Object-Coding for Resolution-Free Musical Audio

Stephen J. Welburn[1], Mark D. Plumbley[1] and Emmanuel Vincent[2]

[1]*Centre for Digital Music, Queen Mary, University of London, Mile End Road, London, E1 4NS. United Kingdom*

[2]*IRISA-INRIA, Campus de Beaulieu, 35042 Rennes cedex, France*

Correspondence should be addressed to Steve Welburn (`stephen.welburn@elec.qmul.ac.uk`)

ABSTRACT

Object-based coding of audio represents the signal as a parameter stream for a set of sound-producing objects. Encoding in this manner can provide a resolution-free representation of an audio signal. Given a robust estimation of the object-parameters and a multi-resolution synthesis engine, the signal can be "intelligently" upsampled, extending the bandwidth and getting best use out of a high-resolution signal-chain. We present some initial findings on extending bandwidth using harmonic models.

1. INTRODUCTION

There are large quantities of legacy 16bit/44.1kHz "CD quality" audio currently in use (as samples, digital recordings and CDs) which need to be fitted into the high-resolution workflow. To maximize the benefits of these libraries, they need to be upsampled to make use of the available bandwidth in the high-resolution domain.

The resolution of a signal is related to its capture or reproduction - it is a property of the interface between the physical world and the digital world (Figure 1). In order to reproduce a signal at its original resolution, the captured data can be stored and then output when required. To use the signal at other resolutions, the signal will need recoding. By selecting an appropriate format in which to store the signal, this recoding can be made as flexible as possible.

Most of the current formats for audio have addressed the compression issue - preserving a suitable signal whilst reducing the storage space required. These encodings have achieved low bit-rates by using redundancy reduction and irrelevancy reduction.

Redundancy reduction takes advantage of structure within the signal to represent the signal in a more compact manner. Information-theoretic compression algorithms allow this to be done efficiently (e.g. Huffman coding[10, 5]). Lossless encoders[8] such as SHORTEN [14] rely solely on redundancy reduction.

Irrelevancy reduction discards elements in the signal which have low impact on the final perception of the signal. Psychoacoustic principles are applied to achieve this - e.g. considering the frequency response of the auditory system, the threshold of hearing and masking effects. Irrelevancy reduction is a lossy compression technique - after discarding the irrelevant elements, the original signal can no longer be produced. A typical lossy compression scheme is MPEG-1 Layer 3 (MP3) [1, 5].

The main motivation behind object-coding work to date has been the production of low bit-rate encoding, as for use with mobile devices. Much of this work has concentrated on the "Harmonics and Individual Lines plus Noise" (HILN) model included in the MPEG-4 audio standard [2]. This models a signal as a harmonic component plus individual sinusoidal lines plus a "noise" element. This work has produced fair quality, low bit-rate (16kbit/s) representations of a signal [13]. More recent work has considered Bayesian extraction of harmonic models from an audio signal [19], again for achieving low bit-rates.

For high-resolution audio, the main concern is quality-of-sound rather than speed-of-transmission. Hence, rather than adopting a low bit-rate approach, an approach which maximimizes the quality available from the data is required. Thus we are looking at coding the signal to produce the highest quality, rather than coding to reduce the bit-rate (compression).

Consider a parameterised model for a signal, continuous in the time domain, and which estimates the evolution

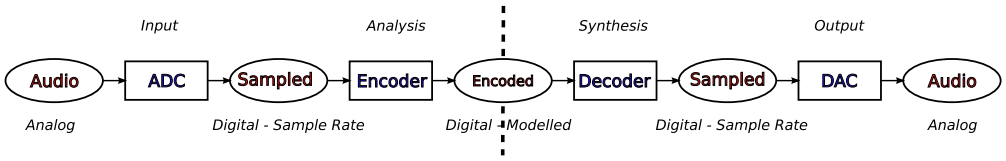


Fig. 1: Basic stages in encoding / decoding audio

of a signal over time. Using this approach, a signal can be coded as a series of parameters values for that model. The parameters can be represented at arbitrary precision, and resolution parameters will only be required to produce output at a chosen resolution - thus making the coded signal, effectively, resolution-free. In order to encode a given signal for use with the model, parameter values must be inferred from that signal. The precision chosen for the parameters will affect the maximum resolution available from the coding, but if a lossless coding scheme is adopted, the original captured signal can be restored and recoded at a higher precision when required, to the maximum working precision of the models used.

