nederlands akoestisch genootschap
NAG

journaal

nr. 184
november 2007

Efficiënte audiocompressie gebaseerd op de perceptieve codering van ruimtelijk geluid

Jeroen Breebaart, Armin Kohlrausch, Steven van de Par – Philips Research
High Tech Campus 36 M/S2 5656 AE Eindhoven

Efficient audio compression based on the perceptual coding of spatial audio

Abstract

Digital compression of audio signals is one of the technological fields that has always been tightly linked to detailed knowledge on human perception. Recent advances have extended the sources for bit-rate reduction from a predominantly masking-oriented approach towards additional exploitation of spatial perceptual irrelevancies. New techniques aim at modeling of perceptually-relevant spatial cues using a parametric approach. These parametric methods result in considerable improvements in compression efficiency for stereo and multi-channel audio coders and have consequently been adopted by standardization bodies such as MPEG (standard for audio and video compression) and 3GPP (standard for mobile transmission). Besides compression efficiency, the parametric approach has shown to be applicable to spatial processing methods as well, such as rendering of virtual auditory scenes over headphones or to provide user interactivity with audio content, such as the freedom to reposition, amplify or equalize individual objects in a complex auditory scene. In this paper, the perceptual basis and methods for parametric spatial techniques will be outlined and examples of the resulting compression and processing technologies will be given.

1. Introduction

Current trends in consumer audio show two important developments. Firstly, there is a shift from conventional stereo to multi-channel audio. Initially this shift was predominantly observed in the movie domain (for example by the introduction of the DVD). More recently, pure audio is also available in multi-channel format on SACD and DVD audio. A second trend that can be identified is the increased popularity of mobile audio. Interestingly, recent models of mobile audio players do not only have audio playback capability, but also provide video playback. A logical consequence is that mobile audio/video players will become multi-channel capable as well.

These trends impose new challenges on audio compression, broadcasting and processing schemes. For example, several audio codecs that are being used are simply not capable of handling multi-channel audio (such as MPEG-1 layer 3) [1]. Secondly, if an audio codec is capable of handling more than two audio channels, the increase in required bit rate scales linearly with the number of audio channels which is in many cases undesirable or even impossible because the required bandwidth is not available. A third important aspect of multi-channel audio codecs is their ability to provide backward-compatible bit streams. For example, if a radio or television broadcast system is upgraded from stereo to multi-channel

audio, it is essential that existing receivers that expect stereo content will still operate as normal. Unfortunately, such backward compatibility often results in a significant additional bit-rate penalty compared to multi-channel codecs without backward compatibility. Last but not least, mobile devices often have a rather limited processing power and battery life. This property makes it quite difficult to introduce multi-channel audio on a mobile device. In most cases, the complexity of audio decoders scales approximately linearly with the number of channels. Moreover, assuming that most consumers will use headphones on a mobile device, conventional methods to create a convincing and realistic virtual multi-channel setup are often quite CPU intensive.

From a perceptual point of view, the various channels in a multi-channel signal may sound very much alike. Interestingly, conventional audio coders only have a limited repertoire to gain efficiency from such perceptual similarity. Besides mid/side transforms [2] to remove potential inter-channel redundancies, the only method to exploit perceptual irrelevancies in the spatial domain is referred to as 'intensity stereo' [3]. Very recently, however, new techniques have been developed to describe spatial audio properties in a perceptually-motivated space rather than a multi-dimensional signal space. These techniques, often referred to as 'spatial audio coding' and 'binaural cue coding' [4, 5, 6, 7] provide several benefits in the field of multi-channel audio compression, broadcasting and rendering. More specifically, as will be outlined in the following sections, spatial audio coding techniques provide unsurpassed compression efficiency and backward compatibility for stereo and multi-channel audio codecs, and enable new and efficient methods for spatial audio rendering.

2. Spatial parameterization

Spatial audio coding employs a perceptually-motivated parameterization of spatial attributes. Instead of describing the spatial sound field by means of two or more *signals* that correspond to separate recording or reproduction positions, a spatial sound field is characterized by a very limited set of audio signals (typically one or two), accompanied by parameters that describe the perceptually-relevant aspects of the stereo or multi-channel content (see [8] for a concise overview). These parameters relate to sound-source localization attributes (such as inter-aural level differences, or ILDs, and inter-aural time differences, or ITDs) as well as spatial 'quality' attributes such as the perceived 'width' (which is closely related to the inter-aural coherence, or IC of the signals arriving at the eardrums) [9].

