Low bitrate object coding of musical audio using bayesian harmonic models

Emmanuel Vincent, Mark Plumbley

Low Bitrate Object Coding of Musical Audio Using Bayesian Harmonic Models

Emmanuel Vincent and Mark D. Plumbley

Abstract— **This article deals with the decomposition of music signals into pitched sound objects made of harmonic sinusoidal partials for very low bitrate coding purposes. After a brief review of existing methods, we recast this problem in the Bayesian framework. We propose a family of probabilistic signal models combining learnt object priors and various perceptually motivated distortion measures. We design efficient algorithms to infer object parameters and build a coder based on the interpolation of frequency and amplitude parameters. Listening tests suggest that the loudness-based distortion measure outperforms other distortion measures and that our coder results in a better sound quality than baseline transform and parametric coders at 8 kbit/s and 2 kbit/s. This work constitutes a new step towards a fully object-based coding system, which would represent audio signals as collections of meaningful note-like sound objects.**

Index Terms— **Object coding, harmonic sinusoidal model, perceptual distortion measure, Bayesian inference.**

I. Introduction

PERCEPTUAL coding aims to reduce the bitrate required to encode an audio signal while minimizing the perceptual distortion between the original and encoded versions. For musical audio, much of the effort to date has concentrated on generic transform coders which encode the coefficients of an adaptive time-frequency representation of the signal. Transform coders such as MPEG-4 AAC (Advanced Audio Coder) [1] typically provide transparent quality at around 64 kbit/s for mono signals but generate artifacts at lower bitrates.

Parametric coders attempt to address this issue by representing the signal as a collection of sinusoidal, transient and noise elements, whose characteristics are more adapted to musical audio. For example, sinusoidal elements are formed by locating sinusoids within short time frames using spectral peak picking or Matching Pursuit [2] and tracking them across frames. Amplitude and frequency parameters are then differentially encoded for each track, while phase may be transmitted or not, depending on the coder. The MPEG-4 SSC (SinuSoidal Coding) parametric coder [3], based on this approach, results in a better quality than AAC at 24 kbit/s. However it is not suited for much lower bitrates.

Object coding is an extension of the notion of parametric coding where the signal is decomposed into meaningful sound objects such as notes, chords and instruments, described using high-level attributes [4]. As well as offering the potential for very low bitrate compression, this coding scheme leads to many other potential applications, including browsing by content, source separation and interactive signal manipulation.

Several authors have proposed to address object coding based on the fact that musical notes contain sinusoidal partials at harmonic frequencies. The MPEG-4 HILN (Harmonic and Individual Lines plus Noise) coder defines *pitched objects* made of harmonic sinusoidal tracks and extracts one predominant object per frame [5], whereas other methods extract several objects per frame [6], [7]. In order to reduce the bitrate needed to represent each object, while preserving its perceptually important properties, the frequency and amplitude parameters of the tracks are jointly encoded using a single fundamental frequency track and a few spectral envelope coefficients. In practice, however, the various algorithms proposed to estimate pitched objects do not succeed in extracting all the sinusoidal partials present in the signal and the remaining partials must be encoded as standalone sinusoidal tracks. Therefore none of these methods is fully "object-based" and this results in a limited compression gain. For instance, at 6 kbit/s, HILN performs only slightly better than a simple parametric coder [5], but not as well as TwinVQ [8].

In this article, we propose a Bayesian approach to decompose music signals into pitched sound objects for very low bitrate object coding purposes. We do not focus on accurately estimating the fundamental frequencies of the notes being played, but rather on using a perceptually motivated analysis-by-synthesis procedure that guarantees a good resynthesis quality without needing complementary standalone sinusoidal tracks. The strength of the proposed approach is the exploitation of both simple psycho-acoustics and learnt parameter priors. We extend our preliminary work [9] in several ways: we investigate other perceptually motivated distortion measures, we design an improved Bayesian marginalization algorithm, we propose a new interpolation and quantization scheme to obtain a specified bitrate, and we provide a rigorous evaluation of our approach by means of listening tests.

The structure of the rest of the article is as follows. In Section II, we discuss in more detail some existing methods for the extraction of pitched objects and reformulate this problem in the Bayesian framework. We define a family of probabilistic signal models involving pitched objects in Section III and describe the associated perceptual distortion measures in Section IV. Then we design an efficient algorithm to infer the object parameters in Section V and derive a very low bitrate coder in Section VI. We select the best distortion measure and evaluate the performance of this coder from listening tests presented in Section VII. We conclude in Section VIII and suggest further research directions.

Manuscript received April 4, 2006; revised August 11, 2006. The associate editor coordinating the review of this manuscript was Dr. George Tzanetakis.

The authors are with the Center for Digital Music, Department of Electronic Engineering, Queen Mary, University of London, London E1 4NS, U.K. (e-mail: emmanuel.vincent@elec.qmul.ac.uk; mark.plumbley@elec.qmul.ac.uk).

II. Methods for the estimation of pitched objects

Object coding can be performed in two steps: first estimate the parameters of the sound objects underlying the signal, then jointly encode these parameters. In the case of pitched objects, the first step amounts to estimating the time-varying fundamental frequency of each object and the time-varying amplitudes and phases of its harmonic partials. Several approaches have been proposed so far to perform this estimation.

A. Sinusoidal track extraction and grouping

A fast approach employed in [6] is to extract sinusoidal tracks [10] and group simultaneous tracks into pitched objects using auditory motivated principles such as proximity of onset times, harmonicity and correlation of frequency modulations [11]. This method shows several drawbacks in an object coding context. First, the quality is often poor due to tracking errors perceived as artifacts, such as spurious sinusoidal tracks not corresponding to actual note partials, upper note partials being transcribed as several tracks separated by a gap, or partials from different notes being joined into a single track. These errors are particularly frequent for music signals, since partials from different notes often overlap or cross each other in the time-frequency plane and partials in the upper frequency range tend to be masked by background noise due to their small amplitude, as illustrated in Figure 1. Moreover, the compression gain is usually limited due to grouping errors resulting in some notes being represented by several objects with redundant information instead of a single object [6].

B. Pitch tracking and estimation of harmonic partials

A more principled approach to obtain pitched objects is to estimate the fundamental frequency tracks underlying the signal and compute the amplitudes and phases of their harmonics. This approach is known to help reducing the above tracking and grouping errors since all the partials of a given note are tracked jointly [12]. However the estimation of several concurrent fundamental frequencies is a difficult problem for which no current algorithm provides a perfect solution. The method used in [7] determines the fundamental frequencies based on the summary autocorrelation of the signal [13] and estimates the parameters of the partials separately for each object on each frame. This method is fast, but it may produce spurious or erroneous fundamental frequencies and temporal discontinuities, which result either in a poor rendering of the signal or in an increase of the number of parameters to encode. Harmonic Matching Pursuit [14] ensures a better resynthesis quality due to its analysis-by-synthesis approach, but often generates fundamental frequency errors since it does not use any information about the amplitudes of the partials.