The captured digital signal is usually considered as the "ground truth" that needs to be reproduced. However, the "best" output signal is actually the one that most closely resembles the original source material. Given suitable models of the original source (for example based on instrument identification or simply prior knowledge of the signal's content) a better approximation of the source may be achievable by inferring model parameters that could have produced the captured signal, and creating a signal using those parameters to drive the model at the full available output resolution. In the speech domain, work has been carried out on modelling speakers voices and thus extending the bandwidth of narrow-band signals [3].

If the model is appropriate for representing the signal, then a large proportion of the signal content will be coded in the parameters, and effective coding of the modelled content will be possible. This coding may result in compression of the signal, as the model encapsulates information regarding the signal. However, our focus will be on the theory that, given an appropriate model for the signal and a set of parameter values inferred from a standard bit-rate signal, the output of the model at a higher resolution should be an appropriate representation of the source signal at that higher resolution.

2. OBJECT-CODING

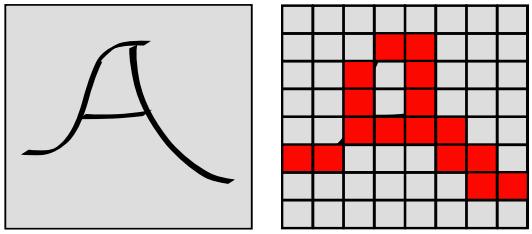
For object-coding, a signal is modelled as the output of a combination of individual, parameterised objects. Typical objects in image coding include Bezier curves [11]. For encoding video, the objects may represent elements in the image - for example representing real-world objects as shaded / textured geometric shapes - and suitable time-varying parameters indicate the positions of the objects and their orientations. Such techniques are used in scene analysis and image tracking [7]. Typically, parameter estimation is a complex procedure usually performed off-line, however real-time models have been developed, e.g. for capturing vector graphics [15].

For an audio signal, our objects are individual "voices" in the signal. Voices may represent individual instruments used to produce the original signal (e.g. plucked strings [17]); groups of instruments (e.g. the violin section); or simply the most practical elements to use to approximate the signal (e.g. sine, sawtooth, square and other basic waves or the MPEG-4 HILN model [12]). Each voice combines a synthesis engine with an appropriate set of parameters - such as onset time, duration, pitch, attack-rate - and an analysis engine to infer suitable parameter values from a signal. Synthesis techniques encapsulated in the voices may vary from simple sinusoidal synthesis models to complete physical models of instruments, and may allow the signal to be represented in a format such as MPEG-4 Structured Audio [16, 18].

As the voices aim to represent relevant phenomena present in the signal, this can allow physically appropriate expansion of a signal to take advantage of the available output bandwidth. For example, the output from the model can use the full bit-depth available, requantizing from 16-bit to 24-bit audio, and higher temporal resolution, increasing the available frequency range to reduce aliasing and other artefacts.

As an analogy, consider a hand-written document - the document itself does not have a "resolution" (Figure 2a).

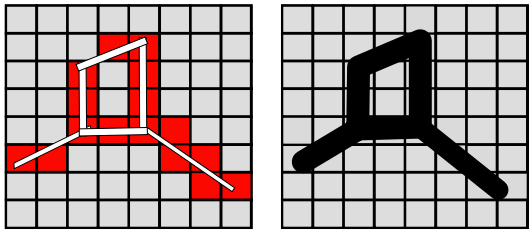
Scanning the document onto a computer, the maximum resolution available is a property of the scanner (Figure 2b).



(a) **(b)**

Fig. 2: (a) Original and (b) Scanned Graphic

Choosing to model the image as a set of lines, a "best-fit" representation can be inferred from the scanned image (Fig 3a) and the original image can be reproduced from those lines at the highest output resolution available (Figure 3b).

