_k} \approx a_k\,\mu(\omega_k)^{\frac{1}{2}}\; h_1(2\delta_k)^{\frac{1}{2}}\; 2^{\frac{1}{4}(\frac{\mathcal{H}^K}{K}-\mathcal{H}_0-\mathcal{H}_\mu)} \\ g_{\Delta_k} \approx a_k\,\mu(\omega_k)^{\frac{1}{2}}\; h_1"(2\delta_k)^{\frac{1}{2}}\; 2^{\frac{1}{4}(\frac{\mathcal{H}^K}{K}-\mathcal{H}_0-\mathcal{H}_\mu)} \\ g_{\Omega_k} \approx a_k\,\mu(\omega_k)^{\frac{1}{2}}\; h_2(2\delta_k)^{\frac{1}{2}}\; 2^{\frac{1}{4}(\frac{\mathcal{H}^K}{K}-\mathcal{H}_0-\mathcal{H}_\mu)}, \end{cases} \tag{26}$$

where $\mathcal{H}^K$ is the total entropy for the K components in the current time-segment and the constant $\mathcal{H}_\mu$ is defined in appendix C.

One can notice that we approximately get the same QCDs as in the single-sinusoid case, but each QCD is now multiplied by $\mu(\omega_k)^{\frac{1}{2}}$. In other words, the QCD is higher, i.e. the quantizer is more precise, when the component is perceptually more significant. Also, the target entropy appears as $\mathcal{H}^K/K$, which corresponds to the average entropy per component. This is consistent with the fact that the number of coding bits is allocated to the whole set of K components. The corresponding compression functions can be derived in the same way as in section III-A2.

IV. Evaluation of the coding system

In this section, we evaluate the performance of our coding scheme in terms of perceived quality vs. entropy on real audio data. As a reference, we choose a constant-amplitude sinusoidal model with the quantization scheme described in [8]. We globally use the same analysis/synthesis framework with dynamic segmentation aligned with onsets and the same psychoacoustic model for both methods. The two differences are: first, the windows. For the standard model, we follow

Id	Author	Identification	Style	Duration	$\bar{\mathcal{H}}$ sin model	$\bar{\mathcal{H}}$ EDS model	Bitrate ratio for damping	Attack sharpness	Attack density
1	P. Simon	Late in the Evening	Percussions	6 s	16 bits	23 bits	30%	+++	+++
2	S. Rollins	Sentimental Mood	Jazz saxophone	7 s	16 bits	22 bits	27%	+	+
3	J.J. Cale	The Problem	Electric Guitar	6 s	15 bits	21 bits	29%	++	++
4	S. Vega	Toms dinner	Singing voice	4 s	18 bits	24 bits	25%	+++	++
5	F. Tarrega	Alhambra	Classical guitar	6 s	15 bits	21 bits	29%	++	+++
6	G.P. Telemann	Fantasia	Flute	8 s	19 bits	27 bits	30%	+++	+
7	W.A. Mozart	Alla Turca	Piano	7 s	15 bits	21 bits	29%	+	+++
8	J.S. Bach	Suite	Cello	6 s	16 bits	22 bits	27%	+	++

TABLE I
AUDIO MATERIAL FOR LISTENING TEST. ATTACK SHARPNESS AND DENSITY GRADES HAVE BEEN OBTAINED THROUGH INFORMAL LISTENING TESTS AND ARE GIVEN ONLY FOR INFORMATION PURPOSES. + : LOW, ++ : MEDIUM, +++ : HIGH.

the same approach as in the literature: sinusoidal windows are used for analysis and synthesis. Second, for the standard model, the estimation algorithm is the perceptually weighted Matching Pursuit [27]. Note that the perceptual weights are identical for both algorithms.

A. Complexity

For each method, we evaluate the computational load required for estimating the parameter set over each analysis segment. We recall that T stands for the length of the analysis segment (in samples) and K stands for the model order.

First, we consider the case of the standard model. For a general Matching Pursuit (MP) algorithm, the complexity is proportional to the size of the dictionary (the set of all possible atoms) [28], and thus depends on the discretization of the parameters. Using the notations introduced in [28], we assume that frequency is discretized using $\alpha_f T$ levels. α_f is usually greater than 1 (which corresponds to the spectral resolution of the Fourier Transform), but does not exceed a few dozens. A fast implementation of the MP for extracting constant-amplitude sinusoids using the FFT algorithm [27] has a complexity in $O(\alpha_f K T \log(T))$. A refinement stage using a Newton algorithm [9] increases the precision of the frequency estimation, but also increases the complexity. We do not consider such a method here.

For the EDS model, some implementations using MP have been proposed [10]. Assuming that damping is discretized using $\alpha_d T$ levels, a fast implementation using the FFT algorithm has a complexity in $O(\alpha_f \alpha_d K T^2 \log(T))$. Again, a refinement stage using a Newton algorithm can be performed to increase the precision of frequency and damping estimation. In contrast, the complexity of subspace methods is not related to the precision of parameter estimation. A fast implementation of ESPRIT [29, Chapter V] has a complexity in $O(KT(K+\log(T)))$. When searching for a few sinusoidal components, i.e. K small, this is similar to the complexity of the standard model using fast MP, and much lower than the complexity of the EDS model using fast MP.