C. Proposed Bayesian approach

This brief review shows that the estimated pitched objects must satisfy two requirements for a coding application: firstly they must minimize the perceptual distortion between the observed and the resynthesized signals, and secondly they must exhibit the same parameter values as typical musical notes

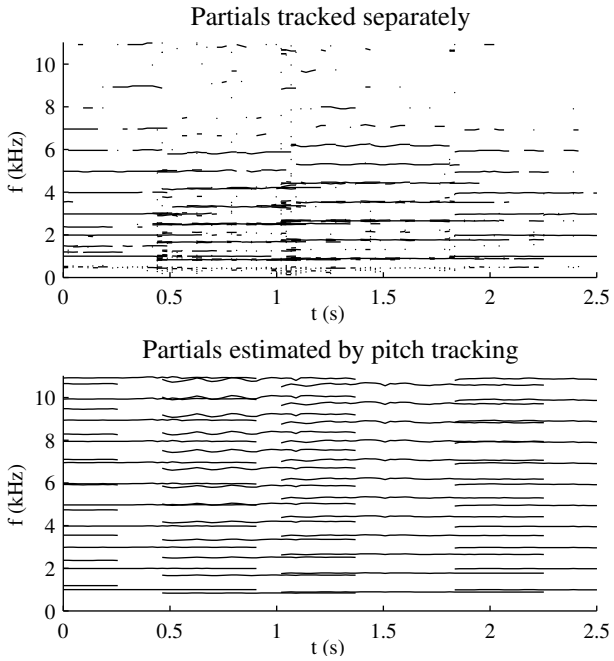


Fig. 1. Comparison of two approaches for the estimation of pitched objects on a solo flute signal.

to avoid spurious or erroneous objects which would result in an increased bitrate. Bayesian estimation theory is a natural framework to solve this problem. It consists in modeling prior belief about the object parameters and the non-pitched residual using probabilistic priors, and in estimating the parameters using a Maximum *A Posteriori* (MAP) probability criterion.

Two families of Bayesian harmonic models have been proposed previously. The models presented in [15], [16] describe each object by its fundamental frequency, amplitude and phase parameters on each time frame. The number of objects and the number of partials per object are allowed to vary across frames [15] or assumed to be fixed over the whole signal [16] and follow exponential [15] or Poisson priors [16]. Amplitudes are modeled by independent uniform [15] or zero-mean Gaussian priors [16] and fundamental frequencies by independent log-Gaussian [15] or uniform [16] priors. The residual follows a Gaussian prior. Parameter inference relies on Markov Chain Monte Carlo (MCMC) algorithms. Another model introduced in [17] models each object in state space form by a fixed number of oscillators with fixed frequencies and damping factors and a Gaussian residual. The initial amplitudes of the oscillators follow a zero-mean Gaussian prior. Object onset and offset times are modeled by a factorial Markov prior. Decoding is achieved by Kalman filtering and beam search.

Both families of models have provided promising results for the estimation of the musical score. However, they suffer some limitations for coding. The models in [15], [16] are not constrained enough to ensure a good resynthesis quality. Firstly, the lack of temporal continuity priors over the parameters and the possible variation of the number of partials per object may produce temporal discontinuities perceived as artifacts. Secondly, the priors over the number of partials favor a small

number of estimated partials independently of the fundamental frequency, which induces a low-pass filtering distortion on low frequency notes containing a large number of partials. Thirdly, the Gaussian prior over the residual corresponds to a power distortion measure which results in low power components such as high frequency partials, onsets and reverberation not being transcribed despite their perceptual significance. Finally, the priors over the amplitudes of the partials do not penalize partials with zero amplitude and can lead to fundamental frequency errors [16]. The model in [17] exhibits similar limitations and appears too constrained to allow perfect resynthesis of realistic musical notes. Moreover, both families of models rely on computationally intensive inference algorithms.

In the following, we seek to address these limitations by defining a new family of Bayesian harmonic models involving learnt priors for amplitude and fundamental frequency parameters and various perceptually motivated priors for the residual. We also design a faster estimation algorithm based on a new Bayesian marginalization technique.

III. Bayesian harmonic models

A. Structure of the proposed family of models

The proposed models exhibit a four-layer dynamic Bayesian network structure shown in Figure 2. The observed signal $x(t)$ is split into several time frames $(x_n(t))_{1\leq n\leq N}$ defined by $x_n(t) = w(t)x(n\tilde{W}+t)$, where $w(t)$ is a window of length W and $\tilde{W}$ is the stepsize. Each layer represents these signal frames at a different abstraction level.

The bottom layer provides a so-called *piano roll* representation of the signal consisting in a sequence of discrete vector states $(S_n)_{1\leq n\leq N}$. In western music, the normalized fundamental frequency f_{pn} of each note generally varies over time but remains close to a discrete pitch of the form

$$\mu_p^f = \frac{440}{F_s} 2^{\frac{p-69}{12}}, \tag{1}$$

where F_s is the sampling frequency in Hz and p an integer value on the MIDI semitone scale, with $p = 69$ corresponding to 440 Hz. Assuming no *unison*, *i.e.* several pitched objects corresponding to the same discrete pitch cannot be present at the same time, each point p on the MIDI scale is simply associated with a binary activity state $S_{pn} \in \{0, 1\}$ determining whether a pitched object corresponding to that discrete pitch is present in frame n or not. The global state and the set of active discrete pitches in frame n are denoted respectively $S_n = (S_{pn})_{p_{\rm low}\leq p\leq p_{\rm high}}$ and $\mathcal{A}_n = \{p \text{ s.t. } S_{pn} = 1\}$.

This piano roll representation can be expressed equivalently as an object-based representation: a subsequence of activity states such that $S_{p,n_{\rm on}-1} = 0$, $S_{p,n_{\rm off}+1} = 0$ and $S_{pn} = 1$ for all $n_{\rm on} \leq n \leq n_{\rm off}$ then corresponds to a pitched object with onset time $n_{\rm on}$, offset time $n_{\rm off}$ and discrete pitch p.

The signal corresponding to each pitched object is defined in the middle layers by

$$s_{pn}(t) = w(t) \sum_{m=1}^{M_p} a_{pmn} \cos(2\pi m f_{pn} t + \phi_{pmn}), \tag{2}$$

where f_{pn} is its normalized fundamental frequency in frame n and (a_{pmn}, ϕ_{pmn}) are the amplitude and the phase of its m-th partial in that frame. We emphasize that f_{pn} may be different from μ_p^f and vary over time. The partials amplitudes are also related to a global amplitude factor r_{pn} defined later in Section III-C. The number of partials per object M_p is constrained to

$$M_p = \min\left(\frac{1}{2\mu_p^f}, M_{\max}\right), \tag{3}$$

so that the partials fill the whole observed frequency range up to a maximal number of partials $M_{\max}$. Finally, the observed signal is modeled in the top layer as

$$x_n(t) = \sum_{p\in\mathcal{A}_n} s_{pn}(t) + e_n(t), \tag{4}$$

where $e_n(t)$ is the residual.

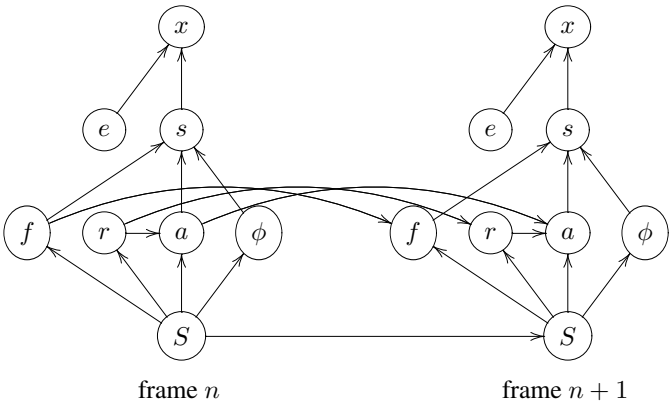


Fig. 2. Graphical representation of the proposed models. Circles represent vector random variables (some of variable size) and arrows denote conditional dependencies. The variables denote the following quantities: x observed signal, s pitched objects, e residual, f fundamental frequencies, r global amplitude factors, a amplitudes of the partials, ϕ phases of the partials, S discrete object states. Subscripts are omitted for legibility.