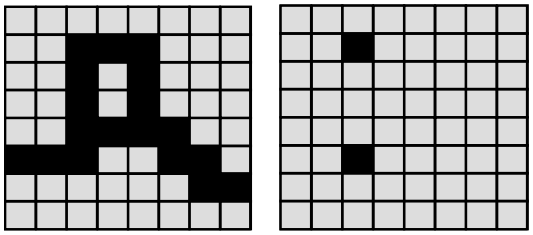


(a) **(b)**

Fig. 3: (a) Modelled and (b) High resolution image

Although the approximation does not match the source material, it exhibits fewer of the artefacts caused by the low-resolution capture process. The "vector" graphic doesn't allow the recovery of the data as originally captured. An additional "residual" signal must be kept, indicating how resynthesizing the modelled data at the original resolution (Figure 4a) differs from the original signal captured (Figure 2b). The residual element (Figure 4b) is *not* resolution-free - it has the same resolution as the original uncoded signal - but the use of suitable models for the signal should give a residual much smaller than the original signal.

Including the residual with the parameters means that it is always possible to recover the original signal as captured by simply regenerating the objects at the original resolution and adding the residual component. This offers a lossless method of encoding - the original signal is always retrievable.



(a) **(b)**

Fig. 4: (a) Reconstructed image and (b) Residual

Returning to audio object-coding, and the voices model, permitting an object-type of "audio sample" allows an audio signal to be represented losslessly as a set of modelled voices, and a residual "sample" object. Real-world audio will not be perfectly represented by the models and a residual element will be necessary, creating a trade-off between the lossless aspects of the coding and the resolution-free component. In particular, noise elements, which vary randomly across time, may be modelled in a perceptually identical manner but will differ from the original captured audio. This can be catered for in the coded file by storing alternative representations of the signal - i.e. *both* the residual signal and a "best modelled approximation" of it (for example using models of white / pink noise or representing inaudible residuals as silence). Resolution-free content can then be used for playback whilst still maintaining the captured audio.

The lossless encoding retains a "master" copy of the original signal. As improved voice models (and/or inference procedures) become available, the original signal can be recovered and then recoded - reducing the residual "unexplained" element and improving the high-resolution output.

3. EXPERIMENTAL METHOD

Given that onset detection [4, 6] is an active research areas, the assumption was made that suitable techniques will eventually become available for use in object-coding. An annotated signal is thus taken as the starting point, for which the onsets are known.

3.1. Voices Used

Initial experiments considered voices encapsulating basic waveforms. The waveforms considered were:

- sine, $s_n = \begin{cases} 1 & n = 1 \\ 0 & \text{otherwise} \end{cases}$
- sawtooth, $w_n = \frac{1}{n}$
- square, $q_n = \begin{cases} \frac{1}{n} & n \text{ odd} \\ 0 & \text{otherwise} \end{cases}$
- triangle, $t_n = \begin{cases} \frac{1}{n^2} & n \text{ odd} \\ 0 & \text{otherwise} \end{cases}$

These were chosen as they are well-known, and have a well-define harmonic structure which can be produced for a given sample-rate by limiting the number of harmonics such that the highest harmonic remains beneath the Nyquist frequency.

Given a signal with known onset times, it can be split into segments in which we expect constant voicing - the same set of voices should be producing the signal across the segment as new voices would have caused additional onsets. An autocorrelation based pitch detection algorithm was applied to the segments, and the modal MIDI pitch for selected as the pitch value for that segment.

Given the pitch values, a single harmonic structure for each segment was then estimated from the signal as follows.

