Spatial Audio Object Coding (SAOC) – The Upcoming MPEG Standard on Parametric Object Based Audio Coding

Jonas Engdegård[1], Barbara Resch[1], Cornelia Falch[2], Oliver Hellmuth[2], Johannes Hilpert[2], Andreas Hoelzer[2], Leonid Terentiev[2], Jeroen Breebaart[3], Jeroen Koppens[4], Erik Schuijers[4] and Werner Oomen[4]

[1]*Dolby Sweden AB, Gävlegatan 12A, 11330, Stockholm, Sweden*

[2]*Fraunhofer Institute for Integrated Circuits IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany*

[3]*Philips Research Laboratories, High Tech Campus 36, 5656 AE, Eindhoven, The Netherlands*

[4]*Philips Applied Technologies, High Tech Campus 5, 5656 AE, Eindhoven, The Netherlands*

Correspondence should be addressed to Jonas Engdegård (`jengd@dolby.com`)

ABSTRACT

Following the recent trend of employing parametric enhancement tools for increasing coding or spatial rendering efficiency, Spatial Audio Object Coding (SAOC) is one of the recent standardization activities in the MPEG audio group. SAOC is a technique for efficient coding and flexible, user-controllable rendering of multiple audio objects based on transmission of a mono or stereo downmix of the object signals. The SAOC system extends the MPEG Surround standard by re-using its spatial rendering capabilities. This paper will describe the chosen reference model architecture, the association between the different operational modes and applications, and the current status of the standardization process.

1 INTRODUCTION

The typical audio production and transmission chain consists of a set of operations that are executed in a very specific order. For example for musical content, various audio objects (instruments, vocals, etc) are first recorded (or synthetically produced), and subsequently mixed for playback on a specific reproduction system. The mixing process is performed by an audio engineer who decides on object positioning, relative levels and effects that are employed according to esthetical and technical objectives. In many applications, the resulting mix is transmitted using lossy compression algorithms. This conventional chain leaves virtually no flexibility in changing the composition at the reproduction side. A similar limitation holds for multiple-talker communication systems (teleconferences). The speech signals are typically combined into a mono downmix which is transmitted to the various far ends, leaving no means to adjust levels or positions in a stereo perspective of the various talkers.

There exists a range of applications that can benefit from user-control of various audio objects at the playback side. Examples are teleconferencing, remixing applications, on-line gaming, and karaoke functionality. Although such functionality can be obtained by transmitting all objects independently, this scenario is undesirable due to large bandwidth requirements and the fact that it is difficult to guarantee a certain esthetical quality level, which is extremely important in the music industry.

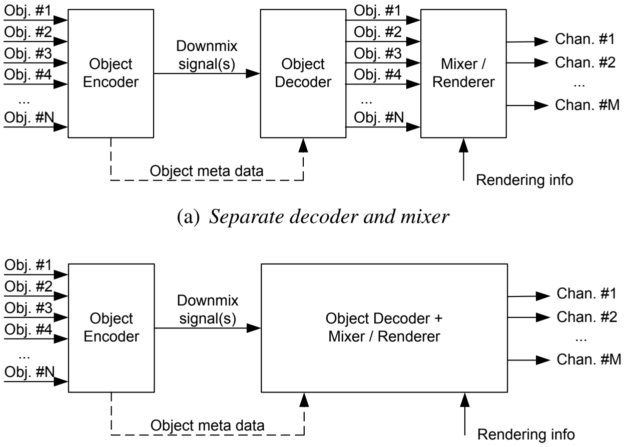
Spatial Audio Object Coding (SAOC) is a parametric multiple object coding technique that overcomes these drawbacks. It is designed to transmit a number N of audio objects in an audio signal that comprises K downmix channels, where $K < N$ and K is typically one or two channels. Together with this backward compatible downmix signal, object meta data is transmitted through a dedicated SAOC bitstream to the decoder side. Although this object meta data grows linearly with the amount of objects, the amount of bits required for coding these data in a typical scenario is negligible compared to the bitrate required for the coded downmix channels.

In a conceptual overview, as illustrated in Figure 1(a), the decoder side can be divided into an object decoding part decomposing the N objects and a rendering part that allows manipulation and mixing of the original audio object signals into M output channels. For those processes, the object decoder requires the object meta data while the renderer requires object rendering information. Figure 1(b) shows a similar system where object decoding and rendering are performed in an integrated fashion, avoiding a costly intermediate upmix to N discrete audio object signals.

Section 2 outlines the envisioned applications for SAOC. Section 3 describes the joint process of object decomposition and rendering as it is the core of SAOC, and how this relates to Spatial Audio Coding (SAC).

2 APPLICATIONS

SAOC serves as the basis for many different applications based on the identical core technology. The three most prominent applications will be described in the following subsections.



(a) *Separate decoder and mixer*

(b) *Integrated decoder and mixer*

Figure 1: *SAOC conceptual overview*

2.1 Teleconferencing

With only few exceptions, current telecommunication infrastructures (telephone, teleconferencing systems, etc.) are all monophonic. This is without disadvantage if only one-to-one communication is considered. If more than one communication partner is transmitted over the monophonic channel, all legacy solutions do not enable access to the individual signal components (e.g. speaker X and speaker Y). SAOC however, allows the free arrangement of all communication partners into a virtual conference room. That way, speakers can easily be distinguished, and volume can be adjusted for each speaker individually. Informal experiments have shown that this leads to increased intelligibility and an overall more comfortable and less exhausting experience due to an added dimension of realism. For communication situations where a visual transmission of the other communication partners is involved, SAOC allows to spatially synchronize the audio signal with the visual representation (e.g. acoustically place the audio signal of speaker X at the same position on the screen where the corresponding video signal is played back). The conference may be rendered on any typical loudspeaker configuration (stereo or multi-channel) or even on headphones using binaural techniques to create a virtual 3D sound image. Whereas SAOC minimally increases the overall bitrate to offer this kind of flexibility, conventional systems would require multiple transport channel capacity. Thus, SAOC allows for cost-effective implementation of enhanced voice services. The SAOC side information is conveyed in a hidden, backward compatible way, allowing legacy terminals to decode the full communication signal without the additional SAOC functionality.

A different way of teleconferencing is the "voice chat" application in Massively Multiplayer Online gaming (MMO) or other social-oriented virtual worlds. In such contexts, there is typically already an IP based communication framework available. In addition, the heart of the communication protocol between the players is their virtual coordinates used in the common game world. This makes it highly suitable to use SAOC as a spatial communication system. As the coordinates from other players get combined with the user's orientation, the necessary SAOC matrix parameters can easily be derived, and hence the voices of the players get rendered and realistically positioned in space. Inherently, such a system would support any playback configuration including e.g. binaural, 5.1 or any other popular game audio setup.