In our worst case, the length of the analysis segment is $T = 2048$ and the model order is $K = 40$. Assuming that $\alpha_f = \alpha_d = 1$, about 10^6 elementary operations are required for both the standard model with fast MP and the EDS model with fast ESPRIT, while the EDS model with fast MP requires about 10^9 elementary operations.

As a conclusion, subspace methods are much more efficient than MP for estimating the parameters of the EDS model. One can also notice that using the damped sinusoidal model does not increase the complexity of the analysis process compared to the standard model.

B. The evaluation process

With both coding methods, two parameters must be set: the model order K and the target entropy of quantization indexes per component $\bar{\mathcal{H}} = \mathcal{H}^K/K$ which controls the precision of the quantization process. The resulting bitrate and signal quality will depend on both parameters. So, maximizing the perceived quality for a given bitrate implies finding the optimal combination between these two parameters. We choose to tune the quantization stage so that the quantization error reaches a "perceptible but not annoying" level. To do so, we set a fixed average bitrate $\mathcal{R} = 20$ kbits/s, which seems to be a reasonable value for parametric audio coding applications. For both methods and for each audio excerpt, we perform the coding process for a discrete set of target entropy $\bar{\mathcal{H}}$, between 10 and 30 bits per component. Assuming that the noiseless coder is ideal i.e. the average number of coding bits per component equals the entropy, the model order is given by $K = \left[\frac{T\mathcal{R}}{\bar{\mathcal{H}} f_s}\right]$ where $[.]$ denotes the nearest integer and f_s the sampling frequency. Then, we apply the PEMO-Q algorithm [30] between the resynthesized unquantized and quantized signals. This method performs an objective evaluation of perceptive difference between two audio samples, and thus measures the distortion introduced by the quantization process. We *a posteriori* select the value of $\bar{\mathcal{H}}$ that corresponds to the desired quality.

For both methods, applying the joint-scalar quantizers implies the estimation of an offset parameter on the entropy, in our case $\mathcal{H}_0 + \mathcal{H}_\mu$. Estimating these parameters is not easy: $\mathcal{H}_0$ depends on $h(P)$, and it is well known that precisely estimating the entropy is a difficult problem [31], especially in multi-dimensional spaces, since a precise estimation sometimes requires a huge number of observations. Obviously, estimating the offset parameter with a reasonably good precision is not possible for each analysis segment individually. So we choose a global approach: we assume that model parameters follow a

probability density function common to all audio files, which defines a global offset parameter. Practically, we build a large audio database made of many short audio segments from various styles, and apply both coding methods with a medium model order corresponding to 20 bits per components and a unitary scale factor, i.e. in our case $2^{\frac{1}{4}(\bar{\mathcal{H}}-\mathcal{H}_0-\mathcal{H}_\mu)}=1$, which implies $\bar{\mathcal{H}}=\mathcal{H}_0+\mathcal{H}_\mu$. Then, we measure the entropy of quantization indexes over the whole database, which happens to be an estimation of the average entropy per component $\bar{\mathcal{H}}$, i.e. of $\mathcal{H}_0+\mathcal{H}_\mu$. We measured an entropy offset of 8.2 bits for the EDS model and of 10.4 bits for the standard sinusoidal model. From equation (26), one can see that a lower entropy offset $\mathcal{H}_0+\mathcal{H}_\mu$ means that a higher target entropy $\mathcal{H}$ is required to get the same scale factor. Thus, it appears that the EDS model requires in average 2.2 more bits per component than the standard model for the same quantization granularity. However, one can not draw conclusions about the final efficiency of both models now, because the scale factor is not necessarily proportional to the perceived distortion.

C. The audio material

We selected 8 short monophonic audio excerpts, with various attack sharpness and density profiles, described in table I. For each one, we give the average entropy of quantization indexes per component $\bar{\mathcal{H}}$ which corresponds to a "perceptible but not annoying" quantization error. We also provide informative grades for attack sharpness and attack density. These have been set through informal listening tests in order to clarify the interpretation of the listening test. The original and decoded audio files are available online at http://www.tsi.telecom-paristech.fr/aao/?p=863.

First, it appears that audio signals with very sharp transients (1, 4 and 6) require more coding bits, especially with the EDS model. This can be explained as follows: the coding quality on sharp transients is obviously crucial from a perceptual point of view, and thus audio material with sharp transients will require more binary information for a given quality. As the entropy offset is common to a large audio database, tracks with exceptionally sharp transients will require a higher target entropy for a given quality.

Then, with the EDS model, each component requires significantly more bits: in average, the classical sinusoidal model requires 16.3 bits per component and the EDS model 22.6 bits per component. Considering that the difference represents the amount of bits required for coding the damping in the EDS model, we obtain 6.3 bits for damping i.e. an average entropy ratio of 28%. Thus, with real audio data and psychoacoustic modeling, the part devoted to damping is close to 1/4, whereas on synthetic data and without a masking model, the results in figure 5 give an entropy ratio close to 1/6 for the same target entropy. Obviously, when psychoacoustics is taken into account, damping becomes much more significant.

Consequently, since the total bitrate is set to 20 kbits/s for both models, the average model order is 21% lower with the EDS model. In other words, there are less sinusoidal components in the synthesized signal with the EDS model than with the constant-amplitude model.