3.2. Estimation of harmonic structure

Consider a real signal $\mathbf{x}$ composed of H harmonically related elements, and with F frames $\mathbf{x}_i$. Let each frame be composed of N components $x_{i,n}$.

$$\begin{aligned} x_{i,n} &= \sum_{h=1}^{H} a_{i,h} \cos(h\omega n + \Phi_{i,h}) \\ &= \sum_{h=1}^{H} a_{i,h} (\cos(h\omega n)\cos(\Phi_{i,h}) - \sin(h\omega n)\sin(\Phi_{i,h})) \end{aligned} \tag{1}$$

Where, ω is the known fundamental frequency, $a_{i,h}$ is the amplitude of harmonic component h in frame i, and $\Phi_{i,h}$ the associated phase.

Then, the DFT of this signal gives:

$$\begin{aligned} X_{i,k} &= \sum_{h=1}^{H} \sum_{n=0}^{N-1} a_{i,h} \cos(\Phi_{i,h}) \cos(h\omega n) e^{-\frac{2\pi i}{N} kn} \\ &\quad - \sum_{h=1}^{H} \sum_{n=0}^{N-1} a_{i,h} \sin(\Phi_{i,h}) \sin(h\omega n) e^{-\frac{2\pi i}{N} kn} \\ &= \sum_{h=1}^{H} a_{i,h} \cos(\Phi_{i,h}) \, \mathrm{DFT}(\cos(h\omega n))_k \\ &\quad - \sum_{h=1}^{H} a_{i,h} \sin(\Phi_{i,h}) \, \mathrm{DFT}(\sin(h\omega n))_k \end{aligned} \tag{2}$$

Let,

$$\begin{aligned} \mathbf{D}_C &= \begin{pmatrix} \mathrm{DFT}(\cos(\omega n))_0 & \dots & \mathrm{DFT}(\cos(H\omega n)))_0 \\ \vdots & \vdots & \vdots \\ \mathrm{DFT}(\cos(\omega n))_{N-1} & \dots & \mathrm{DFT}(\cos(H\omega n)))_{N-1} \end{pmatrix} \\ \mathbf{D}_S &= \begin{pmatrix} \mathrm{DFT}(\sin(\omega n))_0 & \dots & \mathrm{DFT}(\sin(H\omega n)))_0 \\ \vdots & \vdots & \vdots \\ \mathrm{DFT}(\sin(\omega n))_{N-1} & \dots & \mathrm{DFT}(\sin(H\omega n)))_{N-1} \end{pmatrix} \\ \mathbf{D} &= \begin{pmatrix} \mathbf{D}_C & -\mathbf{D}_S \end{pmatrix} \end{aligned} \tag{3}$$

and

$$\begin{aligned} \mathbf{B}_C &= \begin{pmatrix} a_{1,1}\cos(\Phi_{1,1}) & \dots & a_{F,1}\cos(\Phi_{F,1}) \\ \vdots & & \vdots \\ a_{1,H}\cos(\Phi_{1,H}) & \dots & a_{F,H}\cos(\Phi_{F,H}) \end{pmatrix} \\ \mathbf{B}_S &= \begin{pmatrix} a_{1,1}\sin(\Phi_{1,1}) & \dots & a_{F,1}\sin(\Phi_{F,1}) \\ \vdots & & \vdots \\ a_{1,H}\sin(\Phi_{1,H}) & \dots & a_{F,H}\sin(\Phi_{F,H}) \end{pmatrix} \\ \mathbf{B} &= \begin{pmatrix} \mathbf{B}_C \\ \mathbf{B}_S \end{pmatrix} \end{aligned} \tag{4}$$

Then, (2) can be regarded as column i of $\mathbf{X} = \mathbf{DB}$.

Considering the DFT of an observed signal, $\tilde{\mathbf{X}}$, as composed of a harmonic component plus some residual noise ε, there will be some $\mathbf{B}$ such that:

$$\tilde{\mathbf{X}} = \mathbf{DB} + \varepsilon \tag{5}$$

$\mathbf{B}$ can then be estimated by minimizing the square error $\|\varepsilon\|_2^2$. For real $\mathbf{B}$, we can separate the real and imaginary parts of the solution

Source Waveform	Normalised Mean Weights			
	Sine	**Sawtooth**	**Square**	**Triangle**
Sine	0.9999	0.0022	0.0009	-0.0030
Sawtooth	0.0002	0.9983	0.0033	-0.0018
Square	0.0016	0.0164	0.9843	-0.0023
Triangle	-0.0004	0.0028	-0.0015	0.9991