B. State prior

In the context of coding, it is of interest to represent the signal using as few meaningful objects as possible. Thus the prior over the activity states $(S_n)_{1\leq n\leq N}$ must favor inactivity and avoid short duration objects or short silence gaps within notes. We model this using a product of binary Markov priors

$$P((S_n)_{1\leq n\leq N}) = \prod_{p=p_{\rm low}}^{p_{\rm high}} P(S_{p,1}) \prod_{n=2}^{N} P(S_{p,n+1}|S_{pn}). \tag{5}$$

By definition, the values of the transition probabilities $P_{10} = P(S_{p,n+1} = 1|S_{pn} = 0)$ and $P_{01} = P(S_{p,n+1} = 0|S_{pn} = 1)$ are related to the mean duration of activity and inactivity segments for each discrete pitch. Simple computation shows that the mean inactivity probability equals $P_Z = (1+P_{10}/P_{01})^{-1}$.

C. Parameter priors

Given the activity states $(S_n)_{1\leq n\leq N}$, the parameters of different objects are assumed to be independent. To avoid fundamental frequency errors, the parameter priors are based on observation of the empirical distribution of musical note

parameters [18]. We model the normalized fundamental frequency of each object by a product of log-Gaussian priors

$$P((\log f_{pn})_{n_{\rm on}\le n\le n_{\rm off}}) \propto \prod_{n=n_{\rm on}}^{n_{\rm off}} \mathcal{N}(\log f_{pn}; \log \mu_p^f, \sigma^f) \prod_{n=n_{\rm on}+1}^{n_{\rm off}} \mathcal{N}(\log f_{pn} - \log f_{p,n-1}; 0, \tilde{\sigma}^f), \quad (6)$$

where $\mathcal{N}(\cdot; \mu, \sigma)$ is the univariate Gaussian density of mean μ and standard deviation σ. This enforces both proximity to the underlying discrete pitch and temporal continuity. Similarly, we represent the amplitudes of the partials as

$$P((\log a_{pmn})_{n_{\rm on}\le n\le n_{\rm off}}|(r_{pn})_{n_{\rm on}\le n\le n_{\rm off}}) \propto \prod_{n=n_{\rm on}}^{n_{\rm off}} \mathcal{N}(\log a_{pmn}; \log(r_{pn}\mu_{pm}^a), \sigma_p^a) \prod_{n=n_{\rm on}+1}^{n_{\rm off}} \mathcal{N}(\log a_{pmn} - \log a_{pm,n-1}; 0, \tilde{\sigma}_{pm}^a), \quad (7)$$

where $(\mu_{pm}^a)_{1\le m\le M_p}$ is a fixed normalized spectral envelope and r_{pn} a global amplitude factor for this object. This helps to avoid partials with zero amplitude or temporal discontinuities. The global amplitude factor is in turn modeled by

$$P((\log r_{pn})_{n_{\rm on}\le n\le n_{\rm off}}) \propto \prod_{n=n_{\rm on}}^{n_{\rm off}} \mathcal{N}(\log r_{pn}; \log \mu_p^r, \sigma_p^r) \prod_{n=n_{\rm on}+1}^{n_{\rm off}} \mathcal{N}(\log r_{pn} - \log r_{p,n-1}; 0, \tilde{\sigma}_p^r). \quad (8)$$

Finally, we assume that the phases of the partials are independent and uniformly distributed

$$P(\phi_{pmn}) = 1/2\pi. \quad (9)$$

D. Model learning

An accurate way to learn the model hyper-parameters is to use a large database of isolated notes, whose parameters can be easily transcribed without errors and span the whole variation range of several instruments. In the following, we use a subset of the RWC Musical Instrument Database[1] to learn σ^f, $\tilde{\sigma}^f$, $(\mu_{pm}^a)_{1\le m\le M_p}$, σ_p^a, $(\tilde{\sigma}_{pm}^a)_{1\le m\le M_p}$, μ_p^r, σ_p^r and $\tilde{\sigma}_p^r$ for all values of p between MIDI 36 (65.4 Hz) and MIDI 100 (2.64 kHz). We assume that the signal is sampled at 22.05 kHz and we compute signal frames with Hanning windows of length $W = 1024$ (46 ms) and stepsize $\tilde{W} = W/2$. We set Markov transition probabilities manually so that the mean object duration equals 0.5 s and the mean inactivity probability $P_Z = 0.98$. We set $M_{\max}$ to 60 after informal listening tests.

IV. Perceptually motivated distortion measures

Since the eventual receiver of the estimated pitched objects is the human auditory system, it is important to extract the most perceptually salient objects first. Thus the prior over the residual $e(t)$ must be related to the perceptual distortion between the observed signal and the model. We propose a family of distortion measures extending the measure proposed in [19]. These measures are based on splitting the residual and the observed signal into several auditory frequency bands and transforming their time-varying powers into a distortion value taking into account auditory masking effects.

A. Definition of the distortion measures

We define the residual power in band b of frame n by $\tilde{E}_{bn} = \sum_{f=0}^{W-1} v_{bf} g_f |E_{nf}|^2$, where $(E_{nf})_{0\le f\le W-1}$ are the complex discrete Fourier transform coefficients of $e_n(t)$, $(g_f)_{0\le f\le W-1}$ is the frequency response of the outer and middle ear as specified in [20] and $(v_{bf})_{0\le f\le W-1}$ is the frequency response of the gammatone filter modeling band b as given in [21], [19]. Similarly, we define the observed signal power in band b by $\tilde{X}_{bn} = \sum_{f=0}^{W-1} v_{bf} g_f |X_{nf}|^2$. Then we measure the bandwise distortion due to the residual by $D_{bn}^{(\alpha)} = \tilde{E}_{bn}(\tilde{X}_{bn}+\tilde{X}_{\min})^{\alpha-1}$, where $0 \le \alpha \le 1$ is an exponential scaling factor and $\tilde{X}_{\min}$ is a constant modeling the absolute hearing threshold as given in [19]. Finally, we define the total distortion in frame n as $D_n^{(\alpha)} = \sum_{b=1}^{B} D_{bn}^{(\alpha)}$.

B. Interpretation

The meaning of this distortion measure depends on the value of α. We provide below three different interpretations that are valid when the distortion is smaller than the observed signal.

When $\alpha = 0$, the proposed measure is equal to the measure defined in [19], where $D_{bn}^{(0)}$ is interpreted as the probability that a distortion is detected in band b of frame n and $D_n^{(0)}$ as the overall probability that a distortion is detected in that frame. This measure accounts for simple bandwise auditory masking rules [22] stating that the distortion is undetectable in a given band when the power ratio between the observed signal and the residual is above a certain signal-to-mask ratio or when the residual power is below the absolute hearing threshold. This is modeled by the fact that $D_{bn}^{(0)}$ is near zero as soon as $\tilde{E}_{bn}$ is a few decibels smaller than $\tilde{X}_{bn}$ or $\tilde{X}_{\min}$. Note that the signal-to-mask ratio is implicitly approximated as a constant, whereas experimentally it depends on the band and the tonality of the signals [22]. This approximation seems valid in a low bitrate context, since it affects small distortion values in the bands where the residual is close to the masking threshold, but not the overall distortion measure which remains dominated by high distortion values in other bands. This measure is also known to predict accurately more complex auditory masking phenomena [19].

