SPATIAL AUDIO OBJECT CODING WITH ENHANCED AUDIO OBJECT SEPARATION

Cornelia Falch, Leonid Terentiev, and Jürgen Herre

Fraunhofer Institute for Integrated Circuits
Erlangen, Germany
cornelia.falch@iis.fraunhofer.de

ABSTRACT

Spatial sound reproduction on a multi-channel loudspeaker setup indicate a consistent trend in today's audio playback systems. Digital surround sound significantly improves the realism of the spatial sound experience, but also results in a drastic increase in required audio data rate. Spatial Audio Coding (SAC) technology provides means for efficient storage and transmission of multi-channel signals by a downmix signal and associated parametric side information describing the spatial sound image. More recently, SAC has been extended with an object-based concept termed Spatial Audio Object Coding (SAOC) enabling efficient coding and interactive spatial rendering of multiple individual audio objects at the playback side. Due to the underlying parametric coding approach, object level manipulations may affect the produced perceptual sound scene quality, and using extreme object attenuation or boosting may result in unacceptably degraded audio quality.

The paper describes how regular SAOC processing is advanced to ensure high quality sound reproduction even in demanding remix applications.

1. INTRODUCTION

In the area of audio reproduction systems, the trend is towards increasing the fidelity of spatial sound. The introduction of digital multi-channel audio formats has led to an enhanced authenticity in spatial reproduction, providing the listener with a more realistic sound sensation. However, the desire to migrate to multichannel audio has considerably increased the required data rate as compared to using mono or stereo formats. Spatial Audio Coding (SAC) [1] is a technology that facilitates the transmission of high-quality multi-channel audio signals at bitrates that have traditionally been used for transmitting mono- or stereophonic audio material.

Lately the original channel-oriented SAC approach has been developed further yielding an object-based concept termed Spatial Audio Object Coding (SAOC) [2], which aims at processing a number of distinct audio objects rather than channels of a multi-channel signal. SAOC enables efficient coding and permits user-controllable spatial rendering of multiple audio objects at the receiving side. Conceptually, the rendering functionality includes flexible, interactive spatial positioning and gain modification of the audio objects. While the underlying SAOC concept is well suited for spatialization leading to high-quality results, applying significant object level manipulations has been found to be a more challenging process. In case of extreme amplification or suppression of individual audio objects, consistently high audio quality may no longer be produced. The most demanding problem in this sense denotes to entirely mute an audio object. For example, a Karaoke-type playback requires the total suppression of the vocal object. Because all audio objects are contained in the downmix signal, it is essential for the SAOC decoding process to completely extract and remove the particular audio object that is to be suppressed from the downmix signal. Due to the insufficient object separation capability, the SAOC decoder cannot ensure an acceptably high quality for all rendered sound scenes selected by the user.

The paper introduces an enhanced SAOC technique providing high-quality sound scene reproduction even in demanding remix applications. This is achieved by improving the SAOC decoder's capability to separate the audio objects from the downmix. By considering the error of the parametric audio object representation, the proposed processing scheme combines the parametric coding approach with waveform coding. The resulting residual signal is coded and transmitted as part of the SAOC side information yielding an increase in overall bitrate. To this end the additional amout of data can be adjusted to obtain a trade-off between sound scene quality and required bitrate for transmission. A number of subjective listening tests have been conducted revealing a distinctly gain in audio quality already at a moderate bitrate expense.

Section 2 gives an introduction to the SAOC technology. Then the modifications of this "regular" SAOC system are described in Section 3 leading to the advanced coding capability for challenging remix applications. The performance of the enhanced SAOC processing has been evaluated by subjective listening tests, one of them is presented in Section 4.