D. Subjective evaluation

Using the 8 audio excerpts described in table I, we organized subjective listening tests using the comparison category rating (CCR) protocol described by the ITU-T [32]. This test is considered more meaningful than a degradation category rating test at low bitrate [32]. For each audio excerpt, the original signal and two coded signals named A and B are presented. They correspond to both systems under test in random order. The listeners are asked to rate A compared with B on a five-step scale, from -3 (much worse) to +3 (much better). 0 means that both signals are of similar quality. The audio excerpts are presented in a random order, and the test has a symmetric structure, i.e. both A-B and B-A pairs are presented once. We asked 16 listeners to rate the coded signals after a training phase which introduced the typical degradations caused by audio coding. All listeners were professionals of audio signal processing and none of the authors were involved in the test. The mean opinion scores (MOS) and the 95% confidence intervals are presented in figure 6.

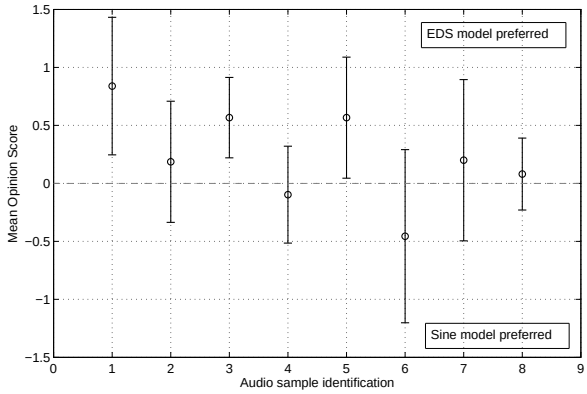

Fig. 6. Relative Mean Opinion Score and 95% confidence interval on audio excerpts described in table I.

The first conclusion is that the average differences between both coding methods are quite small: mostly between -0.5/+0.5 on a -3/+3 scale. This result was predictable since the main structure of both coding schemes is similar. Then, it appears that the EDS modeling was significantly preferred on 3 audio excerpts (1, 3 and 5), and for the other files, no significant difference can be pointed out (absolute MOS lower than 0.5). One can see that the sine model performs best on excerpt 6. No systematic relationship between attack sharpness/density and rating can be established. However, as expected, the EDS model is clearly preferred on audio signals which exhibit frequent and sharp onsets (++/+++). But on excerpt 4, which also exhibits frequent and sharp onsets, no significant improvement can be seen. On excerpt 6, which exhibits very sharp but rare attacks, the EDS model is obviously ineffective.

A general trend can be pointed out: on audio signals with frequent and sharp onsets, the EDS model is usually better, and otherwise, both models are similarly efficient, despite a lower number of sinusoids with the EDS model. However, two exceptions can be found on excerpts 4 and 6, where the EDS model is not as efficient as expected. This can be explained by

the presence of transients with complex amplitude envelopes (e.g. consonants in the singing voice, or plosive blows in a flute). In these situations, the model presumably does not efficiently match the signal.

V. Conclusion

In this paper, we present an alternative to classical sinusoidal modeling for low bitrate parametric audio coding, using exponentially damped sinusoids. We propose an efficient analysis/synthesis framework including a perceptual model and a new entropy constrained quantization method. When a fast algorithm is used, the analysis process has approximately the same complexity as a standard sinusoidal model with a fast Matching Pursuit algorithm, and thus will not lead to a higher calculation load than with a state-of-the-art codec. The design of the quantization stage is crucial in an audio coder, because it generates a significant part of the distortion and highly influences the rate-distortion efficiency of the global coding process. Our method is not as efficient as a vector quantizer, but it is much less complex to implement, and the achievable resolution can be as low as desired with the same complexity. We evaluate our method on real audio data at a target entropy corresponding to an average bitrate of 20 kbits/s. Compared with a similar coding scheme using a classical constant-amplitude sinusoidal model, our method is slightly but significantly better in terms of perceived audio quality. On most audio signals which exhibit frequent sharp onsets, our method is clearly preferred. However, on specific signals where the model does not seem to correctly match some transient parts, our coding scheme is not as efficient as expected and the classical scheme can perform better. Finally, our study shows that the exponentially damped sinusoidal model is an efficient tool for audio coding. We do not present a complete parametric audio codec, but propose some solutions for solving the main difficulties related to the implementation of an EDS-model based codec. However, this solution could be improved for instance by using a variable overlap at onsets, depending on the temporal energy profile, and by developing a specific hearing model for impulsive sounds.