2.2 Backwards Compatible Interactive Remixing

The SAOC technology is also perfectly suited for interactive re-mix applications. This means that mono or stereo content can be provided in arbitrary formats to the consumer along with SAOC data describing the objects present in the mix. The user is now able to create his or her own remix of the music, or the sounds in the mix. SAOC allows managing the audio objects similarly to that of a multi-track mixing desk such as: Adjusting the relative level, changing the spatial position and using effects. In this way one can e.g.,

- Suppress / attenuate certain instruments for playing along (karaoke type of applications)
- Modify the original track to reflect the user's preference (e.g. "more drums and less strings" for a dance party; "less drums and more vocals" for relax music). In case of many different objects it is also possible to group and modify certain instrument classes simultaneously (e.g. the brass or percussions section in an orchestra).
- Choose between different vocal tracks ("female lead vocal versus male lead vocal")

For online music distribution, another interesting application is an upgrade service for existing customers of music content. Similarly as MPEG Surround [1] may offer an upgrade option for the added value of multichannel sound in a backwards compatible way, SAOC may offer object modification features. A commercial service for example, provides two types of downloads, the core content perfect for conventional listening, and an additional (small size) download, upgrading the core content with SAOC side information to enable a karaoke type of application.

In the context of broadcasting or movie playback SAOC allows the same flexibility as described above. In a typical scenario the audio material comprises background music or ambient sounds in combination with a dialog, narration or voice-over. Here, the user can for example change the level of the dialog with respect to the background music and thereby increase the intelligibility of the dialog or voice-over. This is especially interesting for users with hearing impairments or in case of playback in an environment with a fairly high background noise level. As the use of mobile media such as DVB-H is emerging, the same content (i.e. the same audio mix) may be broadcast to the living room and to a mobile receiver in an urban noisy environment. This issue is typically addressed by such an SAOC enhancement application.

2.3 Gaming / Rich Media

For rich media content, SAOC can be used exploiting its high compression efficiency and computationally efficient rendering. The applications of rich media are manifold as they span a wide scope ranging from interactive audio/visual interfaces to computer games. Typical examples of such applications and platforms include Flash or Java based rich media platforms or mobile gaming in general. On those platforms, the audio rendering capabilities are often kept on a simplistic level in order to limit complexity or data size. A legacy decoding/rendering system that handles many separate audio object streams gets increasingly complex with a growing number of audio objects as every object needs its individual decoder instance. Using SAOC technology, the complexity depends less on the number of objects since they are handled parametrically. The complexity of SAOC is essentially defined by the number of downmix and output channels, which is independent of the number of controllable audio objects.

One of the main aspects of game audio is the interactive rendering of audio objects. Here SAOC can serve several functions. Background music in games is often applied as interactive music, which is since long an important concept in gaming technology. The efficient representation of object controllable (interactive) music that SAOC offers results in small audio data size and low computational complexity when the music is decoded and rendered. The handling of environmental sounds can also benefit from SAOC in a similar way, as they often are implemented as short loops in order to save memory.

3 SAOC ARCHITECTURE

Spatial audio coding technology (such as MPEG Surround as standardized in ISO/MPEG) [2, 1] recently opened up possibilities for a new paradigm of user-controllable manipulation. The resulting Spatial Audio Object Coding (SAOC) work item provides a wealth of rendering flexibility based on transmission of a conventional mono or stereo downmix, extended with parametric audio object side information that enables the desired flexibility. In January 2007, Moving Picture Experts Group (MPEG) issued a Call for Proposals (CfP) for an SAOC system. After evaluation of the responses six months later, a reference model zero (RM0) was chosen. Interestingly, the RM0 system re-uses MPEG Surround (MPS) as rendering engine using transcoding and effects extensions. At the receiver side, an SAOC bitstream containing the parametric audio object description is transcoded into an MPS bitstream containing the parametric audio channel description. Another important input to this transcoder is the rendering matrix, describing the (level) mapping from the audio objects to the playback channels. This transcoder functionality is comparable to an audio mixing console with the object signals at the inputs. The system also facilitates the integration of insert- and sum-effects. Furthermore, the SAOC system provides a broad range of playback configurations (inherited from MPEG Surround) including 7.1, 5.1, binaural stereo and plain stereo. The various objects can have a mono, stereo or a multi-channel format (so-called Multi-channel Background Objects, or MBOs).

3.1 Hybrid QMF

An essential part of the spatial audio technologies, MPEG Surround as well as SAOC, are the Quadrature Mirror Filter (QMF) banks [3] which serve as time/frequency transform and are required to enable the frequency selective processing. The QMF bank has near alias-free behavior even when altering the gains of neighboring subbands excessively, which is a fundamental requirement for these systems.

The same QMF is also the native filterbank in Spectral Band Replication (SBR) [4, 5] and Parametric Stereo (PS) [6, 7]. SBR and PS are the vital enhancement tools in the MPEG-4 HE-AAC (V2) codec, and by combining SAOC with SBR one can further improve low bitrate performance. Furthermore, all three cascaded post processes, SBR – SAOC – MPS can be done consecutively in the QMF domain, enwrapped by only one analysis and synthesis stage, and hence allowing significant complexity savings. The SAOC-to-MPS transcoding becomes straightforward and without quality loss as matching filter banks are used.

The 64 channel QMF bank offers linearly spaced frequency bands, hence the effective bandwidth for a sampling rate of 44.1 kHz is approximately 345 Hz per frequency band. At the lowest frequencies this resolution is far lower than proposed bandwidth scales that take auditory perception into account, e.g. the Equivalent Rectangular Bandwidth (ERB) scale [8]. This motivates the introduction of a hybrid filterbank structure for use in parametric channel extension methods [7], comprising additional subband filtering for the lower QMF bands. The complete hybrid filterbank achieves a good approximation of the ERB scale with effective bandwidths of approximately 86 Hz for the lowest bands up to the inherent 345 Hz for the upper bands. In Table 1, the frequency division of the hybrid filterbank is shown.

In order to reduce the amount of parameter data, it makes sense to keep the frequency resolution for the upper range lower than the native QMF bandwidth. Still following the approximated ERB scale, the QMF bands higher up in frequency are therefore grouped and averaged accordingly. This combined structure of frequency bands forms a 28 band base resolution, but can easily be scaled down to subsets by grouping neighboring bands. These bands are referred to as parameter bands.

3.2 SAOC Parameters

As taught in Section 3.1, the hybrid QMF bank is used for enabling frequency selective processing of all parameters. 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

Band	Frequency range	Bandwidth
0	0–86 Hz	86 Hz
1	86–172 Hz	86 Hz
2	172–258 Hz	86 Hz
3	258–345 Hz	86 Hz
4	345–517 Hz	172 Hz
5	517–689 Hz	172 Hz
6	689–861 Hz	172 Hz
7	861–1034 Hz	172 Hz
8–68	1034–22050 Hz	345 Hz

Table 1: *Frequency resolution of the hybrid filterbank. Figures are based on 44.1 kHz sampling rate.*

in time. As in MPEG Surround the frequency resolution can be made dynamic by changing the number of parameter bands to be coded. Also, the update rate can be made adaptive, constituting a flexibly defined time/frequency grid. An important feature of the SAOC parametrization is the fact that it utilizes the same time/frequency grid engine as MPS. These shared properties result in a straightforward and also lossless transcoding of the signaling of time/frequency parameter mapping.