Table 1: Mean Weights for Basic Waveforms

$$\mathbf{X}_R + i\mathbf{X}_I = \mathbf{D}_R\mathbf{B} + i\mathbf{D}_I\mathbf{B} + \varepsilon \tag{6}$$

Then,

$$\begin{aligned} \|\varepsilon\|_2^2 &= \|\mathbf{X}_R + i\mathbf{X}_I - \mathbf{D}_R\mathbf{B} - i\mathbf{D}_I\mathbf{B}\|_2^2 \\ &= \|\mathbf{X}_R - \mathbf{D}_R\mathbf{B}\|_2^2 + \|\mathbf{X}_I - \mathbf{D}_I\mathbf{B}\|_2^2 \end{aligned} \tag{7}$$

To minimize the squared error, set the derivative of $\|\varepsilon\|_2^2$ to zero:

$$\frac{\partial\|\varepsilon\|_2^2}{\partial\mathbf{B}} = \mathbf{D}_R^T(\mathbf{X}_R - \mathbf{D}_R\mathbf{B}) + \mathbf{D}_I^T(\mathbf{X}_I - \mathbf{D}_I\mathbf{B}) \tag{8}$$

This is zero when,

$$\begin{aligned} \mathbf{D}_R^T\mathbf{X}_R + \mathbf{D}_I^T\mathbf{X}_I &= \mathbf{D}_R^T\mathbf{D}_R\mathbf{B} + \mathbf{D}_I^T\mathbf{D}_I\mathbf{B} \\ &= (\mathbf{D}_R^T\mathbf{D}_R + \mathbf{D}_I^T\mathbf{D}_I)\mathbf{B} \\ &= \mathbf{Y}\mathbf{B} \end{aligned} \tag{9}$$

Where $\mathbf{Y} = (\mathbf{D}_R^T\mathbf{D}_R + \mathbf{D}_I^T\mathbf{D}_I)$.

Hence $\mathbf{B}$,

$$\mathbf{B} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T(\mathbf{D}_R^T\mathbf{X}_R + \mathbf{D}_I^T\mathbf{X}_I) \tag{10}$$

Given $\mathbf{B}$, $a_{i,h}$ and $\Phi_{i,h}$ follow from (4).

3.3. Expressing harmonic structure in terms of basic waveforms

We consider the harmonic structure of the signal as a linear combination of the basic waveforms:

$$\tilde{\mathbf{H}} = \mathbf{S}^T\mathbf{W} + \mathbf{E} \tag{11}$$

Where $\tilde{\mathbf{H}}$ is the observed harmonic structure ($H \times F$), $\mathbf{S}$ the structure of the basic wave forms ($n \times H$), $\mathbf{W}$ the proportions of each waveform in the structure ($n \times F$), and $\mathbf{E}$ is an error term ($H \times F$), and n the number of basic waveforms.

Using the voices from 3.1, the structure of $\mathbf{S}$ is:

$$\begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \dots & w_h \\ \frac{1}{1} & 0 & \frac{1}{3} & \dots & q_h \\ \frac{1}{1} & 0 & \frac{1}{9} & \dots & t_h \end{pmatrix} \tag{12}$$

Minimizing the square error estimates $\mathbf{W}$ for the signal:

$$\mathbf{W} = (\mathbf{S}\mathbf{S}^T)^{-1}\mathbf{S}\tilde{\mathbf{H}} \tag{13}$$

Hence, we represent the harmonic part of the signal as a linear combination of the outputs of our basic waveform voices.

4. RESULTS

4.1. Generated basic signals

A signal consisting of 0.5 second 16bit/44100kHz samples of each of the basic waveforms was generated and processed as above using a 2048 bin FFT. The columns of $\mathbf{W}$ were then normalised (L_1 norm) to give the relative weights for each waveform in each frame. The means of these weights over time are given in Table 1.