Appendix A

In the single-sinusoid case, the MSE over the current analysis segment is defined by equation (15). The computation of the integral gives

$$\begin{aligned} d(p,\hat{p}) = {} & a^2\left(\tfrac{1-e^{-2|\delta|}}{2|\delta|}\right) + \hat{a}^2\left(\tfrac{1-e^{-2|\hat{\delta}|}}{2|\hat{\delta}|}\right) \\ & - \tfrac{2a\hat{a}(\delta+\hat{\delta})e^{-\frac{|\delta+\hat{\delta}|}{2}}}{(\delta+\hat{\delta})^2+(\omega-\hat{\omega})^2}\Big[e^{\frac{(\delta+\hat{\delta})}{2}}\cos[(1-\tau)(\omega-\hat{\omega})+(\phi-\hat{\phi})] \\ & \qquad - e^{\frac{-(\delta+\hat{\delta})}{2}}\cos[-\tau(\omega-\hat{\omega})+(\phi-\hat{\phi})]\Big] \\ & - \tfrac{2a\hat{a}(\delta+\hat{\delta})e^{-\frac{|\delta+\hat{\delta}|}{2}}}{(\delta+\hat{\delta})^2+(\omega-\hat{\omega})^2}\Big[e^{\frac{(\delta+\hat{\delta})}{2}}\sin[(1-\tau)(\omega-\hat{\omega})+(\phi-\hat{\phi})] \\ & \qquad - e^{\frac{-(\delta+\hat{\delta})}{2}}\sin[-\tau(\omega-\hat{\omega})+(\phi-\hat{\phi})]\Big]. \end{aligned} \tag{27}$$

This expression implies that δ and $\hat{\delta}$ have the same sign, which means that the damping quantizer is symmetric around 0. Because this equation is not tractable for further calculation, we develop it as a function of $(\omega-\hat{\omega})$ and $(\phi-\hat{\phi})$ in Taylor series around 0. We get an approximation valid for small error values, i.e. under a high resolution hypothesis:

$$\begin{aligned} d(p,\hat{p}) \approx {} & a^2h_1(2\delta) + \hat{a}^2h_1(2\hat{\delta}) - 2a\hat{a}h_1(\delta+\hat{\delta}) \\ & + a\hat{a}(\omega-\hat{\omega})^2h_2(\delta+\hat{\delta},\tau) + a\hat{a}(\phi-\hat{\phi})^2h_1(\delta+\hat{\delta}) \\ & + a\hat{a}(\omega-\hat{\omega})(\phi-\hat{\phi})h_3(\delta+\hat{\delta},\tau), \end{aligned} \tag{28}$$

where functions h_1, h_2 and h_3 are defined as:

$$\begin{aligned} h_1(x) &= \tfrac{1-e^{-|x|}}{|x|} \\ h_2(x,\tau) &= e^{\frac{-|x|}{2}}\left(\tfrac{e^{\frac{x}{2}}(x^2-2x+2)-2e^{\frac{-x}{2}}}{x^3}\right) \\ &\quad + e^{\frac{-|x|}{2}}\left(\tfrac{e^{\frac{x}{2}}(1-x)-1}{x^2}\right)\tau + \left(\tfrac{1-e^{-|x|}}{|x|}\right)\tau^2 \\ h_3(x,\tau) &= 2e^{\frac{-|x|}{2}}\left(\tfrac{e^{\frac{x}{2}}(x-1)+e^{\frac{-x}{2}}}{x^2}\right) - 2\left(\tfrac{1-e^{-|x|}}{|x|}\right)\tau. \end{aligned} \tag{29}$$

The optimal choice for τ is the value that minimizes expression (28). It will appear in further calculation that, because quantization cells are symmetrical, the mean distortion does not depend on odd powers of $(\omega-\hat{\omega})$ and $(\phi-\hat{\phi})$, and thus not on h_3. Assuming that a and $\hat{a}$ have the same sign, which means that the amplitude quantizer is symmetric around 0, the minimum distortion is obtained when $h_2(\delta+\hat{\delta},\tau)$ is a minimum as a function of τ. It is quite easy to prove that h_2 has a single minimum for the value of τ defined by equation (17), and thus the distortion can be written as in equation (16) with

$$h_2(x) = \frac{1}{|x|}\left(\frac{1-e^{-|x|}}{|x|^2} - \frac{e^{-|x|}}{1-e^{-|x|}}\right). \tag{30}$$

We now focus on the calculation of the mean distortion over all quantization cells. We note $\{\hat{p}_n\}, n \in \{0\dots N-1\}$ the reconstruction dictionary. $\mathcal{C}_n$ is the quantization cell associated to the reconstruction value $\hat{p}_n$. The mean distortion over $\mathcal{C}_n$ can be written as

$$d_{\mathcal{C}_n}(\hat{p}_n) = \frac{\int_{\mathcal{C}_n}\rho_P(p)d(p,\hat{p}_n)dp}{\int_{\mathcal{C}_n}\rho_P(p)dp}, \tag{31}$$

where $\rho_P(p)$ is the probability density function (PDF) of EDS parameters. The overall mean distortion is

$$D = \mathbb{E}[d(P,\hat{P})] = \sum_n \rho_n\, d_{\mathcal{C}_n}(\hat{p}_n), \tag{32}$$

where $\rho_n = \text{proba}\{P\in\mathcal{C}_n\} = \int_{\mathcal{C}_n}\rho_P(p)dp$.