When $\alpha = 0.25$, $D_{bn}^{(0.25)}$ models the specific *loudness* [22] of the residual in band b of frame n and $D_n^{(0.25)}$ its overall loudness in that frame, taking into account possible masking by the observed signal. In particular, when the residual is equal to the observed signal itself and well above the absolute threshold, the measured value $D_{bn}^{(0.25)} = \tilde{X}_{bn}^{0.25}$ is consistent with the standard approximate formula for specific loudness in the absence of masking given in [22][2]. When $\tilde{E}_{bn}$ becomes a

[1] http://staff.aist.go.jp/m.goto/RWC-MDB/

[2] This formula involves a slightly larger exponent $\alpha = 0.3$, but experimental data show that loudness grows more slowly at moderate levels.

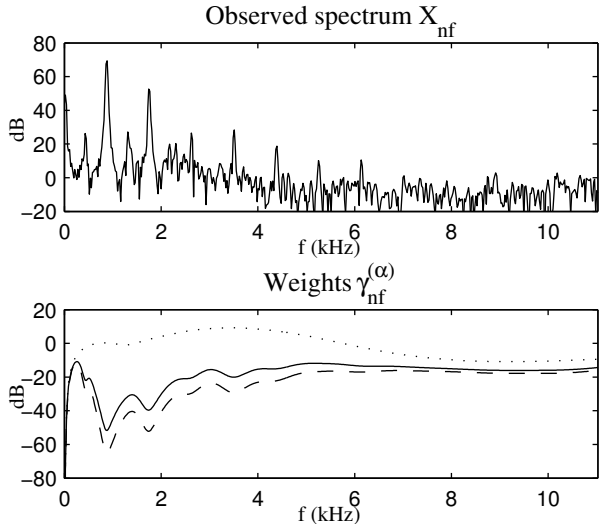


Fig. 3. Comparison of the perceptually motivated frequency weights for a 876 Hz flute note signal (dashed: $\alpha = 0$, solid: $\alpha = 0.25$, dotted: $\alpha = 1$).

few decibels smaller than $\tilde{X}_{bn}$ or $\tilde{X}_{\min}$, $D_{bn}^{(0.25)}$ drops quickly to zero in accordance with bandwise auditory masking rules. It is not known how well this measure approximates the loudness curve for intermediate values of $\tilde{E}_{bn}$, since this curve has not yet been measured experimentally in a masking context.

Finally, when $\alpha = 1$, $D_n^{(1)}$ corresponds to the power of the residual weighted by the frequency response of the outer and middle ear, without any masking effects.

C. Residual priors

For convenience, the above distortion measures can also be expressed equivalently as squared weighted Euclidean norms $D_n^{(\alpha)} = \sum_{f=0}^{W-1} \gamma_{nf}^{(\alpha)} |E_{nf}|^2$, where the weights are given by

$$\gamma_{nf}^{(\alpha)} = \sum_{b=1}^{B} v_{bf} g_f \left(\sum_{f'=0}^{W-1} v_{bf'} g_{f'} |X_{nf'}|^2 + \tilde{X}_{\min} \right)^{\alpha-1}. \quad (10)$$

The weight values corresponding to various values of α are plotted in Figure 3. Following the classical interpretation of Euclidean distortion measures as Gaussian priors, we derive the residual priors by $P(e_n) \propto \exp(-D_n^{(\alpha)}/(2(\sigma^e)^2))$. This results in the family of weighted Gaussian distributions

$$P(e_n) = \prod_{f=0}^{W-1} \mathcal{N}(E_{nf}; 0, \sigma^e (\gamma_{nf}^{(\alpha)})^{-1/2}). \quad (11)$$

V. Efficient Bayesian inference

The probabilistic signal model above can be used to infer the pitched objects representing a given signal using a MAP criterion, once the model hyper-parameters have been learnt. However, due to the complexity of the model, exact inference is intractable. The main issue is that the temporal continuity priors in (6-8) induce long-term dependencies between the parameters of different objects as soon as the temporal support of any object overlaps with the support of at least one other object. To overcome this issue, we propose a three-step approximate inference procedure: first we approximate the state and parameter priors by their marginals on each time frame and we estimate the MAP states on each frame separately; then we refine these estimated states using the exact state priors; finally we estimate the MAP parameters using the exact parameter priors while keeping the states fixed. These steps are described in more details in the following.

A. State and parameter marginal priors

The marginal prior corresponding to the state prior in (5) is a product of independent Bernoulli distributions

$$P(S_n) = \prod_{p \in \mathcal{A}_n} (1 - P_Z) \prod_{p \notin \mathcal{A}_n} P_Z, \quad (12)$$

where P_Z is the mean inactivity probability. Assuming that $\tilde{\sigma}^f \ll \sigma^f$, $\tilde{\sigma}^a_{pm} \ll \sigma^a_p$ and $\tilde{\sigma}^r_p \ll \sigma^r_p$, *i.e.* that the frame-to-frame parameter variation range is much smaller than the overall variation range, it is easy to show that the marginal priors corresponding to the parameter priors in (6-8) are given approximately by the log-Gaussian distributions

$$P(\log f_{pn}) = \mathcal{N}(\log f_{pn}; \log \mu^f_p, \sigma^f), \quad (13)$$

$$P(\log a_{pmn} | r_{pn}) = \mathcal{N}(\log a_{pmn}; \log(r_{pn} \mu^a_{pm}), \sigma^a_p), \quad (14)$$

$$P(\log r_{pn}) = \mathcal{N}(\log r_{pn}; \log \mu^r_p, \sigma^r_p). \quad (15)$$

B. Search within the local state space

The MAP state S_n is estimated on each time frame n via an iterative stochastic jump algorithm [23]. The algorithm starts with a single state hypothesis $\hat{S}_n$ where all the pitches $\hat{S}_{pn}$ are inactive. Then, at each iteration, the past state hypotheses are sorted according to their posterior probability $P(\hat{S}_n | x_n)$ and each of the N_{best} best hypotheses generates several additional state hypotheses: $|\mathcal{A}_n|$ hypotheses where one active pitch $\hat{S}_{pn}$ is deactivated, plus N_{add} hypotheses where one inactive pitch $\hat{S}_{pn}$ is activated. The most promising pitches to be activated are preselected as those giving the largest dot product between the residual spectrum $|E_{tf}|$ and the normalized average note spectra derived from $(\mu^a_{pm})_{1 \le m \le M_p}$. The algorithm stops when additional hypotheses do not improve the posterior probability and the best past hypothesis is selected. This avoids the need to test all possible states, which is not feasible. For instance, there are about 10^8 possible states on a typical scale of 65 semitones for a maximal number of 6 concurrent pitches.