For a correct waveforms, the maximum variance of the relative weights was 0.05 (the triangle wave), and the variances were less than 10^{-3} for the alternative waveforms. The waveforms were all recognized correctly, and variances in the proportions were mainly due to frames which contained data from more than one waveform (i.e. the frames at the 0.5 second boundaries). Subsequent experiments Hence, it should be possible to identify basic waveforms within a signal, and to extend their harmonic range - generating the upper harmonics for a new sample rate from the known harmonic structure.

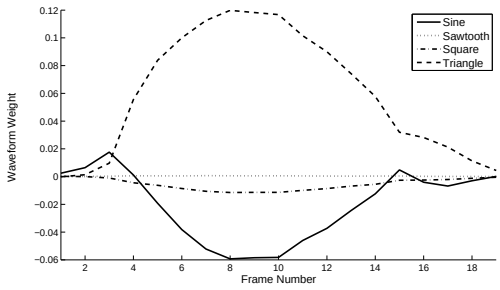


Fig. 5: Waveform weights for Clarinet note

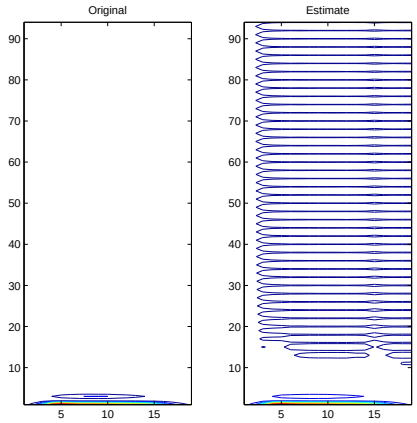


Fig. 6: Original and estimated harmonic structure of Clarinet note

4.2. Clarinet sample

A clarinet signal (Clarinet1) from the Leveau onset detection database [9] was also processed, using the Leveau onset times and a 2048 bin FFT. The harmonic structure was then estimated using as $\mathbf{S}^T\mathbf{W}$.

Considering a typical note from the signal, figure 4.2 shows the variation of the weights of the four waveforms across the FFT frames. This variation appears quite smooth, and could be interpolated using, for example, cubic splines. The variation could then be reproduced at a higher time resolution than the input signal.

Figure 4.2 compares the original harmonic data with the reconstructed form showing that the reconstructed harmonic structure contains levels of the upper harmonics which were not present in the original signal. This is confirmed by informal listening tests in which the reconstructed signal is significantly brighter than the source material.

5. CONCLUSION

Object-coding represents an audio signal as the output of multiple voice objects, with a set of parameters inferred from an original source signal. By associating a residual signal with the modelled data, the source signal can be coded in a lossless manner. Selecting continuous-time models for the voices, the output resolution of the decoder can be chosen as required.

To upsample 16bit/44.1kHz legacy audio, the signal can be encoded as voice parameters and a 16bit/44.1kHz residual signal. A high-resolution version of the signal can then be generated at 32bit/192kHz from the objects and this can be combined with either an upsampled version of the residual. Thus, the full available resolution can be used to extend both the dynamic-range and frequency-bandwidth of the modelled signal. This is made feasible by the encoder, and once the initial encoding process has been completed, bandwidth extension appropriate to any output sample-rate is possible. Assuming appropriate voice models, this bandwidth extension will occur in a physically relevant manner adding relevant upper harmonics to each object in the signal. The bandwidth of the residual element may also be extended by inferring noise parameters and resynthesizing suitable noise from those parameters.

However, the simple harmonic waveforms used in this paper were not suitable for modelling the higher harmonics of the signals. Two fundamental aspects are apparent: the harmonic models chosen and the inference techniques used. Advances in machine-learning, audio modelling, computer power allow both of these to be investigated.

The challenges involved in object-coding are great. It requires not only the development of effective voice models but also algorithms to infer which objects are present in the audio signal and their parameters. Dependencies on ongoing research areas (onset detection, pitch tracking, source separation, instrument identification) position it's general use firmly in the future. Nevertheless, we believe that object-coding will allow audio to be encoded in a flexible and scalable format, promising a content-sensitive approach to upsampling legacy audio for use in high-resolution environments.