2. SPATIAL AUDIO OBJECT CODING

SAOC has been forming the most recent expansion to the SAC technology. The fundamental idea of SAC is to characterize the spatial image of a multi-channel signal by a set of parameters, which are known to be essential for spatial auditory perception, and a downmix signal. A detailed description of SAC and an overview of related technologies (such as e.g. Binaural Cue Coding [3], [4], [5], and Parametric Stereo [6], [7]) is given in [1], [8], [9]. In 2007 SAC had become the basis of an ISO/IEC International Standard termed MPEG Surround (MPS) [10]. The request for delivering high-quality multi-channel audio at moderate bitrates is, amongst other requirements essential to the fields of broadcasting and on-demand services (e.g. Internet streaming), targeting at providing more realistic spatial audio reproduction to the consumers.

SAOC is designed for interactive and personalized audio applications providing the user control over a number of individual audio objects at playback side. With SAOC multi-object audio content is efficiently represented by a downmix signal and object related parametric side information which enables the consumer to create his/her own remix of the music. The goal for developing SAOC was to directly act upon the multiple audio objects in an interactive manner. The technology supports rendering of audio objects in mono, stereo and multi-channel format. Specifically the latter considers multi-channel background objects exhibiting limited rendering capability (i.e. restricted to gain modifications). In the context of personalized audio and musical recordings, the audio objects typically represent individual instruments and vocal tracks, and may be mixed by the user similar to the way a sound engineer performs the mixing process in a professional audio production. Remix applications benefit from SAOC's efficient signal representation (i.e. a downmix signal plus parametric side information instead of multiple discrete audio object signals) in two ways. On the one hand, due to this representation an interactive music content provider does not need to deliver the individual audio objects to the consumer, i.e. there is no need to distribute the original recorded audio tracks. In this way, SAOC supports copyright protection of the original recordings. On the other hand, even remix applications targeting at low bitrate (e.g. wireless) environments and limited terminal computing power (e.g. mobile devices) can be implemented perfectly with SAOC. The advantages offered in bitrate and computational efficiency as well as the rendering interactivity facilitate a wide range of applications with user-control, which can benefit from SAOC, such as network-oriented multi-player gaming and on-demand services (e.g. Internet streaming). Furthermore, SAOC supports broadcasting by offering the user the possibility to change the level of the dialog speech or a narrator with respect to background music or ambient noise in order to increase speech intelligibility. Teleconferencing systems may also benefit from improved intelligibility and yield reduced listening effort since multiple conferencing partners can easily be spatially rendered to a multi-loudspeaker setup of choice. These applications consider speech signals as audio objects.

One of the recent (and almost completed) ISO/MPEG standardization activities is dedicated to specifying SAOC as ISO/IEC International Standard MPEG Spatial Audio Object Coding [11].

2.1. SAOC Architecture

The SAOC system comprises an encoder preparing the audio data for transmission. At the receiver side, the rendered sound scene is processed by a decoder or a transcoder depending on the desired playback specification. SAOC considers a range of loudspeaker setups including mono, stereo and multi-channel (e.g. ITU 5.1) configurations with standard and non-standard loudspeaker positioning. Audio reproduction over headphones is supported with binaural processing by incorporation of a parametric HRTF processing module to enhance out-of-head localization and enable virtual 3D spatialization.

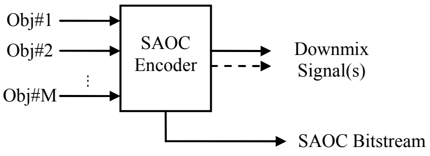

Figure 1: *Basic structure of the SAOC encoder.*

The basic structure of the SAOC encoder is shown in Figure 1. A set of individual audio objects compose the signal input to the encoder, which combines them into a mono or stereo downmix signal and generates the SAOC bitstream. To this end it is responsible for extracting perceptually motivated audio object parameters such as Object Level Difference (OLD) and Inter-Object cross Coherence (IOC) in a frequency selective manner.

