A PROTOTYPE SYSTEM FOR OBJECT CODING OF MUSICAL AUDIO

Emmanuel Vincent and Mark D. Plumbley[*]

Electronic Engineering Department, Queen Mary, University of London
Mile End Road, London E1 4NS, United Kingdom
emmanuel.vincent@elec.qmul.ac.uk

ABSTRACT

This article deals with low bitrate object coding of musical audio, and more precisely with the extraction of pitched sound objects in polyphonic music. After a brief review of existing methods, we discuss the potential benefits of recasting this problem in a Bayesian framework. We define pitched objects by a set of probabilistic priors and derive efficient algorithms to infer active objects and their parameters. Preliminary experiments suggest that the proposed method results in a better sound quality than simple sinusoidal coding while achieving a lower bitrate.

1. INTRODUCTION

Perceptual audio coding aims to reduce the bitrate required to transmit 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 that split the signal into an adaptive number of subbands and time frames and quantize them separately. Existing coders, such as MPEG4 AAC (Advanced Audio Coder), provide a near-transparent quality down to 50 Kb/s for mono signals but generate "birdies" artifacts under 15 Kb/s, caused by sound components appearing and disappearing successively [1]. Parametric coders result in a better quality at low bitrates by representing the signal as a sum of sound atoms whose structure is more adapted to musical audio. For instance, sinusoidal coders decompose the signal into a set of sinusoidal tracks, transients and background noise that are encoded separately. This improves the quality around 10Kb/s, but other kinds of artifacts appear at lower bitrates.

The term *object coding* has been used for parametric coding methods which try to group sound atoms into higher-level hierarchical sound objects. For example, harmonic sinusoidal tracks may be grouped into notes and further into instrumental sources. This potentially leads to a better perceptual quality at very low bitrates by allowing joint encoding of the sinusoidal tracks using attributes such as fundamental frequency and timbre. This also enables advanced coding paradigms such as minimization of the semantic distortion, that is ensuring the musically relevant parts of the signal are encoded prioritarily. For instance, bits may be saved by encoding accompaniment with a lower quality than melody and by removing stationary background noise entirely.

In this article, we focus on the definition and the extraction of pitched sound objects in musical audio. We propose a Bayesian extraction method whose strength is to exploit both psycho-acoustics and learnt parameter priors in order to estimate sets of harmonic sinusoidal tracks within an audio signal. After a brief review of existing object coding methods, we define pitched objects in Section 2 and point out the reasons to consider their extraction as a Bayesian estimation problem. In Section 3 we describe the probabilistic priors associated with these objects and we propose efficient estimation algorithms. Preliminary experiments are carried in Section 4 to validate the relevance of the method from the coding perspective. We conclude by listing further work directions.

2. OBJECT CODING STRATEGIES

2.1. Defining sound objects

Several definitions of sound objects have been proposed in the literature. *Auditory objects* are events perceived as a coherent whole regardless their source or meaning [2], whereas *semantic objects* are meaningful units organized hierarchically such as notes and instruments [3]. These definitions are too restrictive from the coding point of view. An *encoded object* should describe several auditory/semantic objects or only parts of a single auditory/semantic object when this provides a lower bitrate. Also different object properties may correspond to different hierarchies: for example instruments are grouped differently within an orchestra when the aim is to describe timbre or spatial direction. Similarly to [2], we assume that an encoded object is a particular element of a class of signal models described by a set of specific parameters (pertaining to this object only) and a set of shared parameters (pertaining to several objects). Object coding is often performed in two steps: first identify the model class and the model parameters of active objects then jointly encode these model parameters into specific and shared parameter sets.

2.2. Structured sparse coding

Sparse coding with dictionaries containing local sines and cosines [4] and possibly other atoms (damped sines, chirps, wavelets) has proved a successful parametric coding strategy. Each signal frame is decomposed iteratively as a weighted sum of significant atoms using Matching Pursuit with a perceptual stopping criterion. The significance map (*i.e.* the indices of significant atoms) is then encoded and atom weights are quantized. This analysis-by-synthesis method guarantees a good perceptual reconstruction on each frame but "birdies" may appear at low bitrates. Also joint encoding of atom weights is difficult: when frequency spacing between atoms is coarse or when instruments play *vibrato*, a given sinusoidal partial may be represented by several atoms whose number and positions vary between successive time frames.