6. ACKNOWLEDGEMENTS

This work has been made possible by an EPSRC Research Studentship at Queen Mary, University of London.

7. REFERENCES

[1] ISO/IEC JTC 1/SC 29. *13818-3:1998 : Information technology - generic coding of moving pictures and associated audio, part 3: Audio*. MPEG-1 Layer 3 (Audio), MP3. ISO/IEC International Standard, 1998.

[2] ISO/IEC JTC 1/SC 29. *14496-3:2005 Coding of audio-visual objects, part 3: Audio*. MPEG-4 Audio. ISO/IEC International Standard, 2005.

[3] D. Bansal, B. Raj, and P. Smaragdis. Bandwidth Expansion of Narrowband Speech Using Non-Negative Matrix Factorization. In *Eurospeech 2005*, September 2005.

[4] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5), 2005.

[5] K. Brandenburg. MP3 and AAC explained. In *Proceedings of AES 17th International Conference on High-Quality Audio Coding*. AES, 1999.

[6] Nick Collins. A comparison of sound onset detection algorithms with emphasis on psychoacoustically motivated detection functions. In *Proceedings of AES 118th Convention*. AES, 2005.

[7] P. Gerken. Object-based analysis-synthesis coding of image sequences at very low bit rates. *IEEE Transactions on Circuits and Systems for Video Technology*, 4(3):228–235, 1994.

[8] M. Hans and R. W. Schafer. Lossless compression of digital audio. *Signal Processing Magazine, IEEE*, 18(4):21–32, 2001.

[9] Pierre Leveau, Laurent Daudet, and Gael Richard. Methodology and tools for the evaluation of automatic onset detection algoritms in music. In *Proceedings of ISMIR 2004, the 5th International Conference on Music Information Retrieval*, pages 72–75. Universitat Pompeu Fabra, 2004.

[10] David MacKay. *Information Theory, Inference and Learning Algorithms*, chapter 5, pages 98–101. Cambridge University Press, 2003.

[11] Asif Masood and Shaiq A. Haq. Object coding for real time image processing applications. In *Pattern Recognition and Image Analysis*, volume 3687/2005 of *Lecture Notes in Computer Science*, pages 550–559. Springer, Berlin / Heidelberg, 2005.

[12] H. Purnhagen, B. Edler, and C. Ferekidis. Object-based analysis/synthesis audio coder for very low-bit rates. In *Proceedings of AES 104th Convention*. AES, 1998.

[13] H. Purnhagen, N. Meine, and B. Edler. Speeding up HILN-MPEG-4 parametric audio encoding with reduced complexity. In *Proceedings of AES 109th Convention*. AES, 2000.

[14] T. Robinson. SHORTEN: Simple lossless and near-lossless waveform compression. Technical Report CUED/F-INFENG/TR.156, Cambridge University Engineering Department, 1994.

[15] M. Sarfraz and M. A. Khan. An automatic algorithm for approximating boundary of bitmap characters; computer graphics and geometric modeling. *Future Generation Computer Systems*, 20(8):1327–1336, 2004.

[16] Eric D. Scheirer. Structured audio and effects processing in the mpeg-4 multimedia standard. *Multimedia Systems*, (7):11–22, 1999.

[17] Tero Tolonen. Object-based sound source modeling for musical signals. In *Proceedings of 109th AES Convention - SESSION A: DIGITAL SIGNAL PROCESSING, PART 1*, Los Angeles, USA, 2000. AES.

[18] BL Vercoe, WG Gardner, and ED Scheirer. Structured audio: creation, transmission, and rendering ofparametric sound representations. *Proceedings of the IEEE*, 86(5):922–940, 1998.

[19] E. Vincent and M.D. Plumbley. Low Bitrate Object Coding of Musical Audio using Bayesian Harmonic Models. Technical Report C4DM-TR-06-01, Centre for Digital Music, Queen Mary, University of London, 2006.