The required filter bank, which is commonly used in SAC, exhibits a hybrid QMF structure [12], [13]. In addition, parametric information of the downmix process is computed and incorporated into the SAOC bitstream together with the OLDs, IOCs and other side information. After quantization, a lossless coding scheme is applied to all signal parameters being transmitted. This processing leads to the compact description of a complex audio scene consisting of a multitude of audio objects. Although the amount of object metadata increases with the number of audio objects, the bitrate of the parametric side information is marginal compared to that for transmitting the downmix audio signal, even if the latter is further processed by a legacy audio coder, such as HE-AAC [14]. Consequently, the entire audio scene can be stored or transmitted at a moderate bitrate.

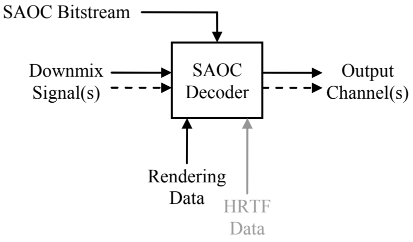

Figure 2: *Basic structure of the SAOC decoder.*

The SAOC decoder, depicted in Figure 2, obtains the transmitted downmix signal and SAOC bitstream and is additionally equipped with a user interface enabling interactive rendering of the audio objects to the channel(s) of the selected mono or stereo loudspeaker arrangement. For binaural headphones reproduction the SAOC decoder features an additional HRTF data interface enabling the user full flexibility in selecting individual HRTF data sets (with particular spatial resolution) rather than providing predefined HRTFs.

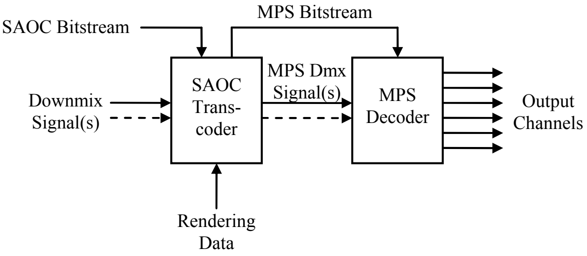

Figure 3: *Basic structure of the SAOC transcoder with MPS decoder.*

Multi-channel rendering is implemented by a two stage decoding structure consisting of an SAOC transcoder followed by an MPS decoder serving as rendering engine. This is shown in Figure 3. The task of transcoding consists of transforming the received SAOC downmix signal, bitstream and rendering information associated with each audio object into a standards compliant MPS downmix signal and bitstream.

3. ENHANCED AUDIO OBJECT SEPARATION IN SAOC

As described in the previous section, the SAOC representation of a multi-object sound scene comprises a downmix signal and object related metadata. On the receiving side, audio object separation from the downmix signal and rendering to the desired playback configuration is implicitly performed in one step.

The object power contributions in the downmix and output signals depend, besides other factors, on the corresponding downmix and rendering parameters. Since the latter can be controlled arbitrarily by the user, the relative object power contributions between downmix and desired output channel can be very different yielding a decrease in perceptual sound quality with increasing object power ratio. At extreme operating points the processing can no longer achieve an adequate subjective sound quality of the resulting sound scene. For example, if solo playback of object *#k* is desired, the power contributions present from all other objects (*i≠k*) act as *interference* in the according output channels. This is due to the underlying downmix / upmix (i.e. rendering) coding approach, which results in a limited object isolation / suppression capability in certain frequency bands. Consequently, parts of the interfering signals remain still perceptible and distort the solistic signal representation. Level amplification and attenuation can be performed well within a certain range, e.g. [-12dB; 12dB]. When trying to operate outside this range, however, the rendered channel signals may be become gradually distorted.

Nonetheless some applications require the total or almost total suppression of specific objects, e.g. for Karaoke playback the vocal object must be muted. For the converse case of playing back only one audio object (i.e. solo playback), all but the desired object must be suppressed. In order to meet these demands, the basic SAOC scheme has been extended by dedicated processing which enhances the resulting sound quality even in these demanding operating conditions. The enhancement is achieved by introducing the concept of *Enhanced Audio Objects* (EAOs) and supports both Karaoke and solo reproduction. Audio objects which are encoded as EAOs exhibit an increased separation capability from the other (regular) audio objects encoded in the same downmix signal (at the expense of an increased side information rate).