Recent methods perform an approximate "molecular" Matching Pursuit that extracts several atoms at each iteration forming objects such as single-frame harmonic stacks [5] or single-frequency

[*]This work is funded by the EPSRC grant GR/S75802/01.

lines [6]. This generally improves the quality and reduces the bitrate required to encode the significance map. But these objects remain too small: time continuity and harmonicity are not exploited jointly to reduce the bitrate. Also no straightforward structure applies to signals with time-varying frequency content since frequencies are discretized from the beginning.

2.3. Pitch tracking and estimation of harmonic partials

Sinusoidal modeling, possibly extended with transient and background noise modeling [7], is another common parametric coding strategy. Spectral peaks are located within each signal frame and tracked along successive frames to generate sinusoidal tracks. Then inaudible tracks are removed, amplitudes and frequencies are quantized differentially for each track while phase is not transmitted. Tracks can further be grouped into harmonic stacks using auditory grouping principles and encoded more efficiently by quantizing the parameters of the upper partials relatively to the fundamental [3]. This yields a good compression factor, but the quality is often poor even before quantizing because upper partials of the notes are badly tracked when instruments play *vibrato* and spurious tracks appear within background noise. Also long time frames are needed for all partials of low pitch notes to be detected, which results in a smoothing of onset transients.

A more principled approach to obtain harmonic stacks is to estimate a set of fundamental frequencies and to compute the amplitudes of their harmonics. For monophonic recordings (*i.e.* without chords) fundamental frequencies can be estimated by smoothing the output of a monophonic pitch tracker [8] but otherwise a polyphonic pitch tracker is needed [9]. The MPEG4 HILN (Harmonic and Individual Lines plus Noise) coder [10] combines a single predominant harmonic stack with other standalone tracks. Amplitudes within a harmonic stack are typically described by quantized MFCCs or LPCs [9, 10] at a few key time frames, other frames being interpolated at the decoder [8, 9]. This top-down method leads to fewer artifacts than the previous bottom-up method because it avoids standalone tracks. Nevertheless, the coding performance remains similar to that of MPEG4 AAC [10]. In [8] note onsets and reverberation are badly rendered for monophonic solo instruments because reverberation induces a time overlap between successive notes, which defeats monophonic pitch trackers. In [9] amplitudes and phases of the harmonic partials may be discontinuous and inaccurate when partials from different notes overlap because they are estimated separately for each note on each frame. Also polyphonic pitch tracking based on the summary autocorrelation function may generate octave errors, which either produce a bad rendering of the original signal or an increase of the number of parameters to encode.

2.4. Proposed definition and estimation strategy for pitched objects

This review shows that encoding pitched musical sounds as objects is not yet a solved problem. We focus on this issue in this article, leaving encoding of percussion and noise sounds for further work.

In the following, we define a *pitched object* spanning several time frames by its fundamental frequency and by the amplitudes and phases of its harmonics on each time frame, and we express the observed signal by a sum of these pitched objects plus a *residual*. We set probabilistic priors on these parameters and estimate jointly the number of objects and all their parameters within the Bayesian framework. In other words, we perform polyphonic pitch tracking, but within a rigorous probabilistic framework that achieves analysis-by-synthesis on each frame. The anticipated advantage is that the parameter priors can be tuned to yield relevant objects for the compression purpose. For instance, we want the object model to be broad enough to synthesize real data with limited loss (allowing *vibrato* for example), while being constrained enough to avoid spurious or inaccurate parameters (preventing lower octave errors or amplitude discontinuities for example). In practice, we set a psycho-acoustically motivated prior on the residual and we learn priors that model spectral envelope and time continuity on a database of isolated notes.

3. BAYESIAN INFERENCE OF PITCHED OBJECTS

3.1. Existing Bayesian harmonic models

A family of Bayesian harmonic models has been proposed in the literature [11, 12]. These models are designed for a polyphonic transcription purpose and suffer two drawbacks for object coding. Firstly the number of partials per note follows a sparse prior which does not depend on the pitch of the note. This may induce aliasing for high pitch notes and low-pass filtering for low-pitch notes. For instance at a sampling rate of 22 KHz the highest violin note (MIDI 100) has only 4 partials below the Nyquist frequency, whereas the lowest cello note (MIDI 36) has 168. Informal listening tests showed that some upper partials can be removed but that at least 60 partials have to be kept for a good timbre rendering of this note. In [11], the number of partials is also allowed to vary between successive time frames, which typically generates “birdies”. Secondly, the distribution of the residual does not correspond to the auditory significance of events. The Gaussian noise model chosen in [11] results in low power components such as upper partials, onsets and reverberation not being transcribed despite their perceptual importance. On the contrary the autoregressive Gaussian model chosen in [12] may give too much importance to events occuring at frequencies where the residual is small. Another drawback is that the parameters of these models are estimated by computationally intensive particle filtering methods.