Due to the fact that sound source positions may change over time, and the fact that various sound sources from various positions may radiate sound simultaneously, the ILD, ITD and IC parameters of the signals at the level of the eardrums typically vary as a function of time and frequency.

Fortunately, the parametric representation of spectro-temporal variations in ILD, ITD and IC can be sampled on a relatively coarse time-frequency grid, since it is well known that the resolution of the human auditory system is limited in time and frequency for its ability to analyze these attributes. The limited temporal resolution is often referred to as 'binaural sluggishness' [10] while the dominance of a signal onset in a reverberant environment is termed the 'precedence effect' [11]. Given the different temporal characteristic in these two cases, the temporal analysis of spatial parameters should preferably be signal dependent. The spectral resolution is limited according to the concept of 'critical bands', which hence

requires a parameterization in non-linearly spaced frequency bands with a bandwidth of approximately 1/3 octave [12].

3. Spatial audio coding approach

The approach pursued with spatial audio coding and processing comprises separate encoding and decoding steps. The encoding process is visualized in Figure 1. A set of input signals is decomposed in separate time/frequency tiles by a time/frequency transform with variable temporal segmentation using a signal-dependent process [13]. The resulting time/frequency tiles of each signal are sent to a second stage that generates a down mix and extracts spatial parameters. Finally, the down-mix signal is processed by an inverse transform to result in a time-domain down-mix signal.

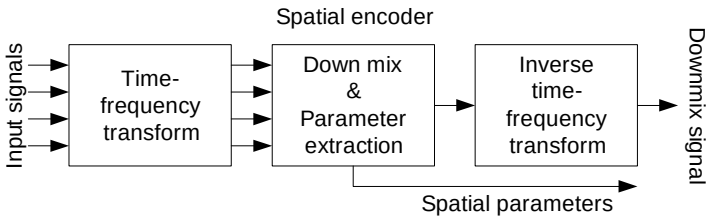

Figure 1: Spatial encoder comprising a time-frequency transform, a down mix and parameter extraction stage, and inverse transform.

The corresponding decoder is shown in Figure 2. The down-mix signal is decomposed into separate time/frequency tiles. Subsequently, a spatial synthesis stage super-imposes the spatial parameters onto the down-mix signal to reconstruct a set of output signals. Finally, an inverse transform results in the output signals.

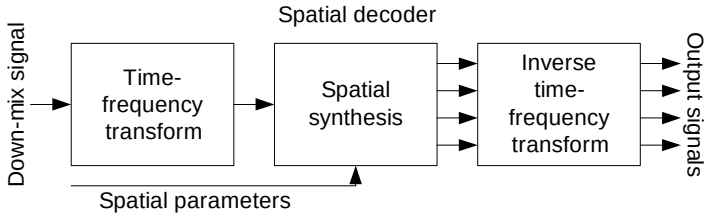

Figure 2: Spatial decoder that reconstructs spatial attributes in separate time/frequency tiles.

The process of spatial encoding and decoding has several benefits in the field of audio compression and processing. A couple of interesting applications will be described in the following sections.

4. Examples

4.1 Parametric Stereo

Parametric Stereo (PS) is the first employment of spatial audio coding technology in international standards and commercially available audio codecs. Within MPEG-4 and 3GPP, PS is supported in aacPlus v2, a codec based on (mono) AAC [1], spectral band replication (SBR) [14] and PS [8]. This codec is regarded as the most efficient (stereo) audio coder available today, delivering ‘good’ quality at bit rates as low as 24-32 kbps, and ‘excellent’ quality around 48 kbps. The incorporation of Parametric Stereo in aacPlus v2 is

shown in Figure 3. The stereo input signal is first analyzed by a spatial encoder generating a mono down mix and spatial parameters. The mono down mix is subsequently encoded using an aacPlus coder. The resulting bit stream is combined with the spatial parameters in a multiplexer to generate the output bit stream.