C. Bayesian marginalization

Given a state hypothesis $\hat{S}_n$, the state posterior $P(\hat{S}_n | x_n)$ is the integral of the joint posterior $P(\hat{S}_n, f_n, a_n, r_n, \phi_n | x_n)$ over the parameters $f_n = (f_{pn})_{p \in \mathcal{A}_n}$, $a_n = (a_{pmn})_{p \in \mathcal{A}_n, 1 \le m \le M_p}$, $r_n = (r_{pn})_{p \in \mathcal{A}_n}$ and $\phi_n = (\phi_{pmn})_{p \in \mathcal{A}_n, 1 \le m \le M_p}$. The computation of this integral is known as the *Bayesian marginalization* problem [23]. Numerical integration by sampling of the joint posterior on a regular grid is intractable since the number of parameters per frame is typically of the order of one hundred or more. MCMC sampling schemes [23] lead to tractable computation, but remain rather slow.

An alternative approach is to estimate the MAP parameters $(\hat{f}_n, \hat{a}_n, \hat{r}_n, \hat{\phi}_n) = \arg\max P(\hat{S}_n, f_n, a_n, r_n, \phi_n | x_n)$ using a standard optimization algorithm[3] and to approximate the joint posterior around these values by a simpler distribution whose integral can be computed analytically. Popular approximations include the "delta" approximation, which is equal to the maximum of the joint posterior $P(\hat{S}_n, \hat{f}_n, \hat{a}_n, \hat{r}_n, \hat{\phi}_n | x_n)$ and mostly relevant for fixed-size models [18], the Laplace approximation [24], which replaces the posterior by a Gaussian distribution with full covariance matrix and performs unbounded integration, and the diagonal Laplace approximation [24], which similarly replaces the posterior by a Gaussian distribution with diagonal covariance. In the following, these approximations are applied to the unbounded log-parameters $\log f_{pn}$, $\log r_{pn}$ and $\log a_{pmn}$ to improve their precision [24]. Note that the diagonal Laplace approximation performs bounded integration over each phase parameter ϕ_{pmn} in $[-\pi, \pi]$.

The latter can be improved without increasing the computational cost by factorizing the posterior as a product of Gaussian and non-Gaussian univariate distributions. Analysis shows that the value of the joint posterior with respect to the phase parameter ϕ_{pmn}, while keeping all other parameters fixed, is proportional to $\exp(-2\hat{c}^{\phi}_{pmn} \sin^2((\phi_{pmn} - \hat{\phi}_{pmn})/2))$, where $\hat{c}^{\phi}_{pmn} = -\partial^2 \log P(\hat{S}_n, f_n, a_n, r_n, \phi_n | x_n)/\partial \phi^2_{pmn}$ denotes the curvature of the log-posterior at its maximum with respect to ϕ_{pmn}. The expression of the posterior with respect to the log-amplitude parameter $\log a_{pmn}$ is more complex and involves four variables. To maintain fast computation, we approximate it as with the diagonal Laplace approximation by the Gaussian shape $\exp(-\hat{c}^{a}_{pmn}(\log a_{pmn} - \log \hat{a}_{pmn})^2/2)$, where $\hat{c}^{a}_{pmn} = -\partial^2 \log P(\hat{S}_n, f_n, a_n, r_n, \phi_n | x_n)/\partial (\log a_{pmn})^2$ is the curvature of the log-posterior at its maximum with respect to $\log a_{pmn}$. Using the delta approximation for f_n and r_n $P(\hat{S}_n | x_n) \approx P(\hat{S}_n, \hat{f}_n, \hat{r}_n | x_n)$, this gives

$$P(\hat{S}_n | x_n) \approx P(\hat{S}_n, \hat{f}_n, \hat{a}_n, \hat{r}_n, \hat{\phi}_n | x_n) \prod_{\substack{p \in \mathcal{A}_n \\ 1 \leq m \leq M_p}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}\hat{c}^{a}_{pmn} y^2} \, dy \int_{-\pi}^{+\pi} e^{-2\hat{c}^{\phi}_{pmn} \sin^2 \frac{y}{2}} \, dy. \quad (16)$$

The two integrals involved in each term of this product are functions of $\hat{c}^{a}_{pmn}$ and $\hat{c}^{\phi}_{pmn}$, computed analytically [24] and by tabulation respectively.

The precision of these approximations is difficult to assess, since the exact state posterior is unknown in general. However, in the case of a single hypothesized note, analysis shows that the parameters $(\log a_{pmn}, \phi_{pmn})$ of each partial are independent a posteriori from those of other partials given f_{pn} and r_{pn}. Thus $P(\hat{S}_n, \hat{f}_{pn}, \hat{r}_{pn} | x_n)$ can be computed exactly but slowly by separate numerical integration over the parameters of each partial. A comparison of various approximations in this case is provided in Figure 4. The delta approximation and the full Laplace approximation provide erroneous pitch estimates, since their maxima do not correspond to the true pitch. This is due respectively to the fact that low pitch notes involve a much larger number of parameters and that phase parameters are bounded. The diagonal Laplace approximation provides a good pitch estimate, but remains significantly different from the exact marginal. The proposed approximation appears the closest to the exact marginal for all hypothesized pitches.

[3]In the following, we use the subspace trust region algorithm implemented in Matlab's lsqnonlin function. Details about this algorithm are available at www.mathworks.com/access/helpdesk_r13/help/toolbox/optim/lsqnonlin.html

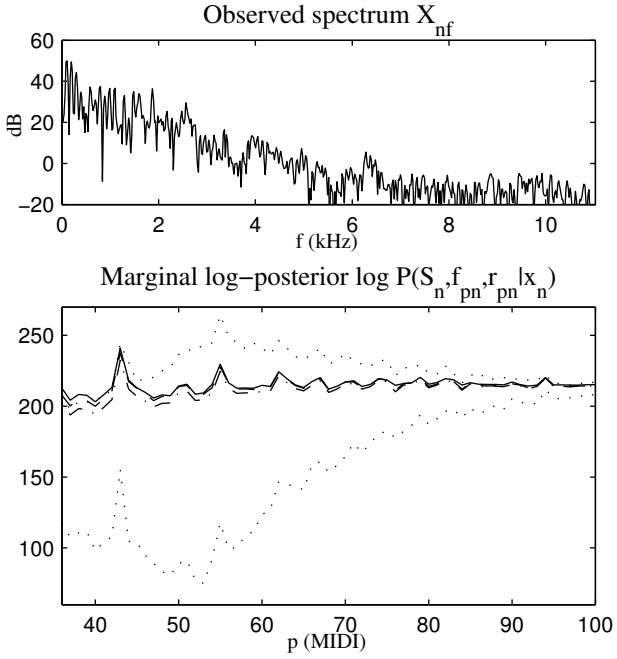


Fig. 4. Comparison of Bayesian marginalization methods for a cello note signal with pitch $p = 43$ as a function of the hypothesized pitch (solid: exact marginal log-posterior, dashed: proposed approximation, dash-dotted: diagonal Laplace approximation, dotted top: full Laplace approximation, dotted bottom: delta approximation).

D. Search within the global state space

Viterbi decoding of the MAP state path $(\hat{S}_n)_{1 \leq n \leq N}$ corresponding to the true state prior in (5) is intractable due to the large size of the state space. We tried using beam search techniques [17] to prune out improbable paths but found them experimentally unreliable: pruning errors sometimes led to important parts of the original signal being omitted from the resynthesized signal. Instead we provide an initial estimate using the MAP state path estimated previously from the marginal state prior in (12) and we iteratively perform an exact Viterbi decoding for each discrete pitch until a local maximum of the posterior has been reached. The associated MAP parameters $(\hat{f}_n, \hat{a}_n, \hat{r}_n, \hat{\phi}_n)_{1 \leq n \leq N}$ corresponding to the marginal parameter priors in (13-15) are also estimated as part of the marginalization algorithm.