This section describes the incorporation of EAO processing into the SAOC system, i.e. at the SAOC encoder residual coding is introduced and the SAOC decoder/transcoder is extended by the EAO processor.

3.1. Concept of Enhanced Audio Objects

To reduce the interference of undesired audio objects in the rendered loudspeaker signals, a sufficiently good separation of retained and discarded audio objects from the downmix signal is essential. Consequently the audio input objects are divided into two groups, namely enhanced and regular audio objects. The set of EAOs comprise those audio objects that are expected to be either totally suppressed (Karaoke playback) or reproduced as solo objects. They can be interpreted as individual foreground objects. The remaining regular audio objects constitute the associated background sound scene.

With regular SAOC decoding/transcoding, object separation is performed by considering the OLDs and IOCs and applying a time and frequency dependent weighting of the downmix signal. In contrast to this, the concept of EAOs introduces additional specific parameters. The underlying processing method is adopted from MPS, where a so-called “two-to-three” (TTT) upmix process is used to derive a center channel from the left and right downmix channels by exploiting two channel prediction coefficients (CPCs) [8], [15]. This is depicted in Figure 4. Effectively, the downmix channels form a linear combination of three channels (i.e. left, right and center) being fed into a first order linear predictor to compute the CPCs.

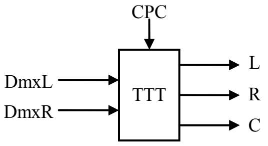

Figure 4: *Conceptual structure of the TTT upmix process in MPS.*

The method is integrated into EAO processing to predict each EAO signal individually, i.e. with SAOC an EAO corresponds to the center TTT channel and a downmix of the remaining regular audio objects corresponds to the left and right TTT channels. Consequently one TTT unit is required to extract one EAO. As illustrated in Figure 5 the SAOC decoder/transcoder incorporates a compact “two-to-N” (TTN) upmix process instead of *N* multiple TTT units.

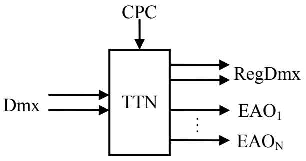

Figure 5: *Conceptual structure of the TTN upmix process in SAOC.*

In SAOC the functionality of the linear prediction model is extended to support also mono downmix signals next to the stereo configuration. Depending on whether a mono or stereo downmix signal is transmitted, either one single CPC or a pair of CPCs per EAO is required. While in MPS the CPCs are integrated into the bitstream, they are directly calculated from the OLDs and IOCs in the SAOC decoder/transcoder avoiding an additional bits load for the SAOC bitstream.

Although this parametric model clearly enhances the object separation capability at the SAOC decoder/transcoder, it does not ensure an adequate perceptual audio quality in any application situation. To this end, 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 requiring only a moderate increase of the bitrate is obtained by considering the error or *residual signal* of the parametric model. This approach has been introduced in MPS and is re-used in SAOC for EAO processing [1], [8]. In the SAOC encoder, the residual signal is calculated, waveform coded and included into the SAOC bitstream as non-parametric side information. As with MPS, the residual signals are perceptually encoded and formatted in close correspondence to the MPEG-2 AAC coding model [16]. Also, the residual signal may be chosen to enhance performance only in parts of the audio frequency range. Typically, residual coding provides the highest advantage in subjective audio quality when applied for the low frequency region.

3.2. Residual Signal Calculation in the SAOC Encoder

The calculation of the residual signals is performed in the SAOC encoder, which is extended by an additional residual processor yielding a two stage SAOC downmix process. First, the regular audio objects are combined into the regular downmix signal,

$$\mathbf{Y}_{reg} = \mathbf{D}_{reg}\mathbf{X}_{reg}\,, \tag{1}$$

with matrix $\mathbf{X}_{reg}$ comprising the regular audio objects and matrix $\mathbf{D}_{reg}$ being part of the SAOC downmix matrix $\mathbf{D}$, i.e. $\mathbf{D} = \begin{bmatrix}\mathbf{D}_{reg} & \mathbf{D}_{eao}\end{bmatrix}^T$.