We now focus on the calculation of $d_{\mathcal{C}_n}(\hat{p}_n)$. As we consider only one quantization cell, we temporarily omit index n. We assume that parameters are quantized with scalar quantizers. Thus, the 4D quantization cell can be seen as the product of 4 scalar quantization cells. Furthermore, it was shown in [25] that, in scalar entropy-constrained quantizers under a high resolution hypothesis, the reconstruction values are in the center of the cells. It is also reasonable to assume that

$\rho_P(p)$ is constant over $\mathcal{C}$. Thus, we get

$$\begin{aligned} d_{\mathcal{C}}(\hat{p}) \approx & \\ & \frac{1}{\Delta_a\Delta_\omega\Delta_\delta\Delta_\phi} \int_{\hat{a}-\frac{\Delta_a}{2}}^{\hat{a}+\frac{\Delta_a}{2}} \int_{\hat{\delta}-\frac{\Delta_\delta}{2}}^{\hat{\delta}+\frac{\Delta_\delta}{2}} \int_{\hat{\omega}-\frac{\Delta_\omega}{2}}^{\hat{\omega}+\frac{\Delta_\omega}{2}} \int_{\hat{\phi}-\frac{\Delta_\phi}{2}}^{\hat{\phi}+\frac{\Delta_\phi}{2}} a^2 h_1(2\delta) \\ & + \hat{a}^2 h_1(2\hat{\delta}) - 2a\hat{a}h_1(\delta+\hat{\delta}) + a\hat{a}(\omega-\hat{\omega})^2 h_2(\delta+\hat{\delta}) \\ & + a\hat{a}(\phi-\hat{\phi})^2 h_1(\delta+\hat{\delta}) \quad da\, d\delta\, d\omega\, d\phi, \end{aligned} \tag{33}$$

where Δ_a, Δ_δ, Δ_ω and Δ_ϕ denote the widths of scalar quantization cells respectively for amplitude, damping, pulsation and phase. The calculation of this integral gives

$$\begin{aligned} d_{\mathcal{C}}(\hat{p}) \approx & \hat{a}^2 h_1(2\hat{\delta}) \\ & + \left(\tfrac{\Delta_a^2}{12} + \hat{a}^2\right) \left[\tfrac{\tilde{h}_1(2\hat{\delta}+\Delta_\delta) - \tilde{h}_1(2\hat{\delta}-\Delta_\delta)}{2\Delta_\delta}\right] \\ & + \hat{a}^2 \left(\tfrac{\Delta_\phi^2}{12} - 2\right) \left[\tfrac{\tilde{h}_1\left(2\hat{\delta}+\frac{\Delta_\delta}{2}\right) - \tilde{h}_1\left(2\hat{\delta}-\frac{\Delta_\delta}{2}\right)}{\Delta_\delta}\right] \\ & + \hat{a}^2 \tfrac{\Delta_\omega^2}{12} \left[\tfrac{\tilde{h}_2\left(2\hat{\delta}+\frac{\Delta_\delta}{2}\right) - \tilde{h}_2\left(2\hat{\delta}-\frac{\Delta_\delta}{2}\right)}{\Delta_\delta}\right], \end{aligned} \tag{34}$$

where $\tilde{h}_1$ and $\tilde{h}_2$ are antiderivatives of h_1 and h_2.

The previous expression can be simplified by using Taylor series expansions in Δ_δ because under a high resolution hypothesis Δ_δ is supposed to be small. Keeping only terms in $O(\Delta_a^2)$, $O(\Delta_\delta^2)$, $O(\Delta_\omega^2)$ and $O(\Delta_\phi^2)$, we get

$$\begin{aligned} d_{\mathcal{C}}(\hat{p}) \approx \tfrac{1}{12} \Big[& h_1(2\hat{\delta})\Delta_a^2 + \hat{a}^2 h_1{}''(2\hat{\delta})\Delta_\delta^2 \\ & + \hat{a}^2 h_2(2\hat{\delta})\Delta_\omega^2 + \hat{a}^2 h_1(2\hat{\delta})\Delta_\phi^2 \Big], \end{aligned} \tag{35}$$

where $h_1{}''$ is the second order derivative of h_1.

At this step, one can verify the consistency with the results given in [8]: with a constant-amplitude sinusoidal model, the mean distortion over each quantization cell is given by

$$d_{\mathcal{C}} \approx \frac{1}{12}\left[\Delta_a^2 + \hat{a}^2\sigma\Delta_\nu^2 + \hat{a}^2\Delta_\phi^2\right], \tag{36}$$

with $\sigma = \frac{T^2}{12}$ for a rectangular analysis window. ν stands for the non-normalized pulsation. In our model, we use a normalized version of the pulsation, which means $\omega = T\nu$ and thus $\Delta_\omega^2 = T^2\Delta_\nu^2$. Furthermore, a constant-amplitude sinusoid corresponds to $\delta = 0$, and thus $\Delta_\delta = 0$, $h_1 = 1$ and $h_2 = \frac{1}{12}$. Finally, we get the same expression for $d_{\mathcal{C}}$.