E. Refining of the parameters

Finally, we fix the state path and we reestimate the MAP parameters corresponding to the true parameter priors in (6-8) using the same optimization algorithm. Rigorous optimization is computationally intensive because all the parameters depend on each other as soon as any object temporally overlaps with at least one other object. Thus we iteratively update the MAP

parameter values for each object until a local maximum of the posterior has been reached.

VI. Coder design

Once the underlying pitched objects have been estimated, the observed signal can be compressed by jointly quantizing the parameters of each object. At low bitrate, phase parameters are generally discarded, since resynthesizing each partial with a random initial phase results in little or no quality degradation [5]. Existing quantization algorithms encode fundamental frequency by differential quantizing [5] and log-amplitudes by differential quantizing [5], attack-decay-sustain-release (ADSR) interpolation [7] or adaptive temporal interpolation [25]. Compression can be increased by grouping high-frequency partials into subbands and encoding the total amplitude in each subband [5] or by replacing individual amplitudes with a small number of coefficients modeling the spectral envelope such as Log-Area Ratio (LAR) coefficients [8] or Mel-Frequency Cepstral Coefficients (MFCC) [7]. These algorithms rely on the estimated pitched objects exhibiting the same properties as musical notes, including frequential and temporal smoothness. In practice, we observed that they often result in large quality degradations since these properties do not hold true for all objects. Instead we perform adaptive linear frequential and temporal interpolation and differentially encode the parameters at interpolation breakpoints.

A. Adaptive frequential and temporal interpolation

The proposed algorithm, inspired from a simpler algorithm in [25], estimates the minimal number of interpolation breakpoints for each object given a maximal distortion threshold $D_{\max}$ and the encodable range of each variable defined by its minimum and maximum quantized values.

Frequential breakpoints are estimated in a first step by scanning the partials in decreasing order. The highest partial is set as a breakpoint $m_1 = M_p$. Then a given partial m is added as a breakpoint $m_j = m$ if either the distortion $D_n^{(\alpha)}$ resulting from frequential interpolation of log-amplitudes between previous breakpoints and $m-1$ is larger than $D_{\max}$ on at least one time frame n, or the log-amplitude difference $\log a_{p,m-1,n} - \log a_{p,m_{j-1},n}$ is outside the encodable range for at least one time frame n. The fundamental is added as the last breakpoint $m_J = 1$.

Similarly, temporal breakpoints are estimated in a second step by scanning the time frames in increasing order. The first frame is set as a breakpoint $n_1 = n_{\text{on}}$. Then a given frame n is added as a breakpoint $n_k = n$ if either the distortion $D_n^{(\alpha)}$ resulting from frequential and temporal interpolation of log-amplitudes between previous breakpoints and $n+1$ is larger than $D_{\max}$ for at least one time frame, or the log-fundamental frequency error resulting from temporal interpolation of the log-fundamental frequency between previous breakpoints and $n+1$ is larger than a fixed threshold $F_{\max}$ for at least one time frame, or the log-amplitude difference $\log a_{p,m_j,n+1} - \log a_{p,m_j,n_{k-1}}$ is outside the encodable range for at least one partial m_j, or the log-fundamental frequency difference $\log f_{p,n+1} - \log f_{p,n_{k-1}}$ is outside the encodable range. The last frame is also added as a breakpoint $n_K = n_{\text{off}}$.

This algorithm is run several times after adapting the distortion threshold $D_{\max}$ by bisection until the target bitrate is reached. The fundamental frequency threshold $F_{\max}$ is set to 10 cents (1/120 octave) in the following.

TABLE I
Bitrate allocation for each object*

Variable	Number of bits
Discrete pitch p	7
Onset frame n_{on}	7 (7)
Duration $n_{\text{off}} - n_{\text{on}} + 1$	8
Number of freq. breakpoints J	$\log_2(M_p)$
Freq. breakpoint m_j	3 (0)
Number of temp. breakpoints K	$\log_2(n_{\text{off}} - n_{\text{on}} + 1)$
Temp. breakpoint n_k	3 (0)
Fundamental frequency f_{p,n_k}	3 (4)
Amplitude a_{p,m_j,n_k}	4 (5)

* For differentially encoded variables, the number of bits for the initial value is indicated in parentheses.

B. Bitrate allocation

The full bitrate allocation scheme is detailed in Table I. Fundamental frequency and amplitude values at the breakpoints are differentially encoded using quantization steps of 10 cents and 3 dB respectively, corresponding to encodable ranges of $[-40, +30]$ cents and $[-24, +21]$ dB. The positions of the frequential and temporal breakpoints and the onset times are also differentially encoded. The delay between onsets and the object duration are limited to 127 frames and 256 frames respectively, which is sufficient for the considered data. Larger delays may be encoded using dummy objects with zero duration. A larger duration limit may be needed for other data.

VII. Evaluation

We evaluated the proposed object coding system on several ten-second items equalized in loudness and sampled at 44.1 kHz: five excerpts of solo instruments (flute, clarinet, oboe, violin, cello) from the SQAM database[4] and five excerpts of chamber music from commercial CDs (flute & clarinet, violin duo, violin & cello, flute & violin & cello, string quartet). The signals were resampled to 22.05 kHz prior to encoding and framed as in Section III-D.

The computational cost of the system is dominated by the cost of Bayesian inference, which depends on the number of tested states. The stochastic state jump algorithm with $N_{\text{best}} = N_{\text{add}} = 3$ (see Section V-B) resulted in an average of 24 tested states per frame and 30 minutes of computation time per second of signal on a 2.8 GHz computer with Matlab. This is about 5 times faster than the computation time reported in [16] on a similar platform, despite the greater complexity and the larger number of parameters of the proposed model. The estimated transcriptions contained up to 5 concurrent objects, with an average of 1.9 objects and 106 parameters per frame.

The performance was measured by means of two listening tests following the MUSHRA standard [26] involving eight

[4] http://www.ebu.ch/en/technical/publications/tech3000_series/tech3253/

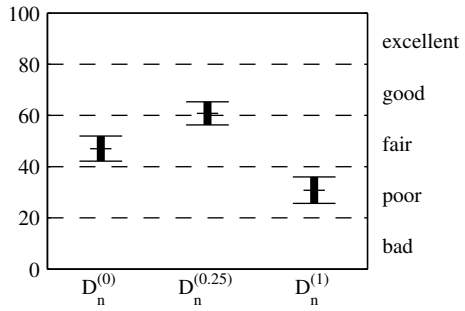


Fig. 5. Subjective quality of the encoded signals before parameter quantization using various distortion measures. Bars indicate 95% confidence intervals over 10 test items and 8 subjects.

and seven subjects respectively[5]. After a training phase, the subjects were asked to rate the quality of the encoded signals compared to the original signals on a scale between 0 and 100, partitioned into five intervals labelled "bad" to "excellent".