Second, the SAOC downmix (of all audio objects) is computed as the sum of the regular downmix and the downmix-weighted EAOs,

$$\mathbf{Y} = \mathbf{Y}_{reg} + \mathbf{D}_{eao}\mathbf{X}_{eao}\,, \tag{2}$$

with matrix $\mathbf{X}_{eao}$ considering the EAOs.

An additional linear combination is computed to yield N_{eao} (denoting the number of EAOs) auxiliary signals,

$$\mathbf{Y}_{aux} = \mathbf{D}_{aux}\begin{bmatrix}\mathbf{Y}_{reg} & \mathbf{X}_{eao}\end{bmatrix}^T, \tag{3}$$

where $\mathbf{D}_{aux}$ is an orthogonal downmix matrix preserving the downmix information defined by matrix $\mathbf{D}$. The combination of both equations results in the extended downmix system

$$\mathbf{Y}_{ext} = \mathbf{D}_{ext}\begin{bmatrix}\mathbf{Y}_{reg} & \mathbf{X}_{eao}\end{bmatrix}^T, \tag{4}$$

with $\mathbf{Y}_{ext} = \begin{bmatrix}\mathbf{Y} & \mathbf{Y}_{aux}\end{bmatrix}^T$ and $\mathbf{D}_{ext} = \begin{bmatrix}\mathbf{D} & \mathbf{D}_{aux}\end{bmatrix}^T$. Provided that $\mathbf{D}_{ext}$ is invertible, full reconstruction of regular downmix and EAOs is ensured by inverting the extended downmix process. Thus, assuming that the auxiliary signals were available at the SAOC decoder/transcoder, the separation operation of the EAOs from the regular downmix would be optimal. However the auxiliary signals are discarded at encoder side since transmitting them (along with SAOC downmix signals and bitstream) leads to an unacceptable increase in bitrate. On the other hand the auxiliary signals can be estimated by a set of CPCs and the SAOC downmix signal, yielding

$$\hat{\mathbf{Y}}_{ext} = \mathbf{C}\mathbf{Y}\,, \tag{5}$$

where the CPC matrix $\mathbf{C}$ is derived from a first order linear predictor. The CPCs are directly related to the OLDs and IOCs, so there is no need for their explicit transmission, but instead they are entirely retrieved in the SAOC decoder/transcoder.

The error of the linear predictor corresponds to the residual signals,

$$\mathbf{S} = \mathbf{Y}_{aux} - \hat{\mathbf{Y}}_{aux}\,, \tag{6}$$

yielding one residual signal for each EAO, which are finally transmitted within the SAOC bitstream.

3.3. EAO Processing in the SAOC Decoder/Transcoder

The EAO processing is integrated into the regular SAOC decoding/transcoding chain in a cascaded way. In the first step, the EAOs are separated from the downmix signal using the SAOC bitstream information and rendered according to the user-specified rendering parameters to yield the decoder output signals originating from the EAOs. If SAOC transcoding is performed, a downmix of the rendered EAOs is required for the subsequent MPS decoder. Similar to regular SAOC transcoding, these two steps (object decomposition and rendering) are implemented in an efficient integrated way, avoiding a costly intermediate upmix to the discrete EAO signals. Second, the resulting downmix signal containing only regular audio objects is fed into the SAOC decoder/transcoder, which provides either SAOC output or MPS downmix signals. Third, both types of rendered output signals are summed up to yield a mono, stereo or binaural sound scene. In case of a multi-channel output setup, the transcoded downmix signals are summed up to obtain the downmix signal for the subsequent MPS decoding.

