AUDIO DATA HIDING WITH HIGH DATA RATES BASED ON INTMDCT

Ralf Geiger, Yoshikazu Yokotani*, and Gerald Schuller***

*Fraunhofer IIS, Erlangen, Germany
**Fraunhofer IDMT and Ilmenau University of Technology, Ilmenau, Germany
Email: [ggr,yki]@iis.fraunhofer.de, shl@idmt.fraunhofer.de

ABSTRACT

This paper describes high data-rate audio data hiding using the IntMDCT. The IntMDCT is an integer approximation of the MDCT with perfect reconstruction. Based on this transform, we describe a straight-forward way to embed data and extract it in a bit-exact manner while perceptual transparency is maintained. Since the IntMDCT spectrum can be used to obtain a closer approximation of the masking threshold compared to time domain approaches, it is possible to achieve higher data rates. In a simple experimental implementation, we found we could embed data at rates up to about 140kb/s without introducing audible distortions.

1. INTRODUCTION

Modern perceptual audio coding schemes, such as MPEG-4 AAC [1], allow for a high compression of audio signals while maintaining high audio quality. For example, the data rate can be reduced from 1.44 Mbit/s (16 bit/sample, 44.1 kHz, 2 channels) to 128 kbit/s. This motivates the approach presented in this paper: The perceptual irrelevancy is exploited to hide data in the original audio signal, in a way that it is inaudible to the human listener.

Several approaches have been proposed for embedding additional data in audio signals.

In [2], an audio watermarking scheme is presented, which shapes the embedding data energy according to the masking threshold. However, this approach is designed to have robustness against attacks, and hence the typical data rate is low, about 1kbit/sec.

In [3], a high-rate buried data channel for the Audio CD is described. In this approach, the additional data is embedded in the time domain as perceptually noise-shaped subtractive dither. On the other hand, compared to such a time domain approach, a transform domain approach can achieve a higher data rate since a closer approximation of the masking threshold is obtained in the transform domain.

As a transform domain approach, in [4], a filter bank and a perceptual model are used to embed additional data to audio signals in an inaudible way. A quantization according to the perceptual model is applied, and the quantization interval is used for adding the additional data. The data is only recovered almost exactly, and hence, it has no bit-exact reconstruction.

In [5], it is proposed to use the wavelet transform. However, the transform used there is a conventional floating-point transform, and thus the subband signals are not integers. As a result, it is difficult to obtain a bit-exact reconstruction. A rounding operation is required after the inverse transform. This could potentially destroy a part of the embedded data. For the case of images, this issue is covered in [6] by using the integer wavelet transform. Such integer transforms allow to embed data on designated LSB values in a straight forward way without the need for an additional rounding stage. This approach will be applied to audio here.

In this paper, we use the Integer Modified Discrete Cosine Transform (IntMDCT) [7] to construct an audio data hiding scheme with a high data rate. The hidden data can be extracted in a bit-exact manner. In addition, since the spectrum can be used to obtain a closer approximation of the masking threshold, it is possible to achieve high data rate embedding in a very straight-forward way while the perceptual transparency is maintained.

The IntMDCT is an integer approximation of the MDCT, which is widely used in state-of-the-art perceptual audio coding schemes, such as MPEG-4 AAC [1]. Thus, choosing the IntMDCT seems a good choice for obtaining an efficient and perceptually transparent audio data hiding scheme. The IntMDCT is also used as a core part of the scalable lossless extension (MPEG-4 SLS) of MPEG-4 AAC.

For applications where embedding at a high data rate into an audio signal is important and a limited complexity and inaudibility of the inserted data is desired, the proposed approach is very useful.

2. THE INTMDCT

Integer transforms like the IntMDCT map integer time-domain input signals to integer subband signals. At the same time they are reversible. Hence, they are a lossless process over the forward and inverse transforms. Transforms like the DFT, DCT, or the MDCT can be decomposed into so-called Givens rotations. Each Givens rotation can in turn be decomposed into lifting steps as follows

$$\begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} = \begin{pmatrix} 1 & \frac{\cos\alpha-1}{\sin\alpha} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \sin\alpha & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{\cos\alpha-1}{\sin\alpha} \\ 0 & 1 \end{pmatrix}$$

Figure 1 illustrates this decomposition (see also [7]). Lifting steps have the advantage that rounding can be included while maintaining the reversibility of each lifting step. The rounding is applied after each multiplication with a factor in Fig. 1. Hence we obtain a reversible integer to integer mapping.