A. Comparison of the distortion measures

The first test aimed to select the best distortion measure $D_n^{(\alpha)}$ among the ones discussed in Section IV-B by comparing the quality of encoded signals before parameter quantization using the same number of frequency and amplitude parameters. The tradeoff between quality and model size depends on the standard deviation of the residual σ^e. Experimentally, the number of parameters and thus the quality always decreased when σ^e increased. We chose a fixed value of σ^e for all items for $D_n^{(0.25)}$ such that it resulted in little or no degradation but that larger values of σ^e resulted in very noticeable degradation according to informal listening tests. Then we estimated by bisection the values of σ^e for $D_n^{(0)}$ and $D_n^{(1)}$ for each item so as to obtain the same number of parameters as with $D_n^{(0.25)}$.

The results of the test are presented in Figure 5. The loudness distortion measure $D_n^{(0.25)}$ resulted in a significantly higher quality than other distortion measures and was selected in the following. This is an important result, since existing parametric coding methods are often based on $D_n^{(0)}$ instead. This result is valid only when the target number of parameters remains close to the critical number set in this test. Indeed, all distortion measures perform equally well when a very large number of parameters is allowed, but this test shows that $D_n^{(0.25)}$ can achieve a fair to good quality using less parameters than other distortion measures. Further experiments are needed to determine whether other values of α further improve quality. However a larger number of subjects may be necessary to obtain significant results.

Note that the proposed object extraction strategy using $D_n^{(0.25)}$ is similar to the loudness maximization principle for parametric coding, introduced in [27] but not validated by formal listening tests. While the former seeks to minimize the loudness of the residual taking into account possible masking by the observed signal, the latter seeks to maximize the overall loudness of the extracted objects independently of the observed signal. In theory, it is possible to find situations where different objects are extracted depending on the strategy. Further experiments are needed to determine how often such situations arise in practice, and which strategy is preferable from a perceptual point of view.

[5]These listening tests were performed using the MUSHRAM interface for Matlab available at http://www.elec.qmul.ac.uk/digitalmusic/downloads/

B. Comparison between the proposed coder and other coders

The second test concerned the comparison of the proposed coder after parameter quantization at 2 kbit/s and 8 kbit/s with baseline transform and parametric coders and with two "anchor" signals: the original signal low-pass filtered at 3.5 kHz and the signal encoded with the proposed method without parameter quantization. We chose a standard MPEG-1 Layer 3 transform coder called Lame[6]. Comparison with standard parametric coders MPEG-4 SSC and HILN could not be conducted since they are not publicly available and their implementation in MPEG-4 reference software is not designed to be competitive[7]. Thus we designed similar coders.

A baseline parametric coder was implemented as follows. First sinusoids are extracted in each time frame n using matching pursuit [2] with $D_n^{(0.25)}$ until the distortion becomes lower than a threshold $D_{\max}$. Then sinusoidal tracks are formed using a simple sinusoidal tracking algorithm [10]. Despite its simplicity, this algorithm is nearly optimal for coding purposes [28]. Frequency and amplitude parameters are differentially encoded with the same bitrate allocation as objects shown in Table I, while phase parameters are discarded. This algorithm is run several times after adapting the distortion threshold $D_{\max}$ by bisection until the target bitrate is reached. We tested several possible modifications of this algorithm, such as using the distortion measure $D_n^{(0)}$, removing short duration tracks or quantizing phase parameters. All these modifications resulted in a lower quality according to informal listening tests and were not incorporated in the following.

A hybrid object/sinusoidal coder similar to HILN was also implemented. Pitched objects are extracted using the proposed object model with $D_n^{(0.25)}$ under the constraint that at most one object be present on each time frame, taken into account by modifying the state prior in (5) [29]. Then sinusoidal tracks are extracted from the residual signal and encoded simultaneously with pitched objects by adapting the same distortion threshold $D_{\max}$ until the target bitrate is achieved.

The resulting sound files are available for listening online[8] and the results of the listening test are summarized in Figure 6. The proposed object coder achieves a significantly better performance than the other coders at the same bitrate, despite the fact that all coders (except the transform coder) are based on the same distortion measure. More precisely, the proposed coder employed at 2 kbit/s results in a fair to good quality, similar to that of other coders employed at 8 kbit/s, whereas the quality of the sinusoidal and hybrid coders at 2 kbit/s is bad to fair. The quality degradation of the object coder due to

[6]http://lame.sourceforge.net/, used with the settings -h –abr 8

[7]The authors of SSC agreed to provide sound files encoded with SSC at its target bitrate of 24 kbit/s, but could not do so for lower bitrates since this would have required a long manual optimization process.

[8]http://www.elec.qmul.ac.uk/people/emmanuelv/coding_demo/

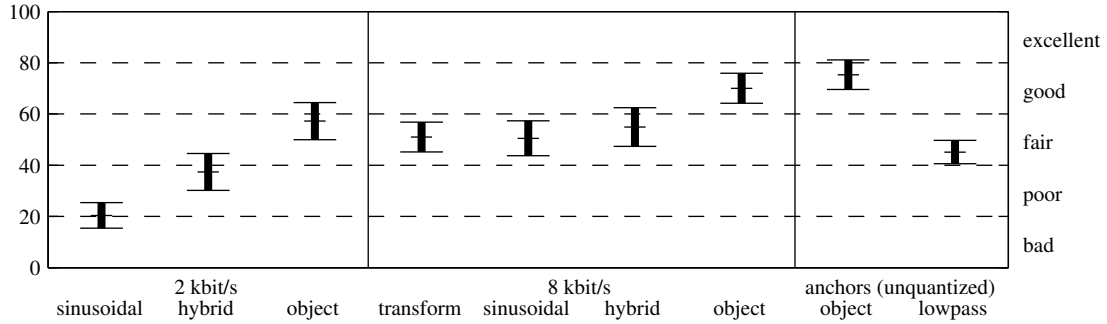


Fig. 6. Subjective comparison of the proposed coder with "anchors" and baseline coders at 2 kbit/s and 8 kbit/s. All coders except the transform coder are based on the same distortion measure. Bars indicate 95% confidence intervals over 10 test items and 7 subjects.

parameter quantization at 8 kbit/s is small, which supports the efficiency of the proposed adaptive interpolation scheme.

Comparison of Figures 5 and 6[9] suggests that the quality increase achieved using harmonic objects instead of standalone sinusoidal tracks is slightly larger than that obtained using $D_n^{(0.25)}$ instead of $D_n^{(0)}$ at 8 kbit/s, and much larger at 2 kbit/s. Thus the performance of the proposed system can be explained both by the object model and the loudness distortion measure. However the former contributes more at very low bitrates.

Detailed results on each test item are not presented here, since the number of subjects participating in the listening test is too small to draw significant conclusions. Nevertheless, results suggest that the quality achieved by the proposed coder for a given bitrate appears to be lower for signals involving low pitch instruments or several instruments, which might be expected since they contain a larger number of sinusoidal partials to be encoded. Also the quality before quantization seems to be lower for instruments exhibiting sharp onsets, bow noise or breath noise, which cannot be encoded in terms of the pitched objects employed in the current system.

It is interesting to note that the polyphonic pitch transcription estimated as part of the proposed coding strategy is not perfect: it contains a few spurious notes with short duration, often located at upper octave intervals of the actual notes, and sometimes short silences within notes. These transcription errors do not seem to affect the rendering of the original sounds, because coding is performed using an analysis-by-synthesis procedure on each frame. We conjecture that transcription errors are necessary to maximize the coding performance by discarding perceptually undetectable notes and rendering additional parts of the signal that do not fit the model, such as harmonic partials whose parameters do not fit the parameter priors (6-8) or transient and noisy parts. Further experiments are needed to verify this conjecture by embedding musical score information into the state prior (5) and measuring the quality of the resulting objects.

