Efficient Parametric Audio Coding for Interactive Rendering: The Upcoming ISO/MPEG Standard on Spatial Audio Object Coding (SAOC)

L. Terentiev, C. Falch, O. Hellmuth, J. Herre

Fraunhofer Institute for Integrated Circuits IIS

leonid.terentiev@iis.fraunhofer.de

Introduction

The production of music usually consists of several steps including the actual mixing of the individual audio sources (tracks). The consumer of music plays back this content passively, leaving no alternative to change even a tiny flavor of the given composition. A similar limitation, however for a completely different application, holds for current multiple-talker communication systems. Each participant receives a mixed signal of all other participants without the possibility to change the level or rendering position of the individual talkers. With the novel MPEG Spatial Audio Object Coding (SAOC) audio codec those and other applications can benefit from user-control of various audio objects at the playback side, for example changing the level or spatial position of the audio objects or switching from stereo to multi-channel reproduction. Although such functionality can be obtained by transmitting all objects independently, this is often not possible due to given bandwidth limitations and backwards compatibility requirements.

MPEG SAOC is an interactive audio concept which describes an efficient transmission chain comprising an object encoder for a number of audio sources (objects) and a corresponding decoder plus rendering unit. The SAOC system is designed to transmit these audio objects mixed into one or two downmix channels. A parametric description of all audio objects is stored in a dedicated SAOC side information bitstream. Although the parametric object related SAOC data grows linearly with the number of objects, the bitrate consumption of the coded object data is negligible compared to that of a coded audio (downmix) channel in a typical scenario. Therefore, the SAOC scheme is capable of transmitting multi-track content at bitrates only slightly exceeding those for mono or stereo signals.

This paper introduces the SAOC concept as a novel interactive audio technology and outlines the envisioned applications that can benefit from the user-control feature of various audio objects at the playback side.

SAOC Architecture

Background and Concept

Spatial audio coding technology, such as the MPEG Surround (MPS) standard [1], has introduced a new paradigm of highly efficient multi-channel audio processing. For complex audio content, this technology provides extremely high compression rates and computationally efficient rendering. Building upon this technology, MPEG issued a Call for Proposals for an SAOC system in January of 2007. The resulting SAOC work item provides a wealth of flexible user-controllable rendering tools based on the transmission of a conventional downmix extended by parametric audio object side information. In a conceptual overview, as illustrated in Figure 1, the sender (encoder) side combines a number N of audio objects in an audio signal that comprises one or two downmix channels. Together with this backward compatible downmix signal, audio object parameters are calculated and transmitted through a dedicated SAOC bitstream to the receiver (decoder) side. The decoder side can be considered as an object decoding part recovering the encoded N objects and a rendering part that allows manipulation and mixing of the original audio object signals into the composite audio output signal which may have several output channels. For these processes, the object decoder requires object metadata describing relevant audio object properties, while the renderer requires user-controllable object rendering information describing the level mapping from the audio objects to the playback channels. In practice, these two steps are performed in an integrated fashion, avoiding a costly intermediate upmix to N discrete audio object signals.

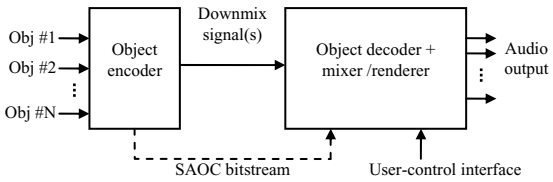

Figure 1: SAOC conceptual overview.

SAOC Decoding and MPS Transcoding Modes

Depending on the intended output channel configuration, the SAOC receiver module may either act as a decoder that directly generates a mono, stereo or binaural (reproduced over headphones) output signal, or it may incorporate a transcoder module and re-use an MPS decoder as a rendering engine yielding a 5.1 multi-channel output signal.

Figure 2 shows the block diagram of the SAOC processing architecture illustrating the SAOC decoder mode. The SAOC parameter processing engine decodes the SAOC bitstream and has an interface for additional input of time-

variant rendering information and Head Related Transfer Function (HRTF) parameters. The downmix processing module directly provides the output signal in a mono, stereo or binaural configuration by applying these user-specified parameters and transmitted SAOC data to the corresponding downmix signal.

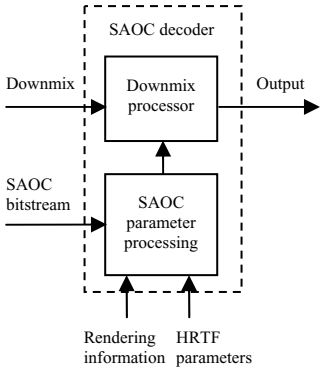

Figure 2: Block diagram of the SAOC decoder mode.

In Figure 3 the transcoding mode structure is depicted. Its architecture consists of the SAOC parameter processing engine and a downmix preprocessing module followed by an MPS decoder. Again, the SAOC parameter processing engine prepares parameters for the downmix preprocessor and provides a standards compliant MPS bitstream to the MPS decoder. This transcoding functionality enables the SAOC to produce a mix of the audio objects rendered in 5.1 audio channel format.