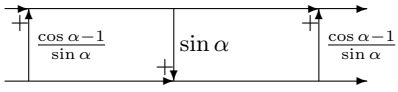

Fig. 1. Givens rotation by three lifting steps

A related approach is to use multi-dimensional lifting [8],[9]. It has the advantage of a lower complexity and of less rounding errors.

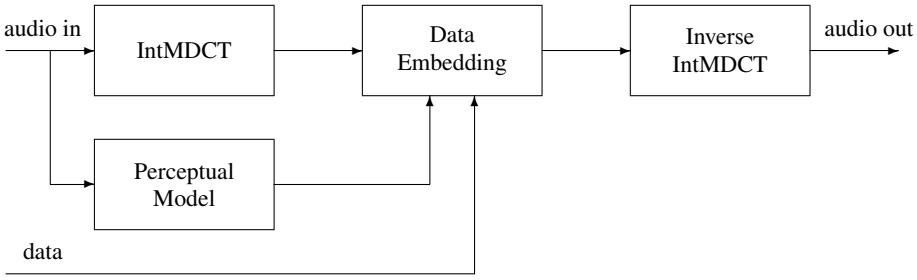

Fig. 2. Embedding algorithm for data hiding using IntMDCT

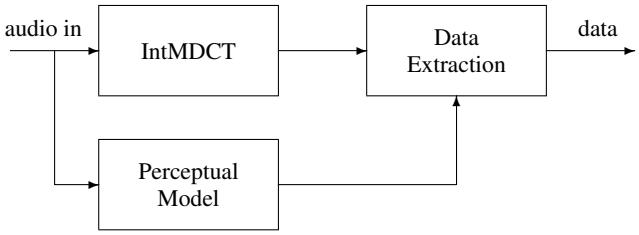

Fig. 3. Extraction algorithm for data hiding using IntMDCT

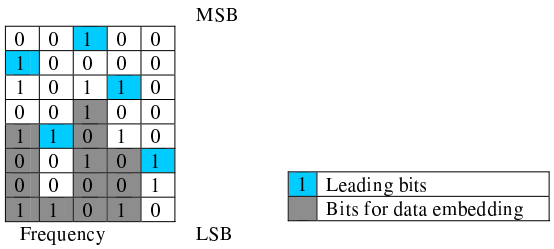

Fig. 4. Data embedding with simple perceptual model (3 bits significant)

3. BASIC PRINCIPLE

The data embedding is performed in frequency domain, allowing the algorithm to exploit perceptual masking phenomena in order to embed data inaudibly at a high data rate. The basic principle of the embedding algorithm is shown in Figure 2. The frequency representation is obtained by using the IntMDCT, representing the audio signal by integer spectral values in an invertible way. A perceptual model determines to which extent each integer spectral value is perceptually significant, e.g. by dividing the binary representation of the absolute value into significant and insignificant bits. The insignificant part can be used for data embedding. Finally, the inverse IntMDCT is applied in order to obtain the audio signal which contains the embedded data. The advantage of this approach is that, even after a modification of the integer spectral values, a transformation to integer audio samples is possible without the need for a lossy rounding operation. Hence the modified integer spectral values can exactly be reconstructed again from the integer samples.

Figure 3 shows the corresponding data extraction algorithm. The embedded data is extracted from the audio signal by using the same transform and the same perceptual model as for the embedding process. The embedded data can be retrieved from the insignificant part of the integer spectral values. Here a constraint for the perceptual model has to be considered: The determined perceptual significance should not change after replacing the insignificant part of the integer spectral values by the embedding data.