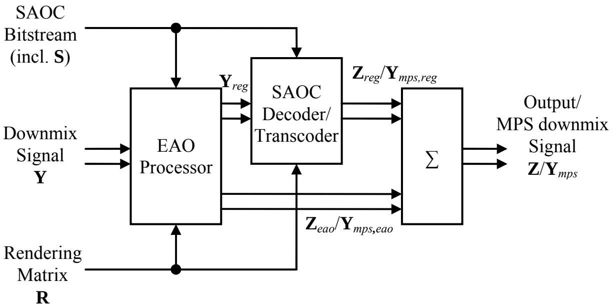

Figure 6: *Basic structure of the SAOC decoder / transcoder with EAO support.*

The conceptual overview of the cascaded EAO decoding/transcoding is depicted in Figure 6 and in the following the corresponding mathematical description of the processing is given. The EAO processor exploits the object parameters provided from the SAOC bitstream to derive exactly the same matrix $\mathbf{C}$ comprising the CPCs as employed in the SAOC encoder, i.e.

$$\mathbf{C} = F\{OLD, IOC\}\,. \tag{7}$$

Considering the residual signals $\mathbf{S}$, which are gathered from the SAOC bitstream, the auxiliary signals are recovered from $\mathbf{S}$, $\mathbf{C}$ and the downmix signal $\mathbf{Y}$ by

$$\mathbf{Y}_{aux} = \mathbf{C}\mathbf{Y} + \mathbf{S}\,. \tag{8}$$

The combination of downmix and auxiliary signals results in the extended signal matrix $\mathbf{Y}_{ext}$. Consequently the EAOs are computed by inverting the extended downmix process, i.e.

$$\begin{bmatrix}\mathbf{Y}_{reg} & \mathbf{X}_{eao}\end{bmatrix}^T = \left(\mathbf{D}_{ext}\right)^{-1}\mathbf{Y}_{ext}\,, \tag{9}$$

where $\mathbf{D}_{ext}$ is derived from the downmix side information. In the absence of signal and parameter quantization and provided $\mathbf{D}_{ext}$ is invertible, this EAO processing approach exhibits perfect reconstruction capability. Effectively the invertibility constraint is ensured in the SAOC encoder by defining an appropriate matrix constellation and SAOC's inherent quantization scheme is designed to render its impact nearly inaudible.

With SAOC decoding, the EAOs are rendered to the desired sound scene according to

$$\mathbf{Z}_{eao} = \mathbf{R}_{eao}\mathbf{X}_{eao}\,, \tag{10}$$

where matrix $\mathbf{R}_{eao}$ comprises the EAO rendering information. For a multi-channel loudspeaker setup, a reduced rendering matrix $\mathbf{R}_{<1,2>,eao}$ is derived from $\mathbf{R}_{eao}$ comprising one or two channels (depending on the number of downmix channels) to yield the EAO part of the MPS downmix signal,

$$\mathbf{Y}_{mps,eao} = \mathbf{R}_{<1,2>,eao}\mathbf{X}_{eao} . \qquad (11)$$

The output / MPS downmix signals of the regular audio objects, $\mathbf{Z}_{reg}$ / $\mathbf{Y}_{mps,reg}$, are obtained from the SAOC decoder/transcoder and added to the appropriate signals from the EAO processor to result in the final mono, stereo or binaural output signal

$$\mathbf{Z} = \mathbf{Z}_{reg} + \mathbf{Z}_{eao} , \qquad (12)$$

and respectively in the MPS downmix signals for subsequent multi-channel MPS decoding, i.e.