The following SAOC parameters are extracted on basis of the native time/frequency grid, as previously outlined.

- Object Level Differences (OLD), describing the relative energy of one object to the object with most energy for a certain time and frequency band.
- Inter-Object Cross Coherence (IOC), describing the amount of similarity, or cross-correlation for two objects in a certain time and frequency band.
- Downmix Channel Level Difference (DCLD), which is only applicable for objects in a stereo downmix, and Downmix Gains (DMG), describing the downmix processing of the input object signals, derived from the gain factors applied to each audio object.
- Object Energies (NRG), specifying the absolute energy of the object with the highest energy for a certain time and frequency band. This parameter is optionally transmitted to enable merging of several SAOC bitstreams in the parameter domain, a feature of special interest for e.g. Multipoint Control Unit (MCU) based teleconferencing applications. This application is discussed in Section 2 and more technical details are presented in 3.5.

SAOC provides quantization mechanisms in combination with a set of sophisticated entropy coding techniques, thus striving to minimize the amount of the SAOC side information while at the same time offering the best possible resulting audio quality. Depending on the parameter type, specific quantization schemes are used. Whereas e.g. Inter-Object Coherence (IOC) parameters are quantized with as few as eight quantization steps, the Object Energies (NRG) used for the MCU bitstream combination have a much higher precision of 64 quantization steps. Both uniform quantization on a logarithmic scale (e.g. for OLD and NRG), and non-uniform quantization (e.g. for IOC) are applied. For further reduction of the SAOC side information, the entropy coding is used for the majority of the quantized parameters, generally as a combination of differential coding and Huffman coding. Separate Huffman code books are trained for every parameter type and coding scheme.

The SAOC side information containing quantized object parameters is transmitted in a low bitrate side channel, e.g. the ancillary data portion of the downmix bitstream. The total SAOC parameter bitrate depends on the number and type of input objects and constitutes of approximately 3 kbit/s per channel of every object and an additional overhead of 3 kbit/s per audio scene.

3.3 SAOC to MPS Transcoding

Since an MPEG Surround (MPS) decoder serves as final rendering engine, the task of transcoding consists in combining SAOC parameters and rendering information associated with each audio object to a standards compliant MPS bitstream. The bitstream parser, which regains the parametric information, and the scene rendering engine, specifying the mapping of N objects to M output channels, denote the two elementary processing blocks of the SAOC to MPS transcoder, as depicted in Figure 2. In order to perform the desired mapping, a rendering matrix is calculated by the rendering matrix generation block exploiting information about the playback configuration, i.e. the number of loudspeakers available for playback together with their spatial positioning, on one hand, and object positioning and amplification information (so called rendering parameters), on the other hand. The rendering parameters can for instance be retrieved interactively from a user interface. As Sections 3.3.1

and 3.3.2 further describe, there are a few structural differences between the handling of mono, Figure 2(a), and stereo, Figure 2(b), downmixes with respect to downmix transcoding.

The rendering matrix A maps N objects to the desired number of output channels (e.g. 5.1). A is typically, but not necessarily constant over frequency and can also be dynamically updated in time. The target rendered output Y can be defined as $Y = AS$, where S denotes the object matrix for N objects of block length L according to,

$$S = \begin{pmatrix} s_1(0) & s_1(1) & \cdots & s_1(L-1) \\ s_2(0) & s_2(1) & \cdots & s_2(L-1) \\ \vdots & \vdots & \ddots & \vdots \\ s_N(0) & s_N(1) & \cdots & s_N(L-1) \end{pmatrix} \qquad (1)$$

From the SAOC parameters (OLD, IOC, DCLD, etc) as described in Section 3.2, one can derive an approximation of the object covariance, $E = SS^*$ where the elements in E contain the object powers and cross-correlations. Thus, the desired target output covariance, R can be expressed as,

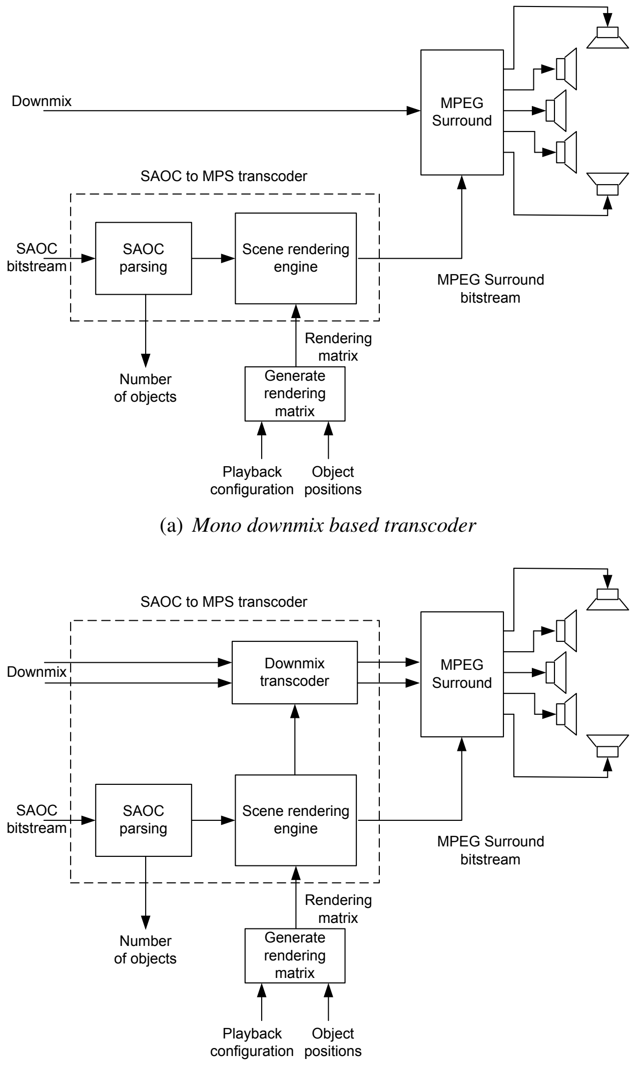
$$R = YY^* = AS(AS)^* = ASS^*A^* = AEA^* \qquad (2)$$

In the SAOC transcoder, R is an essential reference, as it represents the powers and cross-correlations of the output channels the MPS rendering stage should achieve. Hence, R leads to the MPS parameters to be produced by the SAOC transcoder, though different derivation processes need to be applied depending on the active transcoding mode.

The subsequent sections describe the possible transcoding modes: mono, stereo, binaural and the handling of Multi-channel Background Objects (MBOs).