VIII. Conclusion

This article introduced a system for low bitrate coding of musical audio that represents a signal as a collection of pitched sound objects composed of harmonic sinusoidal partials. These objects are extracted using a Bayesian approach and an efficient estimation procedure. Their parameters are then quantized using adaptive frequential and temporal interpolation. Listening tests support the use of the proposed loudness distortion measure within the model. Further listening tests show that the proposed coder outperforms baseline transform and sinusoidal coders at 8 kbit/s and 2 kbit/s.

We are currently considering three further research directions. Firstly, the quality of the encoded signals is limited by the smoothing of note onsets and the non-rendering of bow noise or breath noise. These limitations do not seem fundamental issues at very low bitrates, where most of the quality degradation comes from parameter quantization, but they become critical at higher bitrates when a transparent quality is targeted. Parametric coders address these limitations using various models of onset and noise elements, whose parameters are estimated from the non-pitched residual by a deterministic procedure [3], [5]. However, these models cannot be considered as object models, since they do not separate out contributions from different instruments. For instance, the total noise produced by different instruments is modeled by a single colored noise model. We aim to develop these models into proper onset and noise object models and incorporate them in the current Bayesian framework. Similarly, we plan to incorporate pseudo-pitched objects composed of inharmonic partials, whose frequency relationships follow a learnt prior.

Secondly, the compression performance remains limited by the fact that the parameters of each object are quantized separately. We will investigate grouping objects into higher level instrument-like clusters and jointly encode the objects within each cluster by a limited number of timbre parameters. This may also help browsing the signal structure for indexing or interactive signal manipulation purposes.

Thirdly, while the proposed Bayesian marginalization procedure is faster than MCMC, it is still rather slow due to the very large number of parameters involved. Improved heuristic methods are needed to reduce the number of tested states. We also plan to investigate more flexible Bayesian marginalization procedures by combining the proposed factorial approximation with MCMC approaches and by trying to provide estimation bounds instead of a single value. This would allow a variable tradeoff between estimation accuracy and computational cost.

[9] In theory, Figures 5 and 6 cannot be directly compared since they were obtained from separate tests. Informal listening suggests that fixed perceptual differences correspond to slightly smaller rating differences in Figure 6.

ACKNOWLEDGMENTS

This work is funded by EPSRC grant GR/S75802/01. The authors wish to express their gratefulness to Bert den Brinker for adapting and running MPEG-4 SSC on the test files, Heiko Purnhagen for answering questions about MPEG-4 HILN and all the people who participated in the listening tests.

REFERENCES

[1] International Organization for Standardization, "ISO/IEC 14496-3:2001, Information technology – Coding of audio-visual objects – Part 3: Audio," 2001.

[2] R. Heusdens and S. van de Par, "Rate-distortion optimal sinusoidal modeling of audio and speech using psychoacoustical matching pursuits," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. II–1809–1812.

[3] A. C. den Brinker, E. G. P. Schuijers, and A. W. J. Oomen, "Parametric coding for high-quality audio," in *Proc. AES 112th Convention*, 2002, preprint number 5554.

[4] X. Amatriain and P. Herrera, "Transmitting audio content as sound objects," in *Proc. AES 22nd Conf. on Virtual, Synthetic and Entertainment Audio*, 2001, pp. 278–288.

[5] H. Purnhagen, B. Edler, and C. Ferekidis, "Object-based analysis/synthesis audio coder for very low bit rates," in *Proc. AES 104th Convention*, 1998, preprint number 4747.

[6] K. Melih and R. Gonzalez, "Audio object coding for distributed audio data management applications," in *Proc. Int. Conf. on Communication Systems (ICCS)*, 2002, pp. 727–731.

[7] M. Helén and T. Virtanen, "Perceptually motivated parametric representation for harmonic sounds for data compression purposes," in *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2003, pp. 249–253.

[8] B. Edler and H. Purnhagen, "Parametric audio coding," in *Proc. Int. Conf. on Signal Processing (ICSP)*, 2000, pp. 21–24.

[9] E. Vincent and M. D. Plumbley, "A prototype system for object coding of musical audio," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2005, pp. 239–242.

[10] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[11] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297–336, 1994.

[12] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 1996.

[13] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[14] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with Matching Pursuit," *IEEE Trans. on Signal Processing*, vol. 51, no. 1, pp. 101–111, 2003.

[15] P. J. Walmsley, S. J. Godsill, and P. J. W. Rayner, "Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1999, pp. 119–122.

[16] M. Davy, S. J. Godsill, and J. Idier, "Bayesian analysis of western tonal music," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2498–2517, 2006.

[17] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 679–694, 2006.

[18] E. Vincent, "Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux," Ph.D. dissertation, IRCAM, Paris, France, 2004.

[19] S. van de Par, A. Kohlrausch, G. Charestan, and R. Heusdens, "A new psycho-acoustical masking model for audio coding applications," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. II–1805–1808.

[20] International Organization for Standardization, "ISO 226:2003, Acoustics – Normal equal-loudness-level contours," 2003.

[21] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[22] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models, 2nd Edition*. Heidelberg: Springer, 1999.

[23] G. Casella and C. P. Robert, *Monte Carlo Statistical Methods, 2nd Edition*. New York: Springer, 2005.

[24] D. M. Chickering and D. Heckerman, "Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables," in *Proc. Conf. on Uncertainty in Artificial Intelligence (UAI)*, 1996, pp. 158–168.

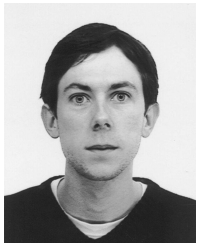
[25] A. K. Malot, P. Rao, and V. M. Gadre, "Spectrum interpolation synthesis for the compression of musical signals," in *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2001, pp. 184–188.

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

[27] H. Purnhagen, N. Meine, and B. Edler, "Sinusoidal coding using loudness-based component selection," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2002, pp. II–1817–1820.

[28] J. Jensen and R. Heusdens, "A comparison of differential schemes for low-rate sinusoidal audio coding," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 205–208.

[29] E. Vincent and M. D. Plumbley, "Predominant-F0 estimation using bayesian harmonic waveform models," in *Proc. Music Information Retrieval Evaluation eXchange (MIREX)*, 2005.



Emmanuel Vincent received the degree from the École Normale Supérieure, Paris, France, in 2001 and the Ph.D. degree in acoustics, signal processing and computer science applied to music from the University of Paris-VI Pierre et Marie Curie, Paris, in 2004.

He is currently a Research Assistant with the Centre for Digital Music at Queen Mary, University of London, Department of Electronic Engineering, London, U.K.. His research focuses on structured probabilistic modeling of audio signals applied to blind source separation, indexing and object coding of musical audio.

Mark D. Plumbley (S'88-M'90) began his research in the area of neural networks in 1987, as a PhD Research Student at Cambridge Unversity Engineering Department. Following his PhD he joined King's College London in 1991, and in 2002 moved to Queen Mary University of London to help establish the new Centre for Digital Music.

He is currently working on the analysis of musical audio, including automatic music transcription, beat tracking, audio source separation, independent component analysis and sparse coding. Dr Plumbley currently coordinates two UK Research Networks: the Digital Music Research Network (www.dmrn.org) and the ICA Research Network (www.icarn.org).