$$\mathbf{Y}_{mps} = \mathbf{Y}_{mps,reg} + \mathbf{Y}_{mps,eao} . \qquad (13)$$

In a typical Karaoke playback application, $\mathbf{R}_{eao}$ or $\mathbf{R}_{<1,2>,eao}$ contain only zero elements, i.e. the vocal object is muted. The background sound scene corresponds to the regular downmix signal modified by one common gain factor for the regular audio objects, retaining their spatial position within the downmix channels unchanged within the output sound scene. The solistic representation of one or a few EAOs is implemented by setting the rendering matrix elements of all regular audio objects to zero, while full rendering flexibility (in terms of spatial position and level modification) is preserved for the EAOs. Besides these two special application situations, the proposed system supports regular SAOC decoding/transcoding too, i.e. the EAO processor is de-activated and the SAOC downmix signal is directly fed into the SAOC decoder/transcoder. So if a user performs remixing avoiding extreme boosting/suppressing of distinct audio objects, (i.e. the relative rendered object powers do not differ significantly from one another), EAO processing can simply be switched off. To this end an actual hard- or software application supporting an enhanced SAOC decoder/transcoder could be equipped with classical solo/mute buttons (just as mixing consoles feature) or a specific “Karaoke/solo” button to enable EAO processing providing improved output sound quality.

The additional computational complexity required for EAO processing in the SAOC decoder/transcoder depends mainly on the number of residual signals[1]. For the MPEG SAOC specification considering a maximum of four EAO signals, the complexity increase is less than one third of the SAOC decoder’s overall computational load. However, in the most typical application scenarios comprising only one mono or stereo EAO this value is much lower.

4. SUBJECTIVE EVALUATION

The improvement in terms of perceptual audio quality of the SAOC system enhanced by the EAO processor described above has been evaluated. This section describes the design and results of one of the subjective listening tests conducted within the scope of the recent standardization activities in the MPEG audio group. Specifically, this test aimed at proving the significant audio quality increase of the enhanced over the regular SAOC system for the two application scenarios Karaoke and solistic vocal playback.

The stereo downmix signals were coded with an AAC core-coder at the commonly used bitrate of 128 kBits per second. A bitrate of 20 kBits per second has been selected for coding the residual signal. In previous subjective listening tests comparing the audio quality of different bitrate settings, this value has been established to yield the best compromise between processing performance and additional data rate consumption.

4.1. Test methodology and design

The tests were arranged in an acoustically isolated listening room ensuring a high-quality listening environment according to ITU recommendation ITU-R BS. 1116-1, specified in [17]. In order to provide optimal means for the evaluation process, the playback of the test stimuli was performed with high-quality headphones. The test method followed the standard procedures used in spatial audio verification tests, based on the “Multiple Stimulus with Hidden Reference and Anchors” (MUSHRA) methodology for the subjective assessment of intermediate audio quality [18]. A total of 9 listeners having wide experience with perceptual listening tests participated in the test.

A set of critical items was selected from typical Karaoke/solo audio material to compare the potential of the EAO processor to that of regular SAOC decoding. The stereo rendering settings were chosen according to the respective reproduction scene. The following two systems have been compared in the test:

1) the enhanced SAOC system with EAO processing (denoted “with EAO proc.” in Figure 7),
2) the regular SAOC decoder (denoted “reg. SAOC” in Figure 7).

Furthermore, the test included 3) a hidden reference signal (“Hidden Ref.”) and 4) a 3.5 kHz low-pass version of the reference signal serving as lower anchor (“Low anchor”). In accordance with the MUSHRA methodology, the listeners were instructed to compare the four systems against the (known) reference. The subjective responses were recorded on the five grade MUSHRA scale. The test conditions were randomized automatically for each test item and for each listener. An instantaneous switching between the systems under test was allowed.

4.2. Listening test results

Figure 7 shows the average MUSHRA grading over all listeners per item and the statistical mean value of all evaluated items together with the associated 95% confidence intervals. For each audio item the results reveal a significantly better performance of the SAOC decoder when it is equipped with the EAO processor. The MUSHRA scores for Karaoke playback are in the range of “excellent” to “good” and for solo reproduction they are slightly below but still clearly graded “good”. These results prove that EAO processing leads to a considerable improvement of the audio signal quality for the considered demanding application scenarios. (The same behaviour has been observed for the SAOC transcoder with a separate listening test.)

[1] In audio coding the computational complexity is commonly specified only for the decoder/transcoder processing.