3.2. Proposed model

Let x_t be the t-th frame of the observed signal x defined by $x_t(u) = w(u)x(tS+u)$ where w is a window of length N and S is the stepsize. We define the signal corresponding to the o-th pitched object in this frame as

$$s_{ot}(u) = \sum_{h=1}^{H_o} a_{oht} w(u) \cos(2\pi f_{ot} h u + \phi_{oht}), \qquad (1)$$

where f_{ot} is its fundamental frequency and (a_{oht},ϕ_{oht}) are the amplitude and phase of its h-th partial. Then we develop x_t as

$$x_t(u) = \sum_{o \in \mathcal{O}_t} s_{ot}(u) + e_t(u), \qquad (2)$$

where $\mathcal{O}_t$ is the set of active objects on this frame and e_t is the residual non-pitched signal. We compute the complex Fourier transform of this residual for positive frequencies $0 \leq f \leq N/2$ by $\tilde{e}_{tf} = \sum_{u=0}^{N-1} e_t(u) \exp(-2i\pi f u/N)$.

Temporal continuity priors on frequency and amplitude parameters are necessary, but they result in a very costly estimation since they introduce dependencies between all parameters as soon as the temporal support of each object overlaps with the support of at least another object. We propose a two-step approximate estimation method, where a local estimation is performed first on each frame and then refined adding duration and continuity priors.

3.3. Estimation with local priors

Each object o is associated with a fixed latent fundamental frequency F_o belonging to the discrete MIDI semitone scale. We suppose that each point on the semitone scale corresponds to at most one object and that this object is active with a probability $1 - Z$ where Z is a sparsity factor. In this case, the number of modelled partials H_o is defined so that $H_o F_o$ is just below the Nyquist frequency. The prior for f_{ot} is set to a log-Gaussian

$$P(\log f_{ot}) = \mathcal{N}(\log f_{ot}; \log F_o, \sigma^f), \tag{3}$$

where $\mathcal{N}(\cdot; \mu, \sigma)$ is the univariate Gaussian density of mean μ and standard deviation σ. Following previous work of the authors on spectral additive models [13], the amplitudes of the partials are described as the product of a fixed normalized spectral envelope (m_{oh}), a latent log-Gaussian amplitude factor r_{ot} and a log-Gaussian residual, *i.e.*

$$P(\log a_{oht}|r_{ot}) = \mathcal{N}(\log a_{oht}; \log(r_{ot} m_{oh}), \sigma_o^a), \tag{4}$$

$$P(\log r_{ot}) = \mathcal{N}(\log r_{ot}; \mu_o^r, \sigma_o^r). \tag{5}$$

The phases of the partials are assumed to be uniformly distributed

$$P(\phi_{oht}) = 1/2\pi. \tag{6}$$

Finally, following recent results in perceptual audio coding [14], the prior for the residual is designed so that the quantitative importance of the signal in each auditory band is roughly proportional to its loudness. We define the excitation power of the signal in the auditory band centered at frequency f on frame t by $E_{tf} = \sum_{b=0}^{N/2} v_{fb} |\tilde{x}_{tb}|^2$ where $(\tilde{x}_{tb})$ is the complex Fourier transform of x_t and (v_{fb}) are coefficients modeling the frequency spread of the auditory band, and we derive the approximate loudness of the signal in this band by $L_{tf} = (g_f E_{tf})^{0.25}$ where g_f is the frequency response of the outer and middle ear at frequency f. We measure the amount of residual error by a weighted Euclidean distance such that the weighted excitation power is proportional to the loudness for each frequency. This distance corresponds to the prior

$$P(\tilde{e}_{tf}) = \mathcal{N}(\tilde{e}_{tf}; 0, \sigma^e (E_{tf}/L_{tf})^{1/2}). \tag{7}$$

Approximate maximum *A Posteriori* (MAP) estimates of active objects and their parameters are obtained via an iterative deterministic jump method. At first, all points on the semitone scale are associated with inactive objects. Then at each iteration at most one object is activated or disactivated to improve the total posterior probability. The MAP parameters for each set of objects are computed by an approximate second order Newton method. This avoids testing all possible sets of active objects, which are about 10^7 for a maximum number of 5 active objects between MIDI 36 and 100. In practice joint MAP estimation of active objects and their parameters does not give the expected results. The chosen priors over-penalize sets of objects containing low pitch notes because they do not match well the parameter distribution for notes with a large number of partials. In order to avoid this, we replace the posterior probability of each set of objects within the jump method by its integral over amplitudes $(\log a_{oht})$ and phases (ϕ_{oht}) using the approximate Laplace integration method [15].