4. EMBEDDING USING SIMPLE PERCEPTUAL MODEL

For the simplest version of the embedding algorithm a constant frame length is chosen (e.g. 256 or 512 spectral values). This allows for a compromise between a good spectral resolution for tonal signals and a good temporal resolution for transient signals. Furthermore, a simple perceptual model is used, demanding a fixed signal to noise ratio for each spectral value. This is achieved by considering the binary representation of each absolute spectral value, and declaring the highest non-zero bit ("leading bit") and a fixed number of lower bits as significant. For example, a signal to noise ratio of about 18 dB is achieved by declaring 3 bits as significant. All lower bits of the absolute integer spectral values are considered as perceptually insignificant and can be utilized for data embedding. This is illustrated in Figure 4.

5. FRAMING DETECTION

For practical purposes it is desirable to start extracting the embedded data at arbitrary samples of the audio signal. For example, it should be possible to truncate the audio signal, to start extracting during an ongoing transmission or to restart the extraction after skipping a part of the audio signal. Hence a mechanism for finding the frame grid of the embedded data is necessary. This can be achieved by adding a small amount of redundancy (e.g. a check sum) to the embedded data. In order to find the correct framing offset, several frames of audio samples have to be transformed repeatedly, with the offset incremented by one sample at each time. By evaluating the check sum of the supposed embedded data, the framing grid can easily be determined. While a false detection of one frame is not fully excluded by using only a small amount of redundancy, a regular occurrence of false detection within the framing grid gets more unlikely, the more frames of audio samples are examined. Hence a small amount of redundancy is sufficient for a reliable detection of the framing grid. Once the framing grid is found, the extraction algorithm can be performed frame by frame.

6. FIRST RESULTS

This simple embedding scheme was evaluated for different frame lengths and for different numbers of significant bits. In this experiment the insignificant bits were not replaced by embedding data, but by a bit mask (1,0,0,...), resulting in a mid-rise quantizer. The number of bits replaced was used to calculate the average bit rate available for embedding.

The perceived audio quality after data inserting is evaluated based on the ITU-R BS.1387 PEAQ (Perceptual Evaluation of Audio Quality) measurement method [10] using the Noise- to-Mask Ratio (NMR) and the Objective Difference Grade (ODG) output values which are briefly described here for better understanding.

Noise-To-Mask Ratio (NMR)

The NMR estimates the ratio between the actual distortion ("Noise") and the maximum inaudible distortion, i.e. the masking threshold ("Mask"). NMR values smaller than 0dB indicate the headroom between the noise and the threshold of audibility whereas values larger than 0dB indicate audible distortions.

Objective Difference Grade (ODG)

The ODG values are designed to mimic the listening test ratings obtained from typical test listeners by means of an objective measurement procedure. The grading scale ranges from -4 ("very annoying") to 0 ("imperceptible difference").

Table 1 presents the resulting bit rates, Objective Difference Grades (ODG) and Worst NMR values for different frame lengths and different numbers of significant bits. The tested audio signals represent the test set of twelve critical items used in the development of MPEG-4 [1] perceptual audio codecs. They contain both very tonal items (e.g. pitch pipe) critical for short transform lengths, and very transient signals (e.g. castanets) critical for long transform lengths. It can be observed that the best compromise is obtained for frame lengths of 256 or 512. Especially for the frame length of 256 a value of 4 significant bits already results in an ODG value of 0.0 and a negative worst NMR, indicating that the signal modification is not audible. It allows an embedding bit rate of 142 kBit/s.