3.3.1 Mono Transcoding

In case of a mono downmix, the MPS decoder employs a tree-structured parameterization based on cascaded One-To-Two (OTT) channel upmix blocks as further described in [2, 1]. The tree-structure leads to an upmix of the mono downmix into: Left-Front (l_f), Center (c), Right-Front (r_f), Left-Surround (l_s), Right-Surround (r_s) and the Low Frequency Effects/Enhancement (lfe) channel. Associated with each OTT element are the MPS parameters Channel Level Difference (CLD) and



(a) *Mono downmix based transcoder*

(b) *Stereo downmix based transcoder*

Figure 2: *SAOC to MPS transcoder with MPS decoder*

Inter-Channel Correlation (ICC). The estimation of all CLDs and ICCs from the SAOC data OLD and IOC is performed separately for each OTT element by deriving corresponding sub-rendering matrices from the rendering matrix A, thus ensuring a correct allocation of the rendering information of all objects and the output channels.

Re-panning objects can be done by means of the CLD and ICC parameters in the MPS renderer, while object attenuation/amplification requires incorporating Arbitrary Downmix Gains (ADGs) for a "virtual" modification of the downmix signal energy. ADG is a set of parameters part of the MPS bitstream that specifies gain modification. They are defined for each time/frequency tile that also is used by the other parameters. The computation of the ADGs is based on the rendering matrix and the SAOC parameters OLD and DMG.

Considering the elements of the target covariance matrix R, extracting the MPS parameters (CLD, ICC, ADG) for the mono downmix case is trivial. The diagonal elements r_{ii} representing the object powers, easily translate into the objects' absolute powers and channel distribution, ADGs and CLDs. Furthermore, the off-diagonal elements r_{ij} ($i \neq j$) translate into the ICC parameters corresponding to the channel pair i and j.

3.3.2 Stereo Transcoding

In the stereo downmix case, the Two-To-Three (TTT) channel upmix stage of MPS needs special considerations. The TTT process can be seen as a first order predictor of the center downmix, where the center downmix contains the center and LFE channel. In MPS, the Channel Prediction Coefficients (CPCs), along with an accompanying cross-correlation parameter describing the prediction loss, complete the mapping from the input stereo downmix to the three channel set of a combined left downmix ($l_\mathrm{f} + l_\mathrm{s}$), a combined right downmix ($r_\mathrm{f} + r_\mathrm{s}$) and a combined center downmix ($c + \mathrm{lfe}$), thus representing a left, right and center branch.

In an application where a part of the spectrum is coded without preserving phase information such as Spectral Band Replication (SBR) [4, 5], this frequency region can be upmixed substituting the CPC based scheme with an upmix based on CLD parameters which is more robust against inter-channel phase distortion.

Even though MPEG Surround is an almost fully generic rendering engine, two issues need to be addressed by special techniques: Firstly, since MPEG Surround performs an energy preserving upmix process it is not straightforward to apply the desired object gain values. However, the Arbitrary Downmix Gain (ADG) tool offers an elegant solution to this as outlined in Section 3.3.1. As the transcoder converts object parameters to channel parameters, the object gain values can be integrated into the ADGs. It can be noted that by the method of transcoding the object level parameters to the ADGs, the downmix signal can remain untouched by the transcoder and directly conveyed to the MPEG Surround decoder.

While this solution is adequate for the case where SAOC operates in a mono downmix based configuration as shown in Figure 2(a), the stereo downmix based configuration points to another issue. MPEG Surround, when operating in stereo downmix mode, expects the downmix (L_0, R_0) to be fixed and time invariant according to:

$$L_0 = l_\mathrm{f} + \alpha l_\mathrm{b} + \beta c \tag{3}$$

$$R_0 = r_\mathrm{f} + \alpha r_\mathrm{b} + \beta c \tag{4}$$

where α and β are the constant downmix coefficients usually set to $\alpha = \beta = 1/\sqrt{2}$. This basic downmix is referred to as the ITU downmix [9]. Even though the ADG tool has the freedom to alter the object gains, there are no means to move objects between the left and right branches in the upmix tree. As an example, an object that is solely present in L_0 can never be rendered to the right front or surround channel by MPS. To overcome this limitation a stereo processor is converting the SAOC downmix, (L_0, R_0) to a new modified downmix, (L'_0, R'_0) which enables MPS to upmix to the target covariance. This is illustrated in Figure 2(b) and a more detailed view of the stereo processor is shown in Figure 3.

The MPS upmixing can be seen as a two stage process – firstly the Two-To-Three (TTT) upmix stage yielding the combined left, right and center channel,

$$\begin{pmatrix} l_\mathrm{f} + l_\mathrm{s} \\ r_\mathrm{f} + r_\mathrm{s} \\ c + \mathrm{lfe} \end{pmatrix} \tag{5}$$

and secondly to fulfill the rendering target with respect to the front/surround and center/LFE relations, their corresponding One-To-Two (OTT) upmix parameters are derived in the same manner as in the mono downmix case. This upmix stage yielding,

$$\begin{pmatrix} l_{\mathrm{f}} \\ l_{\mathrm{s}} \end{pmatrix}, \begin{pmatrix} r_{\mathrm{f}} \\ r_{\mathrm{s}} \end{pmatrix} \text{ and } \begin{pmatrix} c \\ \mathrm{lfe} \end{pmatrix} \tag{6}$$

for the left, right and center branches, respectively. It is therefore appropriate to first consider the target signal after the TTT stage, $Y_3 = A_3 S$, where A_3 defines the rendering matrix mapped to the three channels from (5). Let D be the downmix matrix mapping N objects to K downmix channels. Then, given the downmix signal $X = DS$, the prediction matrix C can approximate the target rendering:

$$CX \approx Y_3 \tag{7}$$

$$CDS \approx A_3 S \tag{8}$$

and hence C can be obtained by the least squares solution and using the relation $E = SS^*$,

$$C \approx A_3 E D^* (D E D^*)^{-1} \tag{9}$$

If A_3 would be restricted to what is achievable with the native TTT upmix function of MPS, then C would equal its corresponding TTT prediction matrix, C_{TTT}, which includes the parameters that need to be conveyed to the MPS decoder. However, for the general case the prediction matrix needs to be factorized into,

$$C = C_{\mathrm{TTT}} C_2 \tag{10}$$

where C_2 is a downmix modifier, part of the stereo processor.