3.4. Reestimation with duration and continuity priors

Multi-frame objects are built by grouping single-frame objects associated with the same discrete frequency along successive frames. This does indeed form longer objects, but many short duration objects or silence segments remain. For each discrete frequency, we replace the independent Bernoulli priors on the activity states by a two-state Markov prior [13] and we reestimate the MAP activity states and parameters. Exact estimation by Viterbi decoding is intractable since the resulting factorial Markov chain evolves in a large state space. We tried beam search techniques but found them experimentally not reliable: estimation errors sometimes led to important parts of the original signal missing. Instead we perform an exact Viterbi decoding for each discrete frequency iteratively until a local maximum of the posterior has been reached.

After this reestimation, frequency and amplitude parameters within each object may still be inaccurate and contain temporal discontinuities because the Newton estimation method falls into local maxima of the posterior when badly initialized. Thus we keep activity states fixed and we reestimate the parameters only adding the temporal continuity priors

$$P(\log f_{ot}|f_{o,t-1}) = \mathcal{N}(\log f_{ot}; \log f_{o,t-1}, \sigma^{f'}), \tag{8}$$

$$P(\log a_{oht}|a_{oh,t-1}) = \mathcal{N}(\log a_{oht}; \log a_{oh,t-1}, \sigma_{oh}^{a'}), \tag{9}$$

$$P(\log r_{ot}|r_{o,t-1}) = \mathcal{N}(\log r_{ot}; \log r_{o,t-1}, \sigma_o^{r'}). \tag{10}$$

4. PROTOTYPE EVALUATION

We built a prototype object coding system that implements the estimation strategy described in Section 3, discards phase parameters, encodes frequency and amplitude parameters by differential quantizing and resynthesizes each partial with a random initial phase. We compare this prototype with a simple sinusoidal parametric coder, where spectral peaks are tracked using SMSTools [1] with a detection threshold of -65 dB and encoded and resynthesized in the same way, and with a transform coder called FAAC [2].

We evaluate the performance of these three methods by performing informal listening tests on ten-second excerpts taken from commercial CDs and resampled at 22 KHz, including four solos (cello, clarinet, oboe, violin) and a duo (cello and flute). We use a scale of 65 semitones for F_o between MIDI 36 and MIDI 100. Hyper-parameters have the same values for all excerpts: σ^f, (σ_o^a), (σ_o^r), $\sigma^{f'}$, $(\sigma_{oh}^{a'})$ and $(\sigma_o^{r'})$ are learnt on a subset of the RWC Musical Instrument Database [3] whereas σ^e, Z and the Markov transition probabilities are set manually. Signal frames are computed with half-overlapping Hanning windows of length 1024 (46 ms). Bitrate allocation is the following for each object: 18 bits for onset and offset times, 11 bits for initial frequency, 3 bits for successive frequencies, 5 bits per partial for initial amplitudes and 4 bits per partial for successive amplitudes.

The resulting sound files are available for listening online on http://www.elec.qmul.ac.uk/people/emmanuelv/WASPAA05/ and the results are summarized in table 1. The mean bitrate provided by our prototype is 9.3 Kb/s. Compared with the baseline parametric coder, it provides a better quality and a coding gain of 1.5 to 5. Frequency sweep artifacts are removed since harmonicity constraints allows better tracking of the sines and birdies

[1] http://www.iua.upf.es/~sms

[2] http://www.audiocoding.com/

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

Coder	Kb/s	Quality
Transform	24 – 30	Medium to good. A few birdies.
Parametric	17 – 30	Bad. Very noticeable birdies and frequency sweep artifacts.
Object	5 – 18	Medium. No birdies but minor inaccuracies (smoothing of note onsets, timbre modification for cello, non-rendering of flute breathing noises).

Table 1: Summary of preliminary results

are removed since duration priors favor long duration objects. Harmonic constraints also help ruling out non-relevant sines and halving the bitrate by encoding only the fundamental frequency of each object instead of the frequency of each track. Compared with the transform coder, our prototype still achieves a coding gain of 1.5 to 5 but its quality is slightly lower. It is unlikely that the current version of the prototype achieves a higher quality because it does not encode inharmonic or fast-varying objects. However the quality of FAAC would be lower for the same bitrate, despite the fact that it uses more precise psycho-acoustical knowledge.