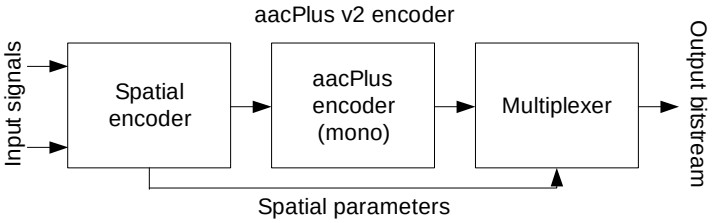

Figure 3: Outline of an aacPlus v2 encoder

The aacPlus v2 decoder basically performs the inverse process (not shown) of demultiplexing, mono decoding, and spatial decoding. The spatial decoding process involves re-instating the extracted parameters onto the mono down mix. This process involves scaling, phase modification and 'decorrelation' [8] in various frequency bands. The compression gain of Parametric Stereo in this coder has been shown to amount to approximately 33% [8] at low bit rates (24-32 kbps). One of the reasons for aacPlus v2 to be extremely efficient is the fact that the amount of bits required for the spatial parameters amounts to only 1 to at most 8 kbps, which is about one magnitude less than the bit rate required to transmit another audio channel.

4.2 MPEG Surround

MPEG Surround [15, 17] is the most recently standardized audio coder employing spatial audio coding technology. It extends the spatial audio coding approach to multi-channel audio. The approach is quite similar to the approach outlined for Parametric Stereo, with the exception that the input, down-mix and output channel configurations can all be chosen independently and may comprise more than two channels. The supported input channel configurations vary from standard 5.1-channel audio to more exotic channel configurations with for example 10 loudspeakers. The down-mix may be mono, stereo, matrixed-surround compatible stereo (to facilitate compatibility with matrixed-surround coders), or even 5.1 if for example a 7.1 input signal is employed. The decoder channel configuration can be stereo, binaural stereo (for headphone playback), 5.1, 7.1, etc.

To facilitate such flexibility, spatial coding in MPEG Surround is based on a set of elementary 'building blocks'. These building blocks come in different flavours (see [15] for a detailed overview). The most important building blocks are spatial encoding and decoding blocks, as outlined in the left and right panel of Figure 4, respectively. A spatial encoding block may have 2 or 3 input signals, and has 1 or 2 output signals, and parameters. The spatial decoding block basically comprises the inverse process. A spatial encoding block with 2 channels as input and one mono output is essentially equivalent to a parametric stereo encoder. The three-channel encoding block is a combination of perceptual parameterization and signal prediction methods (see [15, 16]).

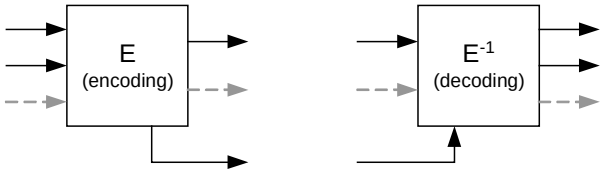

Figure 4: Spatial encoding and decoding block as present in MPEG Surround.

The building blocks can be connected to build an arbitrary encoder or decoder *tree*. An example for a 5.1 input and a stereo down mix is provided in Figure 5. The 6 input channels (L_f, L_s, R_f, R_s, C and LFE, for the left-front, right-front, left-surround, right-surround, center and low-frequency effects channels, respectively) are stepwise reduced to a stereo down-mix (S_L, S_R) using 3 encoding blocks with two inputs and one output (E_0, E_1, and E_2), and one final encoding block E_3 with three input channels and two output channels. Additionally, 4 parameter sets (P_0 to P_3) are extracted.

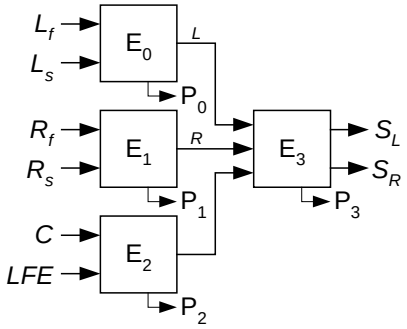

Figure 5: Spatial encoder tree with 6 channel input and a stereo down mix.