As shown in Figure 3, C_2 is applied in the SAOC stereo processor while the C_{TTT} matrix is sent to the MPS decoder. C_2 can be seen as the mixing matrix that assures the objects to be correctly positioned in a left-right perspective. An example that clearly illustrates the importance of the stereo processor is the case of a music/vocal decomposition where vocals are mixed into L_0 and the background music (without vocals) is mixed into R_0. This mix which is, for direct listening slightly awkward, can be suitable for a karaoke application or any application where full separation of a certain audio object is crucial. Here, a typical choice of C_2 would be:

$$C_2 = \begin{pmatrix} 0.707 & \sigma_{\mathrm{L}} \\ 0.707 & \sigma_{\mathrm{R}} \end{pmatrix} \tag{11}$$

resulting in the vocals being positioned in (phantom) center and the background music would become upmixed in a parametric stereo fashion using its left and right gain factors, derived from the corresponding CLD parameters (here denoted ΔL),

$$\sigma_L^2 = \frac{10^{\Delta \mathrm{L}/10}}{1 + 10^{\Delta \mathrm{L}/10}} \tag{12}$$

$$\sigma_R^2 = 1 - \sigma_{\mathrm{L}}^2 \tag{13}$$

However, upmixing with C_2 does not necessarily lead to the correct inter-channel cross-correlation of the background music. For that reason, a signal from a decorrelator needs to be mixed into the output. The bottom branch in Figure 3 contains the creation and addition of the decorrelated signal through the pre-mix matrix Q, the decorrelation process $h(n)$ and the post-mix matrix P. Here, Q and P are calculated to make the covariance matrix of the final output match the target covariance, (AEA^*), yielding a better match of the off-diagonal elements, i.e., the cross-correlations. The function of the decorrelator is to output a signal orthogonal to the input with a similarly perceived character, spectral envelope and subjective quality. The design of the decorrelation process, $h(n)$, is technically related to artificial reverberators, but also includes special concerns due to the dynamic mixing among other things, as discussed in [10]. The stereo processor in the SAOC transcoder reuses the decorrelator specified in MPS for efficient integration.

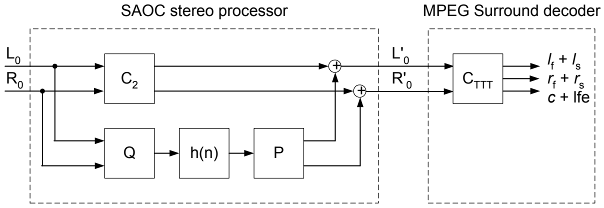


Figure 3: *Stereo downmix pre-processor*

3.3.3 Binaural Transcoding

Headphones as reproduction device have gained significant interest during the last decade. Mobility and social constraints in most cases dictate headphones as a reproduction device on mobile players. If conventional stereo material that was produced for loudspeaker playback is reproduced over headphones, the audio content is perceived inside the head [11, 12] due to the absence of the effect of the acoustical pathway associated with a certain sound source position.

An approach that is often pursued to resolve the lack of ’out-of-head’ localization is to simulate a virtual loudspeaker setup over headphones by means of Head-Related Transfer Functions (HRTFs) [13, 14]. This approach seems however suboptimal since it inherits all drawbacks of loudspeaker systems having a limited number of speakers and hence does not fully benefit from the full 3D positioning freedom that in principle exists when using headphones [15].

One of the challenges for 3D sound reproduction on headphones is the subject dependency of HRTFs [16, 17, 18, 19]. It has been shown that incorporation of head movement relaxes the requirements of using individualized HRTFs [20, 21, 22, 23]. Furthermore, especially in mobile devices, the large storage requirements for HRTF databases are undesirable [24].

To allow full flexibility in terms of HRTF database spatial resolution and individualization, the SAOC transcoder provides a parametric HRTF database *interface* rather than predefined HRTFs. This allows the implementer or end user to adapt the HRTF characteristics as well as their spatial sampling according to the application at hand. The parametric approach ensures minimum storage requirements for HRTFs given their compact representation [25]. The spatial positions for which HRTFs are available in the HRTF database can be freely selected, while SAOC objects can be mapped to any (combination of) HRTFs present in that database. In accordance with the method to map sound sources to channels for loudspeaker reproduction by means of a rendering matrix, the 3D binaural mode employs the same concept to map objects to HRTFs in the database. This method also allows the incorporation of head tracking by dynamically updating the render matrix according to head movements.

The SAOC transcoder for 3D binaural reproduction is outlined in Figure 4. The SAOC parameters are fed into the SAOC transcoder. Furthermore, an HRTF parameter database is connected to the transcoder that comprises HRTF parameters for a defined set of sound source positions. The mapping of objects to HRTF database entries is provided by means of a render matrix. The output of the SAOC transcoder consists of an MPEG Surround bitstream, accompanied by dynamic HRTF parameters that are fed to the MPEG Surround decoder that operates in binaural mode [26].

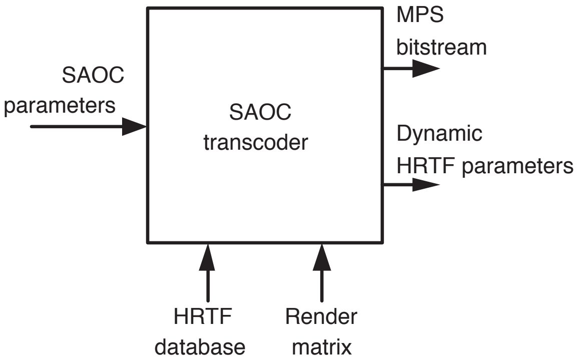


Figure 4: *3D binaural SAOC transcoder.*

The parametric approach allows the use of anechoic HRTFs only; the effects interface that is described in Section 3.6 provides means to incorporate room acoustic simulation in a very flexible manner.

3.3.4 MBO Transcoding

There are some application scenarios where the audio input to the SAOC encoder contains not only regular (mono or stereo) sound sources, but also a so-called Multi-Channel Background Object (MBO). An example for such an MBO is a 5.1 channel mix of different sound sources. Generally, the MBO can be considered as a complex sound scene involving a large and often unknown number of sound sources, for which no controllable rendering functionality is required. Individually, these audio sources cannot be handled efficiently by the SAOC encoder/decoder architecture. The concept of the SAOC architecture is extended in order to deal with these complex input signals, i.e. MBOs together with the typical mono and stereo audio objects.

The combination of the regular SAOC audio objects and MBO is achieved by first incorporating the MPEG Surround encoder yielding an MBO stereo downmix. Then this downmix serves as a stereo input object to the SAOC encoder together with the controllable SAOC objects producing a combined stereo downmix that is transmitted to the SAOC transcoder. The SAOC bitstream including an MPS data for the MBO, is fed into the SAOC transcoder which, depending on the particular MBO application scenario, provides the appropriate MPS bitstream for the MPEG Surround decoder. This task is performed using the rendering information and employing the downmix pre-processing unit. This extension of the SAOC architecture in a two-way analysis scheme is illustrated as a block diagram in Figure 5.

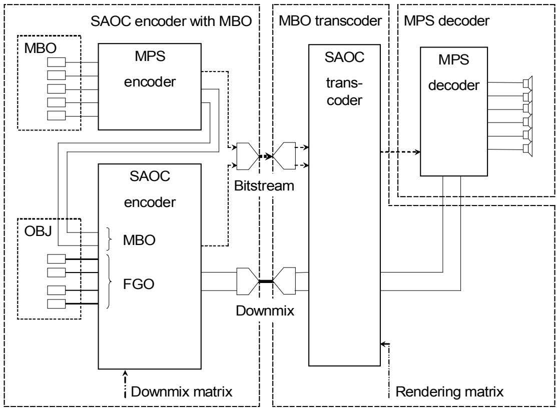


Figure 5: *Architecture of the SAOC system handling the MBO*

3.4 Karaoke / Solo Mode with Residual Coding

With the SAOC technology level manipulations of the individual audio objects are advisable just up to a certain amount. Extensive amplification or attenuation leads to an unacceptable decrease in audio quality. A special "karaoke-type" application scenario requires total suppression of specific objects, typically the lead vocal, while keeping the perceptual quality of the background sound scene unharmed. Vice versa, the utilization of one or a few predominant ForeGround Objects (FGOs) is referred to as "solo" mode and presumes full suppression of the accompanying BackGround Objects (BGOs). Thus, the conventional functionality of "mute" and "solo" can be represented by the notion of FGO and BGO.