As parameters are assumed to be quantized with scalar quantizers, the 4D QCD can be factorized with 4 scalar functions, whose values at the reconstruction points are:

$$g_A(\hat{p}) = \frac{1}{\Delta_a},\ g_\Delta(\hat{p}) = \frac{1}{\Delta_\delta},\ g_\Omega(\hat{p}) = \frac{1}{\Delta_\omega},\ g_\Phi(\hat{p}) = \frac{1}{\Delta_\phi}. \tag{37}$$

Thus, using the result of equation (35), the mean distortion defined by equation (32) can be written as

$$\begin{aligned} D \approx \tfrac{1}{12} \textstyle\sum_n \rho_n \Big[& \tfrac{h_1(2\hat{\delta}_n)}{g_A^2(\hat{p}_n)} + \tfrac{\hat{a}_n^2 h_1{}''(2\hat{\delta}_n)}{g_\Delta^2(\hat{p}_n)} \\ & + \tfrac{\hat{a}_n^2 h_2(2\hat{\delta}_n)}{g_\Omega^2(\hat{p}_n)} + \tfrac{\hat{a}_n^2 h_1(2\hat{\delta}_n)}{g_\Phi^2(\hat{p}_n)} \Big]. \end{aligned} \tag{38}$$

Assuming that $\rho_P(p)$ is constant over each quantization cell leads to the following approximation:

$$\rho_n \approx \rho_P\left(\hat{p}_n\right)\ \Delta_n, \tag{39}$$

where Δ_n is the volume of quantization cell $\mathcal{C}_n$, yielding

$$\begin{aligned} D \approx \tfrac{1}{12} \textstyle\sum_n \rho_P\left(\hat{p}_n\right) \Big[& \tfrac{h_1(2\hat{\delta}_n)}{g_A^2(\hat{p}_n)} + \tfrac{\hat{a}_n^2 h_1{}''(2\hat{\delta}_n)}{g_\Delta^2(\hat{p}_n)} \\ & + \tfrac{\hat{a}_n^2 h_2(2\hat{\delta}_n)}{g_\Omega^2(\hat{p}_n)} + \tfrac{\hat{a}_n^2 h_1(2\hat{\delta}_n)}{g_\Phi^2(\hat{p}_n)} \Big] \Delta_n. \end{aligned} \tag{40}$$

The sum can be approximated by an integral, and we get equation (18).

APPENDIX B

The set of quantizers which minimizes the mean distortion defined by equation (18) under the entropy constraint $H(I) \leq \mathcal{H}$ is approximately reached when the constraint is saturated. This optimization problem can be conveniently solved with a Lagrange optimization technique. The Lagrangian functional is defined as

$$\mathcal{L} = D + \lambda\left[H(I) - \mathcal{H}\right], \tag{41}$$

where λ is the real-valued Lagrange multiplier. The Euler-Lagrange equations give the optimal QCD expressions as functions of λ:

$$\begin{cases} \frac{\partial\mathcal{L}}{\partial g_A} = 0 \quad \Leftrightarrow \quad g_A(a,\delta) \approx \left(\frac{\ln(2) h_1(2\delta)}{6\lambda}\right)^{\frac{1}{2}} \\ \frac{\partial\mathcal{L}}{\partial g_\Delta} = 0 \quad \Leftrightarrow \quad g_\Delta(a,\delta) \approx a\left(\frac{\ln(2) h_1{}''(2\delta)}{6\lambda}\right)^{\frac{1}{2}} \\ \frac{\partial\mathcal{L}}{\partial g_\Omega} = 0 \quad \Leftrightarrow \quad g_\Omega(a,\delta,\omega) \approx a\left(\frac{\ln(2) h_2(2\delta)}{6\lambda}\right)^{\frac{1}{2}} \\ \frac{\partial\mathcal{L}}{\partial g_\Phi} = 0 \quad \Leftrightarrow \quad g_\Phi(a,\delta,\phi) \approx a\left(\frac{\ln(2) h_1(2\delta)}{6\lambda}\right)^{\frac{1}{2}}. \end{cases} \tag{42}$$

The optimal value for λ can be obtained from the constraint. Equation (19) can be rewritten as

$$\begin{aligned} \mathcal{H} \approx h(P) + \int & \rho_P(p) \\ & \times \log_2\left(\frac{a^3 \ln(2)^2\ h_1(2\delta)\ h_1{}''(2\delta)^{\frac{1}{2}}\ h_2(2\delta)^{\frac{1}{2}}}{(6\lambda)^2}\right) dp. \end{aligned} \tag{43}$$

Defining the following constant:

$$\begin{aligned} \mathcal{H}_0 = h(P) + \int \rho_\Delta(\delta) \log_2\left(h_1(2\delta)\ h_1{}''(2\delta)^{\frac{1}{2}}\ h_2(2\delta)^{\frac{1}{2}}\right) d\delta \\ + 3\int \rho_A(a)\log_2(a) da, \end{aligned} \tag{44}$$

where $\rho_A(a)$ and $\rho_\Delta(\delta)$ are respectively the marginal PDFs of amplitude and damping, the optimal value for λ is

$$\lambda \approx \frac{\ln(2)}{6} 2^{\frac{\mathcal{H}_0 - \mathcal{H}}{2}} \tag{45}$$

and the optimal QCDs are finally given by equation (21).