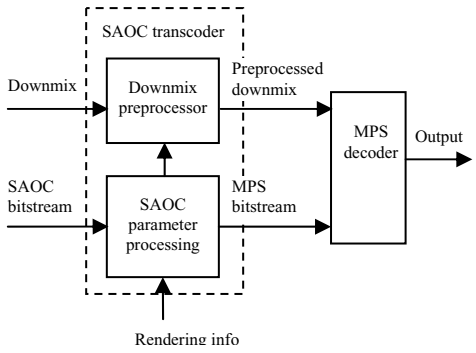

Figure 3: Block diagram of the SAOC transcoder mode.

SAOC Parameters

The SAOC system processes different types of audio objects (e.g. mono, stereo or multi-channel) into a specified (mono or stereo) downmix signal and SAOC parameter data. A hybrid Quadrature Mirror Filter (QMF) filterbank is used enabling frequency selective processing of all SAOC parameters [1]. The filterbank structure and frequency grid defining the parameter bands on which the parameters operate, are re-used from previous MPEG technologies such as MPS [1]. Each data frame of the SAOC bitstream contains one or more sets of parameters for each parameter band, where every set corresponds to a certain block of samples in time. The concept of SAOC is to extract perceptually relevant cues like Object Level Differences (OLD) and Inter-Object cross Coherences (IOC) from the input audio objects. The downmix information is retained by Downmix Gains (DMG) and Downmix Channel Level Differences (DCLD). These parameters are quantized and entropy coded such that the SAOC data can be transmitted in the ancillary data portion of the downmix bitstream. The overall size of the SAOC bitstream depends on the number and type of input audio objects. It approximately consumes 3 kbit/s per object (for high parameter resolution), but can be decreased significantly for certain applications.

Binaural Decoding

The binaural decoding mode facilitates the use of headphones as a playback device. The SAOC processing scheme provides a parametric Head Related Transfer Function (HRTF) database interface and is computationally more efficient than the conventional approach of performing HRTF processing on a multi-channel signal. In SAOC the binaural rendering for all audio objects is realized by processing the single mono downmix signal with two HRTFs (two pairs of HRTFs for a stereo downmix signal), which are derived from a combination of rendering information, SAOC and HRTF-database parameters. Full freedom in 3D positioning (including positioning outside the horizontal plane) is possible by mapping the objects to any combination of virtual loudspeakers represented by the HRTFs. Consequently, the incorporation of head tracking can be implemented by dynamically updating the rendering information according to head movements. The parametric approach ensures a compact representation of HRTFs yielding minimum storage requirements for mobile devices.

Computational Complexity

Different application scenarios have a wide range of complexity requirements. For battery powered mobile devices it is desired to utilize low complexity solutions, sometimes even at the price of a slight compromise in audio quality or functionality. Some application scenarios envisioned for SAOC typically target such complexity-constrained platforms. Typically, the overall SAOC decoding/transcoding complexity is dominated by the complex valued analysis/synthesis filterbank processing and signal decorrelation. These most critical parts in the SAOC processing share the LP MPS (Low Power MPEG Surround) solutions like partially complex QMF filterbanks, low power decorrelator for the mono and its exclusion for the stereo downmix case [3]. The complexity relation between High Quality (HQ) and Low Power (LP) SAOC processing modes resembles the corresponding relation between the HQ and LP mode of MPS and saves up to 50% of the computational power. This supports a successful adoption of the SAOC technology by the industry for a wide range of applications.

Latency

In order to apply the SAOC as an efficient interactive object based parametric coding scheme for teleconferencing applications, it is essential that SAOC processing does not add a significant delay on top of the core coder used for the downmix signal. The overall latency of the system using a Low Delay (LD) downmix core coder and employing SAOC as pre-/post-processor consists of the core coder delay and algorithmic delay of the SAOC processing. One can represent the whole LD SAOC system using two parallel

synchronized data paths, i.e. the downmix signal path and the SAOC parameter transmission path. In the downmix signal path the delay of the SAOC decoder directly adds on top of the core coder delay. On the other hand, parts of the parameter delay of the SAOC encoder can be "hidden" in the downmix delay, since these parameters are transmitted in parallel with the downmix signal. The LD SAOC decoder employs a LD QMF analysis/synthesis filterbank processing to transform signals from time domain to frequency domain and vice versa. With an appropriate choice of the core coder (i.e. MPEG-4 AAC ELD without SBR) a typical algorithmic delay of LD SAOC is about 16 ms, which makes it perfectly suitable for telecommunication applications.

Parameter Mixing of SAOC Bitstreams

Many telecommunication applications request the possibility to combine a number of distributed clients. In general, this task is assigned to a so-called Multipoint Control Unit (MCU). The SAOC concept incorporates the MCU functionality, which is able to merge several audio object streams on a parameter level without de/re-encoding of the corresponding downmix signals. As illustrated in Figure 4 the MCU combines the input SAOC bitstreams into one common SAOC bitstream such that the resulting output bitstream represents the properties of all audio objects from the two input bitstreams. The downmix signals are combined in parallel by the audio combiner. The computational efficiency of this process and low bitrate consumption makes SAOC technology ideally suited for teleconferencing applications.