These two particular application modes are supported by an enhanced SAOC architecture that exploits residual coding to improve the perceptual quality of the desired audio output.

3.4.1 Enhanced SAOC Architecture

The enhanced karaoke / solo technique employs a Two-To-N (TTN) upmix element at the transcoder and its corresponding inverse (TTN^{-1}) at the encoder. The latter is responsible for combining the BGO and FGO signals into a common SAOC stereo downmix signal and generating the required residual signals that are incorporated in the SAOC bitstream. While the TTN^{-1} element allows arbitrary positioning of all individual FGOs in the downmix signal, the BGO is considered a static stereo object, hence led unchanged to the downmix.

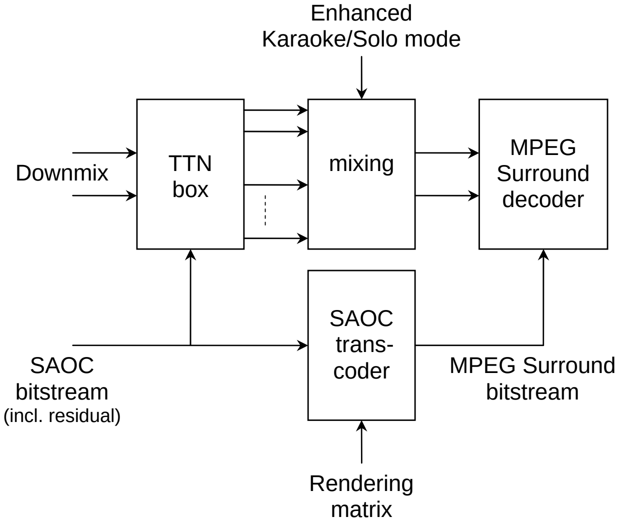


Figure 6: *The enhanced SAOC architecture*

As depicted in Figure 6, both the BGO and FGOs are extracted from the stereo downmix by the TTN element, which utilizes the corresponding information from the transmitted bitstream. Depending on the selected operating mode (i.e. karaoke or solo), the mixing stage produces a proper pre-processed downmix for the subsequent MPEG Surround decoder.

When encoding a Multi-channel Background Object (MBO) it is pre-processed as explained in Section 3.3.4 yielding a stereo signal that serves as the BGO to be input to the enhanced SAOC encoder. When decoding it in karaoke mode, the additional MPEG Surround bitstream that is encapsulated into an MBO, is provided to the MPEG Surround decoder.

3.4.2 TTN Parameters

The recreation of the N objects (i.e. BGO and FGOs) is carried out in the TTN element, a generalized, more flexible version of the TTT box known from the MPEG Surround specification [1]. The TTN element has two inputs, the stereo downmix (L_0, R_0) and residual signals

(r_i), which are extracted from the bitstream, decoded and transformed to the QMF domain. The element's output, the stereo BGO ($\hat{l}_\mathrm{B}, \hat{r}_\mathrm{B}$) and up to four FGO signals ($\hat{s}_{\mathrm{F},i}, i = 1\dots4$), form a linear combination of the input according to:

$$\begin{pmatrix} \hat{l}_\mathrm{B} \\ \hat{r}_\mathrm{B} \\ \hat{s}_{\mathrm{F},i} \end{pmatrix} = M_\mathrm{TTN} C_\mathrm{TTN} \begin{pmatrix} L_0 \\ R_0 \\ r_i \end{pmatrix} \tag{14}$$

The number of residual signals corresponds to the number of FGOs pre-defined at the encoder, a maximum of four individual FGO signals is supported by the reference model. To obtain M_TTN, the downmix matrix D is extended to yield a number of auxilary signals ($s_{0,i}$) in addition to (L_0, R_0), which also form a linear combination of the input objects:

$$\begin{pmatrix} L_0 \\ R_0 \\ s_{0,i} \end{pmatrix} = \widetilde{D} \begin{pmatrix} l_\mathrm{B} \\ r_\mathrm{B} \\ s_{\mathrm{F},i} \end{pmatrix} \tag{15}$$

Assuming four FGOs, $\widetilde{D}$ is given by:

$$\widetilde{D} = \begin{pmatrix} 1 & 0 & m_{1,1} & m_{1,2} & m_{1,3} & m_{1,4} \\ 0 & 1 & m_{2,1} & m_{2,2} & m_{2,3} & m_{2,4} \\ m_{1,1} & m_{2,1} & -1 & 0 & 0 & 0 \\ m_{1,2} & m_{2,2} & 0 & -1 & 0 & 0 \\ m_{1,3} & m_{2,3} & 0 & 0 & -1 & 0 \\ m_{1,4} & m_{2,4} & 0 & 0 & 0 & -1 \end{pmatrix} \tag{16}$$

where ($m_{1,i}, m_{2,i}$) denote the downmix weights for FGO i.

At the encoder, these auxilary signals are used to compute the residual signal for each FGO. M_TTN is equal to the inverse of the extended downmix matrix $\widetilde{D}$ and can therefore be derived from the DMGs and DCLDs at the transcoder. Two channel prediction coefficients ($c_{1,i}, c_{2,i}$) for each FGO are calculated from the OLDs and optionally IOCs and compose the prediction matrix C_TTN (e.g. for four FGOs):

$$C_{TTN} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ c_{1,1} & c_{2,1} & 1 & 0 & 0 & 0 \\ c_{1,2} & c_{2,2} & 0 & 1 & 0 & 0 \\ c_{1,3} & c_{2,3} & 0 & 0 & 1 & 0 \\ c_{1,4} & c_{2,4} & 0 & 0 & 0 & 1 \end{pmatrix} \tag{17}$$

3.4.3 Residual Coding

The parametric model introduced for reconstructing the FGO channels at the transcoder sometimes does not give an adequate perceptual audio quality. On the other hand, an audio signal representation that guarantees perfect (transparent) perceptual quality generally requires a fully discrete multi-channel coding technique resulting in significantly higher bitrates. An intermediate method that more efficiently improves audio quality by only a moderate increase of the bitrate is obtained by incorporating waveform coding of the prediction error, i.e. coding of the residual signal from the parametric model to be included in the SAOC bitstream as non-parametric side information.

The TTN^{-1} element at the encoder calculates the residual signal. Similarly to MPEG Surround, the residual signals are perceptually encoded and formatted using the MPEG-2 AAC low-complexity profile syntax [27].