APPENDIX C

In the multiple-sinusoids case, we write the mean distortion as

$$D = \mathbb{E}[d(\mathbf{P}, \hat{\mathbf{P}})] = \sum_k \sum_n \mu(\hat{\omega}_{k,n})\ \rho_{k,n}\ d_{\mathcal{C}_{k,n}}(\hat{p}_{k,n}), \tag{46}$$

where $\mathcal{C}_{k,n}$ stands for the n-th quantization cell for component k, associated to the reconstruction value $\hat{p}_{k,n}$. We define $\rho_{k,n} = \text{proba}\{P \in \mathcal{C}_{k,n}\} = \int_{\mathcal{C}_{k,n}} \rho_P(p)\ dp$, and $\mu(\omega)$ is the perceptual weight depending on the pulsation. Like in [8], we assume that the total distortion is simply a linear combination of the distortions per component, and we do not take the cross-terms into account. This assumption is

valid when the components do not significantly overlap in the frequency domain, which is usually the case in parametric audio coding, where the signal is modeled by a relatively small number of sinusoids.

We apply the calculation developed in appendix A to $d_{\mathcal{C}_{k,n}}(\hat{p}_{k,n})$ and get the expression of the mean distortion:

$$D \approx \frac{1}{12}\sum_k \sum_n \mu(\hat{\omega}_{k,n})\rho_P\left(\hat{p}_{k,n}\right)\left[\frac{h_1(2\hat{\delta}_{k,n})}{g^2_{A_k}(\hat{p}_{k,n})} + \frac{\hat{a}^2_{k,n}h_1{}''(2\hat{\delta}_{k,n})}{g^2_{\Delta_k}(\hat{p}_{k,n})} + \frac{\hat{a}^2_{k,n}h_2(2\hat{\delta}_{k,n})}{g^2_{\Omega_k}(\hat{p}_{k,n})} + \frac{\hat{a}^2_{k,n}h_1(2\hat{\delta}_{k,n})}{g^2_{\Phi_k}(\hat{p}_{k,n})}\right]\Delta_{k,n}, \tag{47}$$

where $\Delta_{k,n}$ is the volume of quantization cell $\mathcal{C}_{k,n}$. The sum on n can be approximated by an integral, and we get equation (24).

Minimizing the distortion under the entropy constraint defined by equation (25) can be achieved in the same way as in the single-sinusoid case. We obtain the QCDs defined by equations (26) with

$$\mathcal{H}_\mu = 2\int \rho_\Omega(\omega)\log_2\left(\mu(\omega)\right)d\omega. \tag{48}$$

REFERENCES

[1] 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," ISO, Tech. Rep., 1993.

[2] ——, "13818-7: Generic Coding of Moving Pictures and Associated Audio: Advanced Audio Coding," ISO, Tech. Rep., 1997.

[3] R. McAulay and T. Quatiery, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[4] ISO/IEC, "14496-3 subpart 2: Information Technology - Very Low Bitrate Audio-Visual Coding, Part 3: Audio, Subpart 2: Parametric Coding," ISO, Tech. Rep., 1998.

[5] ——, "14496-3/AMD-2: Information Technology - Coding of Audio-Visual Objects, Amendment 2: Parametric Coding for High Quality Audio," ISO, Tech. Rep., 2004.

[6] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. Springer Verlag, 1990.

[7] R. Vafin and W. Kleijn, "Entropy-constrained polar quantization and its application to audio coding," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 220–232, Mar. 2005.

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

[9] J. Jensen and R. Heusdens, "A comparison of sinusoidal model variants for speech and audio representation," in *Proc. EUSIPCO'02*, Toulouse, France, Sept. 2002.

[10] M. Goodwin, "Matching pursuit with damped sinusoids," in *Proc. ICASSP'97*, Munich, Germany, Apr. 1997.

[11] S. VanHuffel, H. Park, and J. Rosen, "Formulation and solution of structured total least norm problems for parameter estimation," *IEEE Trans. Signal Process.*, vol. 44, no. 10, pp. 2464–2474, Oct. 1996.

[12] J. Nieuwenhuijse, R. Heusdens, and E. Deprettere, "Robust exponential modeling of audio signals," in *Proc. ICASSP'98*, Seattle, WA, USA, May 1998.

[13] O. Derrien, R. Badeau, and G. Richard, "Entropy-constrained quantization of exponentially damped sinusoids parameters," in *Proc. ICASSP'2011*, Prague, Czech Republic, May 2011.

[14] J. Jensen, R. Heusdens, and S. Jensen, "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 121–132, Mar. 2004.

[15] K. Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. van Huffel, "Perceptual audio modeling with exponentially damped sinusoids," *Signal Processing*, vol. 85, no. 1, pp. 163–176, Jan. 2005.

[16] M. Christensen, "Estimation and modeling problems in parametric audio coding," Ph.D. dissertation, Aalborg University, Denmark, 2005.

[17] R. Badeau, B. David, and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 492–504, Feb. 2006.

[18] R. Badeau, R. Boyer, and B. David, "EDS parametric modeling and tracking of audio signals," in *Proc. DAFx'02*, Hamburg, Germany, Sept. 2002.

[19] R. Roy, A. Paulraj, and T. Kailath, "Esprit - a subspace rotation approach to estimation of parameters of sinusoids in noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 34, no. 5, pp. 1340–1342, Oct. 1986.

[20] R. Badeau, G. Richard, and B. David, "Performance of ESPRIT for estimating mixtures of complex exponentials modulated by polynomials," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 492–504, Feb. 2008.

[21] B. David, V. Emiya, R. Badeau, and Y. Grenier, "Harmonic plus noise decomposition: Time-frequency reassignment versus a subspace-based method," in *Proc. 120th AES Convention*, Paris, France, May 2006.