The modular architecture of encoding blocks does not only provide flexibility in terms of input, down-mix and output channel configurations; it also provides means for format conversion and advanced rendering. For example, the MPEG Surround decoder features a so-called 'enhanced matrix mode' (EMM) to convert conventional or matrixed-surround compatible stereo to multi-channel audio. The MPEG Surround decoder in EMM mode is visualized in Figure 6. A stereo signal is analyzed using an analysis stage (A) that generates spatial parameters that are used for spatial synthesis. Extensive listening tests have shown superior sound quality of the MPEG Surround EMM mode compared to conventional matrixed surround systems (cf. [17]).

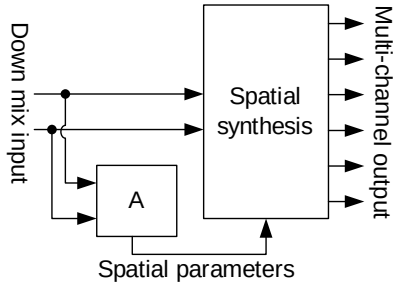

Figure 6: MPEG Surround EMM mode. Spatial parameters are estimated based on the analysis (A) of a stereo down mix.

A second feature of MPEG Surround that is especially interesting for mobile applications is the so-called 'binaural decoding mode' [18]. This mode provides a multi-channel audio experience over legacy stereo headphones. Due to the incorporation of novel methods to

represent head-related transfer functions, and the tight coupling of multi-channel decoding and binaural rendering, the computational complexity of the MPEG Surround binaural rendering system is significantly lower than obtained for conventional, convolution based methods. This property is especially important for mobile applications with limited processing power and battery life.

4.3 Spatial audio object coding

One of the most recent developments in the field of spatial audio coding is the provision of user-interactivity in spatial imaging of individual auditory objects that are present in a down mix [19]. This technology is also referred to as 'spatial audio object coding' (SAOC). Application scenarios that may benefit from such user interactivity are teleconferencing (adjust level and position of each talker individually), gaming (for dynamic rendering of various sound sources depending on user behavior), broadcasting (to increase the level of a voice over for better speech intelligibility) or remixing (modifications of instruments or vocals in a song). The repertoire of feasible modifications to each sound source include (1) level modifications (roughly up to +/- 12 dB), spatial repositioning, sound source equalization, and effect processing of individual objects (for example the addition of reverberation).

A generic SAOC coding scheme is outlined in Figure 7. A set of object signals is combined to generate a down mix in an SAOC encoder. Additional SAOC parameters describe the relative powers and pair-wise correlations between the input object signals. The SAOC decoder receives the down mix and SAOC parameters. The SAOC parameters are combined with object rendering properties (such as the level and desired spatial location of each object) to estimate spatial parameters. Finally, a spatial synthesis stage generates the rendered signals.

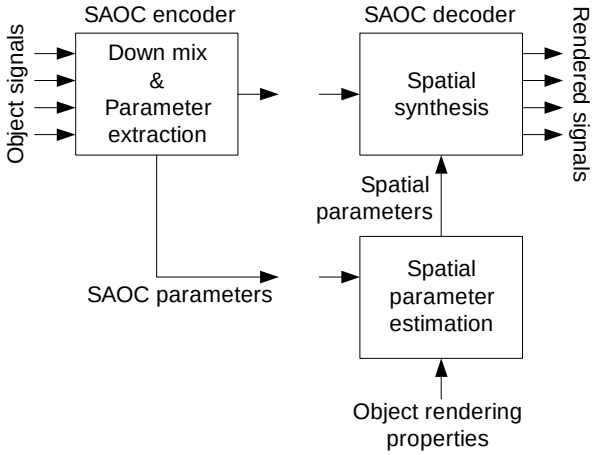

Figure 7: SAOC encoder and decoder.

Currently MPEG is developing a standard for SAOC functionality that is based on re-use of MPEG Surround as spatial rendering engine. In other words, the spatial parameter estimation stage produces an MPEG Surround compliant parameter bit stream that can be fed into an existing MPEG Surround decoder ([20, 21]).