3.5 Parameter Mixing of SAOC Bitstreams

In order to use the SAOC concept for teleconferencing applications a so-called Multipoint Control Unit (MCU) is required, which enables the combination of signals from several communication partners without decoding/re-encoding their corresponding audio objects. As illustrated in Figure 7 the MCU combines the input SAOC side information streams into one common SAOC bitstream in a way that the parameters representing all audio objects from the input bitstreams are included in the resulting output bitstream. These calculations can be performed in the parameter domain without the need to analyze the downmix signals and without introducing additional delay in the signal processing chain. In order to operate independently from the downmix signals the transmission of NRG parameters is needed the in combined SAOC bitstreams. The transmission of the resulting NRG parameters in the output SAOC bitstream can be activated to allow for nested bitstream combinations. In the case when the downmix signals are mixed with a certain level adjustment, the corresponding global mixing factor must be specified for the MCU combiner. Among other information, the MCU combination parameters contain the mapping function describing the required subset of audio objects and their order in the resulting SAOC bitstream.

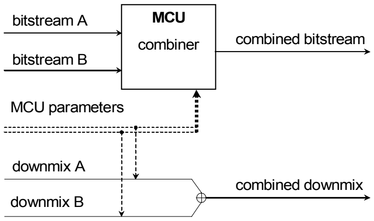


Figure 7: *Outline of the MCU combiner*

3.6 Effects Interface

In some use-cases the provided flexibility with respect to object location and relative volume levels may be insufficient. Audio engineers often employ effects to improve the artistic value of the reproduction. Such effects can be e.g., equalization, dynamic processing (compression), pitch effects, vocoder, delay, reverb, etc. In the binaural mode, the spatialization performance may benefit from additional reverberation as a complement to the anechoic behavior of the parameterized HRTFs.

In audio processing two types of effects processing can be discerned as shown in Figure 8. The first type, *insert effects*, is applied serially to the signal chain. Whereas the second effect type, *send effects*, creates a new parallel bus that can be fed by the mix of several signals. At a later stage the send effect bus (after effects processing) can be mixed together with other signals (or the main bus). Hence, using the same concept as in traditional mixing applications.

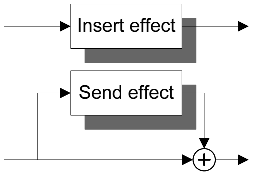


Figure 8: *Two types of effects processing.*

In order to provide the receiving end with the means to apply a wide range of effects in a way similar to mixing consoles, an effects interface is provided that gives access to the objects in the SAOC stream for both types of effect processing. Hence, SAOC provides handles to implement effects processing in the system but does not provide the effects itself.

Because insert effects change the properties of the objects in the downmix before rendering, the effects interface is a part of the downmix transcoder and is placed in front of the stereo processor described in Section 3.3.2. Figure 9 shows the downmix transcoder with the effects interface and stereo processor modules with corresponding inputs and outputs.

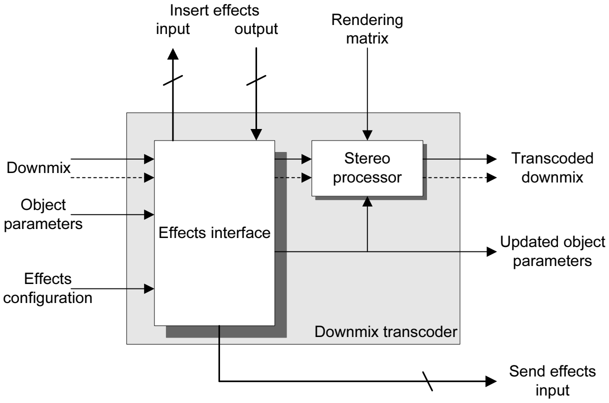


Figure 9: *Input and output data channels of the downmix transcoder.*

The effects interface requires configuration data which contains the following information

- **Object split vector** indicating which objects are used for insert effects.
- **Object extraction matrix** indicating the send effects signal construction as linear combination of objects.

Using this configuration information and the object parameters from the SAOC bitstream, the effects interface creates the desired signals and provides these as input to the effect modules where the actual effects are processed.

Next to the objects for insert effects processing the effects interface also provides the object parameters. The insert effects will in most cases change the spectro-temporal properties of the objects and therefore affect the validity of the object parameters. The insert effects modules are responsible for updating the parameters accordingly. This ensures a correct handling in the further processing and hence optimal performance.

After insert effect processing the resulting object signals are fed back into the effects interface where the objects are combined with the unprocessed objects into a downmix using the updated parameters. The updated parameters are used in the remainder of the transcoding. Figure 10 shows the effects interface in more detail where the insert effects are shown as a single module.

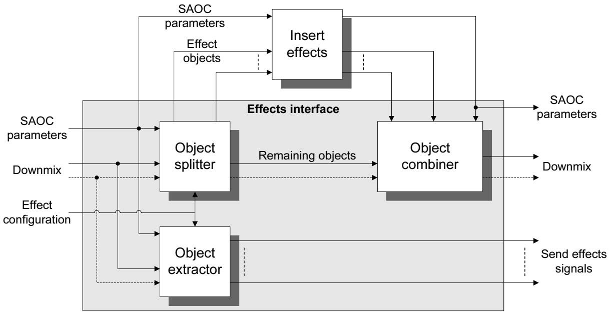


Figure 10: *The effects interface module.*

The send effects signals are generated by the object extractor and fed to the external effects for further processing. In stereo output mode, the send effects signals can be mixed with the output signal of the SAOC transcoder. In other modes where rendering is done by MPEG Surround, the signals can be mixed with the output of the MPEG Surround decoder. This is illustrated in Figure 11 which shows an overview of an SAOC system employing send effects.

The send effects signals are processed by effect modules and consecutively mapped to the reproduction configuration (e.g. 5.1 multi-channel or binaural). The mapping is in fact the mixing of the created send effect tracks, and can be seen as additional columns in the rendering matrix. The actual mixing with the object tracks is implemented by the addition with the MPEG Surround output.

It can be seen from Figure 11 that the downmix transcoding operations take place in the hybrid QMF domain. In this example the send effects are also applied in the hybrid QMF domain. This, however, is a choice of implementation. The hybrid QMF synthesis can be done at any stage after extraction of the send effect signals. For instance, for a time domain effect the hybrid QMF synthesis can be shifted in front of the effect module while at the same time using a hybrid QMF domain effect directly followed by the synthesis operation. The mapping to output channel configuration is then done in the time-domain, declining the possibility for frequency dependent mixing. Applications that require a low computational complexity, mixing with the MPEG Surround output can even be done in the hybrid QMF domain before the MPEG Surround synthesis filterbanks. This way, no additional filterbanks are needed for the send effects.