frame length	Significant bits	Embedding bitrate	ODG	Worst NMR
128	2	259 kBit/s	-1.6	6.7 dB
128	3	204 kBit/s	-0.7	-0.2 dB
128	4	157 kBit/s	-0.2	-7.0 dB
128	5	117 kBit/s	0.0	-13.3 dB
128	6	84 kBit/s	-0.1	-19.0 dB
128	7	58 kBit/s	0.0	-19.1 dB
128	8	38 kBit/s	0.0	-19.0 dB
256	2	241 kBit/s	-1.0	0.9 dB
256	3	187 kBit/s	-0.3	-6.1 dB
256	4	142 kBit/s	0.0	-14.6 dB
256	5	105 kBit/s	0.0	-18.0 dB
256	6	75 kBit/s	0.0	-19.1 dB
256	7	51 kBit/s	0.0	-19.4 dB
256	8	33 kBit/s	0.0	-19.9 dB
512	2	225 kBit/s	-0.8	12.9 dB
512	3	172 kBit/s	-0.1	5.8 dB
512	4	128 kBit/s	0.0	0.1 dB
512	5	94 kBit/s	0.0	-7.6 dB
512	6	66 kBit/s	0.0	-14.9 dB
512	7	45 kBit/s	0.0	-19.2 dB
512	8	29 kBit/s	0.0	-19.7 dB
1024	2	213 kBit/s	-1.5	25.5 dB
1024	3	161 kBit/s	-0.5	19.8 dB
1024	4	119 kBit/s	-0.2	14.9 dB
1024	5	86 kBit/s	-0.2	7.9 dB
1024	6	60 kBit/s	-0.3	1.5 dB
1024	7	41 kBit/s	0.0	-4.0 dB
1024	8	26 kBit/s	0.0	-9.7 dB

Table 1. Embedding bit rate and perceptual quality for different frame lengths and significant bits

7. ADVANCED PERCEPTUAL MODEL AND BLOCK SWITCHING

In the embedding algorithm presented so far the perceptual model is very simple and demands a constant signal to noise ratio for each spectral value. Nevertheless, a more signal adaptive perceptual model can also be used, considering e.g. the different demands of tonal and non-tonal maskers and masking across spectral values. It only has to be assured that the discrimination between significant and insignificant part of the integer spectral values operates still in the same way after the insignificant part is replaced by the embedded data. This can, for example, be achieved by taking into account only the leading bit of each spectral value in the perceptual model and by assuring that the leading bits are not modified. In case of smaller, masked spectral values, more bits can be embedded in this way. This is illustrated in Figure 5.

With the approach presented so far, the maximum possible bitrate of the embedded signal highly depends on the level of the audio signal. Especially in quiet parts of the audio signal the bitrate can even decrease to zero. To avoid this and to assure a certain minimum bitrate, a fixed threshold in quiet can be considered additionally, allowing to embed at least some data especially in the high and low frequency range.

So far only a fixed transform length has been discussed, forcing a compromise between tonal and transient signal portions. Never-

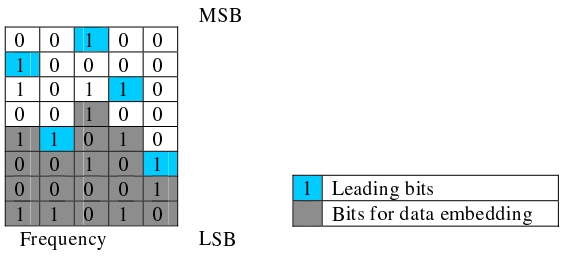

Fig. 5. Data embedding with perceptual model based on leading bits

theless, an adaptive block length, as for perceptual audio coding, can also be considered in this context. To enable this, the extractor has to know the current block length without retrieving any additional side information. This can be achieved by evaluating the redundancy in the embedded data necessary for the framing detection. The extractor just has to try several transforms with different possible block lengths and window shapes. The correct choice can then be found by evaluating the check sums.

8. APPLICATIONS

The embedding technique presented in this section is generally interesting in the context of applications where an unmodified transmission of the digital audio samples can be assumed. In such environments additional data with high data rates can be embedded without loss of audio quality.

The Audio CD is an example system where this technique could be included. While maintaining full backwards compatibility, the additional data can be extracted from the digital audio stream, which is usually available via the SPDIF connection in consumer devices.

The technique of finding the framing grid in the embedded data also allows to proceed with extracting the embedded data after a part of the playback is skipped.

The channel coding used on the Audio CD allows correction of a certain amount of errors occurring in the raw data. However, if the error rate is increased, some errors can not be corrected and an error concealment technique is applied, trying to make the error inaudible. In the data embedding technique put forward here, this concealment can destroy the embedded data. Nevertheless, this can be counteracted by adding an error correcting channel coding technique, which is adapted to the error characteristics of the Audio CD, to the embedded data stream.