[22] H. Chen, S. V. Huffel, and J. Vanderwalle, "Bandpass prefiltering for exponential data fitting with known frequency regions of interest," *Signal Processing*, vol. 48, pp. 135–154, 1995.

[23] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proc. DAFx'02*, Hamburg, Germany, Sept. 2002.

[24] A. Gersho, "Asymptotically optimal block quantization," *IEEE Trans. Inf. Theory*, vol. 25, no. 4, pp. 373–380, July 1979.

[25] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*. Kluwer Academic, 1992.

[26] P. Chou, T. Lookbaugh, and R. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 1, pp. 31–42, Jan. 1989.

[27] T. Verma and T. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," in *Proc. ICASSP'99*, Phoenix, AZ, USA, Mar. 1999.

[28] B. Mailhe, R. Gribonval, F. Bimbot, and P. Vandergheynst, "A low complexity orthogonal matching pursuit for sparse signal approximation with shift-invariant dictionaries," in *Proc. ICASSP'2009*, Tiapei, Taiwan, April 2009.

[29] R. Badeau, "High resolution methods for estimating and tracking modulated sinusoids. Application to music signals." Ph.D. dissertation, École Nationale Supérieure des Télécommunications, ENST2005E007, Paris, France, Apr. 2005, in French.

[30] R. Huber and B. Kollmeier, "PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[31] T. Schürmann, "Bias analysis in entropy estimation," *J. Phys. A: Math. Gen*, vol. 37, pp. 295–301, 2004.

[32] ITU-T, "Recommendation p.800: Methods for subjective determination of transmission quality," ITU, Tech. Rep., 1996.

Olivier Derrien (M'10) was born in Aix-en-Provence, France, in 1974. He recieved the State Engineering degree in 1998 and the Ph.D. degree in audio processing in 2002, both from Télécom ParisTech (formerly ENST), Paris, France.

From 2002 to 2003, he was a Teaching and Research assistant at the University of Paris-XI, Orsay, France. He joined the University of Toulon, Toulon, France, in 2003 as an Associate Professor in the field of telecommunications. In 2008, he joined the Laboratory of Mechanics and Acoustics (LMA), Marseille, France, as an Associate Researcher. His research interests include audio signal processing, especially audio coding, audio synthesis for virtual reality, music synthesis and audio effects. He is a co-author of 3 journal papers and 14 international conference papers.

Roland Badeau (M'02-SM'10) was born in Marseille, France, in 1976. He received the State Engineering degree from the École Polytechnique, Palaiseau, France, in 1999, the State Engineering degree from Télécom ParisTech (formerly ENST), Paris, France, in 2001, the M.Sc. degree in applied mathematics from the École Normale Supérieure (ENS), Cachan, France, in 2001, and the Ph.D. degree from Telecom ParisTech in 2005, in the field of signal processing. He received the ParisTech Ph.D. Award in 2006, and the Habilitation degree from the Université Pierre et Marie Curie (UPMC), Paris VI, in 2010.

In 2001, he joined the Department of Signal and Image Processing of Télécom ParisTech, CNRS LTCI, as an Assistant Professor, where he became Associate Professor in 2005. From November 2006 to February 2010, he was the manager of the DESAM project, funded by the French National Research Agency (ANR), whose consortium was composed of four academic partners. His research interests focus on statistical modeling of non-stationary signals (including adaptive high resolution spectral analysis and Bayesian extensions to NMF), with applications to audio and music (source separation, multipitch estimation, automatic music transcription, audio coding, audio inpainting). He is a co-author of 19 journal papers, 50 international conference papers, and 2 patents. He teaches in the Master of Engineering of Télécom ParisTech and in the Master of Sciences and Technologies of UPMC. He is also a Chief Engineer of the French Corps of Mines (foremost of the great technical corps of the French state) and an Associate Editor of the EURASIP Journal on Audio, Speech, and Music Processing.

Gaël Richard (SM'06) received the State Engineering degree from Télécom ParisTech, France (formerly ENST) in 1990, the Ph.D. degree from LIMSI-CNRS, University of Paris-XI, in 1994 in speech synthesis, and the Habilitation Diriger des Recherches degree from the University of Paris XI in September 2001. After the Ph.D. degree, 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.

From 1997 to 2001, he successively worked for Matra, Bois d'Arcy, France, and for Philips, Montrouge, France. In particular, he was the Project Manager of several large scale European projects in the field of audio and multimodal signal processing. In September 2001, he joined the Department of Signal and Image Processing, Télécom ParisTech, where he is now a Full Professor in audio signal processing and Head of the Audio, Acoustics, and Waves research group. He is a coauthor of over 120 papers and inventor in a number of patents and is also one of the experts of the European commission in the field of audio signal processing and man/machine interfaces. He was an Associate Editor of the IEEE Transactions on Audio, Speech and Language Processing between 1997 and 2011 and one of the guest editors of the special issue on "Music Signal Processing" of IEEE Journal on Selected Topics in Signal Processing (2011). He currently is a member of the IEEE Audio and Acoustic Signal Processing Technical Committee, member of the EURASIP and AES and senior member of the IEEE.