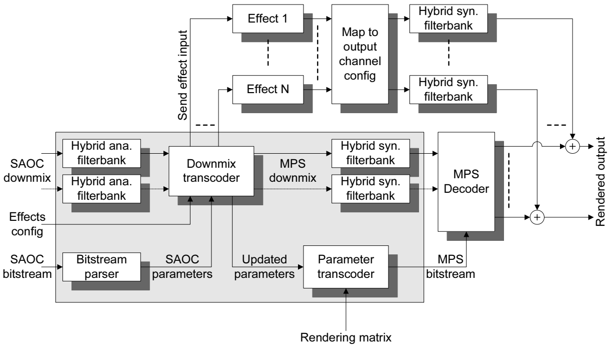


Figure 11: *Example of application of send effects.*

3.7 Operational Modes

In previous sections a number of operation modes were outlined and how they relate to the different processes in the SAOC transcoder. These modes can be more clearly illustrated if broken down to encoder and decoder options.

The following list summarizes the options available at the SAOC encoder as they affect what is conveyed in the bitstream or in the downmixed signal:

- The number of objects and what they are.
- The object type, such as: mono, stereo, MBO or enhanced with residual coding.
- The downmix type, such as: the number of downmix channels and downmix matrix.
- Other system properties, such as: the resolution of the time/frequency grid, sample rate, frame length, enable bitstream mixing, underlying core codec, random access frame rate, etc.

The following list summarizes the options available at the SAOC decoder as they, given a certain bitstream, affect the output presentation format or audio content:

- The playback configuration, such as: mono, stereo, binaural (headphones), 5.1 or even up to MPS' largest possible configuration of 32 loudspeaker channels.
- The rendering matrix, describing how to mix the objects given the chosen playback configuration.
- Addition and routing of effects, such as: compressor, equalizer, delay, reverb, flanger, etc.

An interesting feature given the listing above, is the full native support of all encoding options at the decoder, even at every chosen configuration. For instance, a minimalistic two object encoded SAOC bitstream can be decoded and rendered into a 5.1 surround mix, and a complex encoded bitstream even containing MBOs can be decoded and rendered into a plain stereo mix. From an application perspective, it is also of conceptual significance that the encoder is perfectly agnostic to the playback configuration of the decoder and the mixing interaction.

4 CONCLUSIONS

A novel Spatial Audio Object Coding (SAOC) system has been presented, that is currently under standardization in ISO/MPEG. SAOC successfully builds upon the rendering engine from MPEG Surround and is employed as a transcoding stage prior to the MPEG Surround decoder. With its joint handling of object decomposition and rendering on playback side, this parametric object based audio coding system offers unsurpassed efficiency in bitrate as well as complexity for the transmission of independent 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 those applications being already well established, SAOC can add new compelling features and functionality in a backward-compatible manner.

5 REFERENCES

[1] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, L. Villemoes, and K. Chong, "MPEG Surround - The ISO/MPEG Standard for Efficient and Compatible Multi-Channel Audio Coding," in *AES 122nd Convention*, Vienna, Austria, May 2007, Preprint 7084.

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

[3] P. Ekstrand, "Bandwidth extension of audio signals by spectral band replication," in *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio (MPCA-2002)*, Leuven, Belgium, Nov. 2002, pp. 73 – 79.

[4] ISO/IEC, "Coding of audio-visual objects – Part 3: Audio, AMENDMENT 1: Bandwidth Extension," ISO/IEC Int. Std. 14496-3:2001/Amd.1:2003, 2003.

[5] M. Dietz, L. Liljeryd, K. Kjörling, and O. Kunz, "Spectral band replication, a novel approach in audio coding," in *AES 112th Convention*, Munich, Germany, May 2002, Preprint 5553.

[6] ISO/IEC, "Coding of audio-visual objects – Part 3: Audio, AMENDMENT 2: Parametric coding for high quality audio," ISO/IEC Int. Std. 14496-3:2001/Amd.2:2004, 2004.

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

[8] B. Moore and B. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 750–753, Sept. 1983.

[9] ITU-R, "Multichannel stereophonic sound system with and without accompanying picture," ITU-R Recommend. BS.775-1, 1994.

[10] J. Engdegård, H. Purnhagen, J. Rödén, and L. Liljeryd, "Synthetic ambience in parametric stereo coding," in *AES 116th Convention*, Berlin, Germany, May 2004.

[11] S. P. Lipshitz, "Stereo microphone techniques; are the purists wrong?," *J. Audio Eng. Soc.*, vol. 34, pp. 716–744, 1986.

[12] J. Blauert, *"Spatial hearing: the psychophysics of human sound localization"*, The MIT Press, Cambridge, Massachusetts, 1997.

[13] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. I. Stimulus synthesis," *J. Acoust. Soc. Am.*, vol. 85, pp. 858–867, 1989.

[14] W. M. Hartmann and A. Wittenberg, "On the externalization of sound images," *J. Acoust. Soc. Am.*, vol. 99, pp. 3678–3688, 1996.

[15] J. Breebaart and E. Schuijers, "Why phantoms need to materialize on headphones," *IEEE Trans. On Audio, Speech and Language processing*, p. Submitted, 2008.

[16] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions," *J. Acoust. Soc. Am.*, vol. 94, pp. 111–123, 1993.

[17] A. Bronkhorst, "Localization of real and virtual sound sources," *J. Acoust. Soc. Am.*, vol. 98, pp. 2542–2553, 1995.

[18] H. Møller, M. F. Sørensen, C. B. jensen, and D. Hammershøi, "Binaural technique: Do we need individual recordings?," *J. Audio Eng. Soc.*, vol. 44, pp. 451–469, 1996.

[19] H. Møller, D. Hammershøi, C. B. Jensen, and M. F. Sørensen, "Evaluation of artifical heads in listening tests," *J. Audio Eng. Soc.*, vol. 47, pp. 83–100, 1999.

[20] F. L. Wightman and D. J. Kistler, "Resolution of front-back ambiguity in spatial hearing by listener and source movement," *J. Acoust. Soc. Am.*, vol. 105, pp. 2841–2853, 1999.

[21] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J. Audio Engineering Society*, vol. 49, pp. 904–916, 2001.

[22] P. J. Minnaar, S. K. Olesen, F. Christensen, and H. Møller, "The importance of head movements for binaural room synthesis," in *Proc. ICAD*, Espoo, Finland, July 2001.

[23] P. Mackensen, *Head movements, an additional cue in localization*, Ph.D. thesis, Technische Universitaet Berlin, Berlin, 2004.

[24] D. R. Begault, "Challenges to the successful implementation of 3-D sound," *J. Audio Engineering Society*, vol. 39, 1991.

[25] J. Breebaart and C. Faller, *"Spatial audio processing: MPEG Surround and other applications"*, John Wiley & Sons, Chichester, 2007.

[26] J. Breebaart, L. Villemoes, and K. Kjörling, "Binaural rendering in MPEG Surround," *EURASIP J. on Applied Signal Processing*, vol. Accepted, 2008.

[27] M. Bosi, K. Brandenburg, S. Quackenbush, L. Fielder, K. Akagiri, H. Fuchs, M. Dietz, J. Herre, G. Davidson, and Oikawa, "ISO/IEC MPEG-2 Advanced Audio Coding," in *J. Audio Eng. Soc.*, Oct. 1997, Vol. 45, No. 10, pp. 789-814.