It is interesting to note that the polyphonic pitch transcription estimated as part of the proposed coding strategy is never perfect in these examples: it contains a few spurious notes with short duration, upper harmonics of the actual notes and short silences within notes. These transcription errors do not seem to affect the rendering of the original sounds, because transcription is performed using an analysis-by-synthesis procedure on each frame. Transcription errors may even be necessary to render parts of the signal that do not fit the harmonic signal model exactly. In practice we found that imposing a minimal note duration could decrease the rendering quality, even when it improved the transcription performance.

5. CONCLUSION

This article discussed the estimation of pitched objects for low bitrate coding of musical audio. We defined pitched objects as harmonic sinusoidal models whose parameters follow a set of probabilistic priors and we proposed efficient algorithms to estimate active objects and their parameters. Then we built a prototype coder based on simple differential quantizing of the estimated parameters. Informal listening tests suggested that it resulted in a better coding performance than a simple sinusoidal parametric coder.

We are currently considering three further research directions. Firstly, the differential encoding strategy used in the prototype is very suboptimal. We expect a coding gain of about 10 by joint-encoding frequency and amplitude parameters exploiting temporal evolution (continuity, attack-sustain-decay, vibrato) and instrument-specific spectral envelopes. Also some of the estimated objects are actually not heard and could hopefully be removed using a more detailed psycho-acoustical model after transcription. Secondly, the proposed approximate estimation algorithm is faster than particle filtering methods but still quite slow (between 5 and 10 hours per example with MATLAB on a recent PC). Heuristic methods are needed to reduce the number of tested parameters. Thirdly, the rendering quality provided by the prototype seems to suffer three limitations, as underlined in table 1. The smoothing of note onsets may be addressed by upsampling the amplitude and frequency parameters and reestimating them on shorter time frames. We found that frames of length 256 (12 ms) were sufficient to restore part of the attack strength of clarinet and oboe without reducing the quality of other parts of the signal, as can be heard on `http://www.elec.qmul.ac.uk/people/emmanuelv/WASPAA05/`. But specific onset objects seem to be needed for an improved rendering and there is no consensus to date on the definition and encoding of such objects. The modification of the timbre of low-pitch instruments is related to the phaseless resynthesis of the partials. However encoding all the phase parameters results in a much higher bitrate and determining which phase parameters are auditorily relevant is a challenging problem. Finally, the non-rendering of breathing noises could be coped with in an extended framework involving noise objects, but again the definition and encoding of such objects is quite an open question. Slightly inharmonic pitched objects should also be added for some instruments such as piano.

6. REFERENCES

[1] M. Erne, “Perceptual audio coders ‘what to listen for’,” in *Proc. AES 111th Convention*, 2001.

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

[3] K. Melih and R. Gonzalez, “Audio object coding for distributed audio data management applications,” in *Proc. ICCS*, 2002, pp. 727–731.

[4] T. Verma, “A perceptually based audio signal model with application to scalable audio compression,” Ph.D. dissertation, Stanford University, 1999.

[5] 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.

[6] L. Daudet, S. Molla, and B. Torrésani, “Towards a hybrid audio coder,” in *Proc. ICWAA*, 2004, pp. 13–24.

[7] S. Levine, “Audio representations for data compression and compressed domain processing,” Ph.D. dissertation, Stanford University, 1998.

[8] A. Malot, P. Rao, and V. Gadre, “Spectrum interpolation synthesis for the compression of musical signals,” in *Proc. DAFx*, 2001.

[9] M. Helén and T. Virtanen, “Perceptually motivated parametric representation for harmonic sounds for data compression purposes,” in *Proc. DAFx*, 2003.

[10] H. Purnhagen, “Advances in parametric audio coding,” in *Proc. WASPAA*, 1999, pp. 31–34.

[11] P. Walmsley, S. Godsill, and P. Rayner, “Polyphonic pitch tracking using joint Bayesian estimation of multiple frame parameters,” in *Proc. WASPAA*, 1999, pp. 119–122.

[12] M. Davy and S. Godsill, “Bayesian harmonic models for musical pitch estimation and analysis,” Cambridge University, Tech. Rep. CUED/F-INFENG/TR.431, 2002.

[13] E. Vincent and X. Rodet, “Music transcription with ISA and HMM,” in *Proc. ICA*, 2004, pp. 1197–1204.

[14] R. Der, P. Kabal, and W.-Y. Chan, “Towards a new perceptual coding paradigm for audio signals,” in *Proc. ICASSP*, 2003.

[15] D. McKay, “Choice of basis for Laplace approximation,” *Machine Learning*, vol. 33, no. 1, pp. 77–86, 1998.