Examples for additional data, which could be embedded with this technique, are:

- Video data
  By encoding additional video data with a modern, highly efficient video codec, such as MPEG-4 AVC [11] a video stream with reasonable quality could be embedded in the audio signal.
- Compressed audio data
  With the embedding technique presented here a compressed version of the original audio signal could also be embedded in the audio signal itself. In this way a compressed version of the audio signal could already be produced and embedded during the production process of the Audio CD. Hence the compressed audio signal just has to be extracted from the uncompressed audio signal rather than encoding it with higher computational cost.
- Spatial sound information
  The embedding technique can be used to add spatial cue information to the audio signal. According to [12] a rather small amount of information is necessary to extend an audio signal to a 5.1 channel representation. If these spatial cues are embedded in the stereo audio signal, a device capable of extracting this information can output a 5.1 signal while a stereo playback device still operates in the stereo mode.

9. CONCLUSIONS

It was shown that the IntMDCT is a simple and efficient tool to insert data into an audio stream. First results show that, with a relatively simple system, an embedded data rate of 142 kb/s is possible without audible changes to the audio signal. The audibility was measured with the PEAQ audio quality measurement tool. Even higher data rates can be expected by using a more advanced perceptual model.

10. REFERENCES

[1] "Information technology - Coding of audio-visual objects - Part 3: Audio," International Standard 14496-3:2001, ISO/IEC JTC1/SC29/WG11 Moving Pictures Expert Group, 2001.

[2] F. Siebenhaar, C. Neubauer, R. Bäuml, and J. Herre, "New High Data Rate Audio Watermarking based on SCS (Scalar Costa Scheme)," in *113th Convention of the AES*, Los Angeles, USA, October 5-8 2002, preprint 5645.

[3] A. W. J. Oomen, M. E. Groenewegen, R. G. van der Waal, and R. N. J. Veldhuis, "A Variable-Bit-Rate Buried-Data Channel for Compact Disc," *J. Audio Eng. Soc.*, vol. 43, no. 1/2, pp. 23–28, January/February 1995.

[4] W. R. Th. ten Kate, L. M. van de Kerkhof, and F. F. M. Zijderveld, "A New Surround-Stereo-Surround Coding Technique," *J. Audio Eng. Soc.*, vol. 40, no. 5, pp. 376–383, May 1992.

[5] J. Chou, K. Ramchandran, and A. Ortega, "High Capacity Audio Data Hiding for Noisy Channels," in *Proc. IEEE Information Technolgoy Coding and Computing*, April 2001.

[6] G. Xuan, J. Chen, J. Zhu, Y. Q. Shi, Z. Ni, and W. Su, "Lossless Data Hiding Based on Integer Wavelet Transform," in *IEEE 5th Workshop on Multimedia Signal Processing*, St. Thomas, US Virgin Islands, December 9-11 2002.

[7] R. Geiger, T. Sporer, J. Koller, and K. Brandenburg, "Audio Coding based on Integer Transforms," in *111th AES Convention*, New York, 2001.

[8] R. Geiger, Y. Yokotani, and G. Schuller, "Improved integer transforms for lossless audio coding," in *Proc. of the Asilomar Conf. on Signals, Systems, and Computers*, 2003.

[9] R. Geiger, Y. Yokotani, G. Schuller, and J. Herre, "Improved Integer Transforms using Multi-Dimensional Lifting," in *Proc. of the ICASSP*, Montreal, Canada, May 17-21 2004.

[10] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerens, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality," *Journal of the AES*, vol. 48(1/2), pp. 3–29, 2000.

[11] "Advanced Video Coding," International Standard 14496-10, ISO/IEC JTC1/SC29/WG11 Moving Pictures Expert Group, 2003.

[12] C. Faller and F. Baumgarte, "Binaural Cue Coding Applied to Stereo and Multi-Channel Audio Compression," in *112th Convention of the AES*, Munich, Germany, May 10-13 2002, preprint 5574.