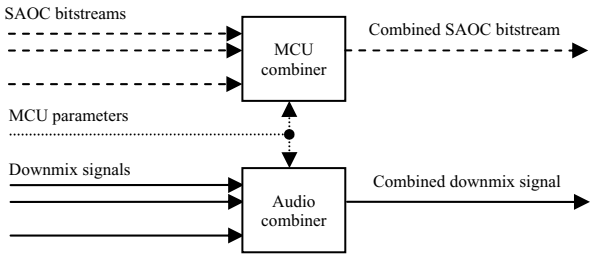

Figure 4: Outline of the SAOC MCU combiner.

Multi-Channel Background Objects

There are some application scenarios for which the audio input to the SAOC encoder contains a so-called Multi-Channel Background Object (MBO). The MBO can be considered as a complex sound scene (e.g. 5.1 channel mix) involving a large and often unknown number of different sound sources, for which no controllable rendering functionality is required. The concept of the SAOC architecture can process these complex input signals together with the typical mono and stereo audio objects. The combination of the regular SAOC audio objects and the MBO is achieved by first applying the MPS encoder yielding an MBO downmix. This downmix then serves as an input object to the SAOC encoder together with the controllable SAOC objects which results in a combined downmix that is transmitted to the SAOC transcoder. The SAOC bitstream including MPS data for the MBO is fed into the SAOC transcoder which provides the appropriate MPS bitstream for the MPEG Surround decoder. This task is performed using the downmix preprocessor based on the rendering information.

Enhanced Karaoke/Solo Processing

For application scenarios requiring very strong amplification or attenuation of individual audio objects (e.g. Karaoke-type applications), an enhanced downmix preprocessor mode can be activated at the SAOC decoder/transcoder side. Figure 5 shows the architecture of a so-called Enhanced Karaoke/Solo (EKS) processor comprising an object separation block and a rendering unit. Using the EKS processor, it is convenient to classify all input audio objects into a static background object (BGO) (e.g. background music scene, MBO, etc.) and flexibly controllable foreground objects (FGOs). These groups can be efficiently recovered from the common downmix signal by a "One-To-N"/"Two-To-N" (OTN/TTN) unit, which supports partially waveform coded objects and in this way achieves a high quality of separation [1]. In the subsequent stage they are appropriately rendered by the successive rendering unit. In practice, the efficient implementation allows to perform these two processing steps in a combined manner thus saving computational costs and memory.

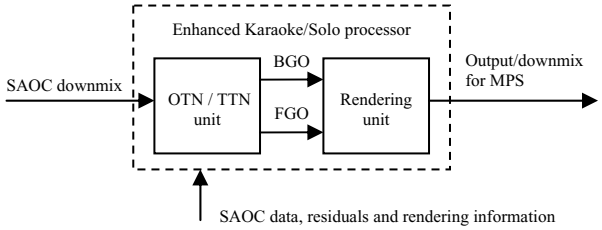

Figure 5: Architecture of the EKS processor.

Effects Interface

In addition, the flexibility of the SAOC technology is extended by an effects interface providing means for insert- and send-effects applied either in series or parallel to the signal processing chain. The SAOC effects interface provides means to incorporate room acoustic simulation in a very flexible manner.

Conclusions

An innovative Spatial Audio Object Coding system has been presented, that is currently under standardization in ISO/MPEG. SAOC successfully builds upon the rendering engine of MPEG Surround and is employed as a transcoding stage prior to MPEG Surround rendering. Because of its one-step object decomposition and rendering on the playback side, the SAOC parametric object based audio coding system offers unsurpassed efficiency in bitrate as well as complexity for the transmission of audio objects over a low bandwidth channel. The interactive nature of the rendering engine makes SAOC especially suited for a variety of attractive applications in the field of teleconferencing, remixing and gaming. Even for already well established applications, SAOC can add new compelling features and functionality in a backward-compatible manner.

The recent development activities in the SAOC technology at Fraunhofer IIS are performed in collaboration with the EU project "Together Anywhere, Together Anytime" (TA2) of the 7th Framework Programme (FP7) of European Commission [4].

References

[1] "MPEG audio technologies – Part 1: MPEG Surround", ISO/IEC Int. Std. 23003-1:2007, 2007.

[2] J. Engdegård, B. Resch, C. Falch, O. Hellmuth, J. Hilpert, A. Hölzer, L. Terentiev, J. Breebaart, J. Koppens, E. Schuijers, W. Oomen, "Spatial Audio Object Coding (SAOC) - The Upcoming MPEG Standard on Parametric Object Based Audio Coding" 124th AES conv., Amsterdam, Netherlands, May 2008, pp. 7377.

[3] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegåard, "Low complexity parametric stereo coding", in AES 116th Convention, Berlin, Germany, May 2004.

[4] http://www.ta2-project.eu