5. Conclusions

Developments in the field of parametric spatial representations have resulted in new methods for compression, reproduction, processing and conversion of stereo and multi-channel audio. The technology is tightly coupled to spatial psychoacoustics and exploits known limitations of binaural processing of the human auditory system. The combination of unsurpassed compression gain, flexibility in channel configurations and the possibility to merge audio compression, processing and rendering stages in a unified parametric approach makes the approach suitable for a very wide set of applications.

References

1. K. Brandenburg: MP3 and AAC explained. Proc. 17th International AES conference, Florence, Italy (1999).
2. R. G. van der Waal, R. N. J. Veldhuis: Subband coding of stereophonic digital audio signals. Proc. ICASSP (1991).
3. J. Herre, K. Brandenburg, D. Lederer: Intensity stereo coding. Preprint 3799, proc. 96th AES convention, Amsterdam, The Netherlands (1994).
4. C. Faller, F. Baumgarte: Binaural cue coding: A novel and efficient representation of spatial audio. Proc. ICASSP (2002).
5. J. Breebaart, S. van de Par, A. Kohlrausch, E. Schuijers: High-quality parametric spatial audio coding at low bit rates. Proc. 116th AES convention, Berlin, Germany (2004).
6. J. Herre, H. Purnhagen, J. Breebaart, C. Faller, S. Disch, K. Kjörling, E. Schuijers, J. Hilpert, F. Myburg: The reference model architecture for MPEG spatial audio coding. Proc. 118th AES convention, Barcelona, Spain (2005).
7. C. Faller: Parametric coding of spatial audio. PhD thesis, EPFL, Lausanne, Switzerland (2004).
8. J. Breebaart, S. van de Par, A. Kohlrausch, E. Schuijers: Parametric coding of stereo audio. EURASIP J. Applied Signal Proc. **9** (2005) 1305-1322.
9. J. Blauert: Spatial hearing: The psychophysics of human sound localization. MIT Press, Cambridge, Massachusetts (1997).
10. B. Kollmeier, R. H. Gilkey: Binaural forward and backward masking: Evidence for sluggishness in binaural detection. J. Acoust. Soc. Am. **87** (1990) 1709-1719.
11. R. Y. Litovsky, H. S. Colburn, W. A. Yost, S. J. Guzman: The precedence effect. J. Acoust. Soc. Am. **106** (1999) 1633-1654.
12. B. R. Glasberg, B. C. J. Moore: Derivation of auditory filter shapes from notched-noise data. Hearing Research **47** (1990) 103-138.
13. E. Schuijers, J. Breebaart, H. Purnhagen, J. Engdegård: Low complexity parametric stereo coding. Proc. 116th AES convention, Berlin, Germany (2004).
14. M. Dietz, L. Liljeryd, K. Kjörling, O. Kunz: Spectral band replication, a novel approach in audio coding. Proc. 112th AES convention, Munich, Germany (2002).
15. J. Breebaart, G. Hotho, J. Koppens, E. Schuijers, W. Oomen, S. van de Par: Background, concept and architecture for the recent MPEG Surround standard on multi-channel audio compression. J. Audio Eng. Soc. **55** (2007) 331-351.
16. G. Hotho, L. Villemoes, J. Breebaart: A stereo backward compatible multichannel audio codec. IEEE Trans. Audio, Speech, Language Proc. (2007) accepted.
17. J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, K. S. Chong: MPEG Surround – The ISO/MPEG standard for efficient and compatible multi-channel audio coding. Proc. 122th AES convention, Vienna, Austria (2007).
18. J. Breebaart, J. Herre, L. Villemoes, C. Jin, K. Kjörling, J. Plogsties, J. Koppens: Multi-channel goes mobile: MPEG Surround binaural rendering. Proc. 29th AES international conference, Seoul, South Korea (2006).
19. C. Faller: Parametric joint-coding of audio sources. Proc. 120th AES convention, Paris, France (2006).
20. J. Breebaart, C. Faller: Spatial audio processing: MPEG Surround and other applications. Wiley, Chichester, UK (2007).
21. ISO/IEC JTC1/SC29/WG11 (MPEG) document N8853: Call for proposals on spatial audio object coding, 79th MPEG meeting, Marakech (2007).