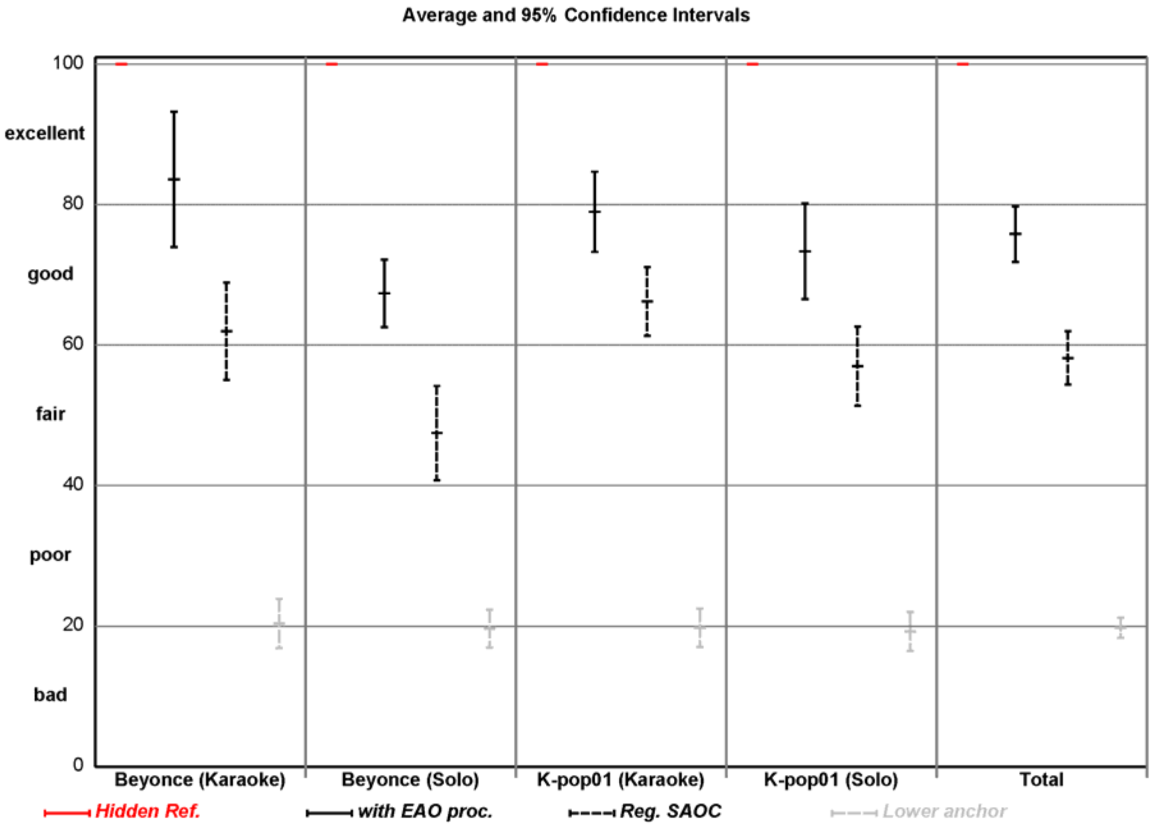

Figure 7: *Average MUSHRA scores of the subjective listening test.*

5. CONCLUSION

An enhanced decoder/transcoder processing for SAOC has been presented, that is dedicated to improving the performance of the SAOC system in demanding operating conditions comprising extreme relative level modifications of the audio objects in the rendered output scene. The introduced parametric processing scheme allows for a better separation of specific enhanced audio objects (EAOs) from the downmix signal by employing a predictive model for each EAO in the SAOC decoder/transcoder. Considering the prediction error signals yields an additional important gain in perceptual audio quality at the expense of a moderate increase in side information to be transmitted. Nevertheless the enhanced SAOC system still ensures an efficient representation of multiple audio objects compared to the multitude of discrete audio object signals requiring considerably higher bitrates.

Due to the significant increase in perceptual audio quality, having been proved by subjective listening tests, the proposed EAO processing scheme has successfully been included into the MPEG SAOC International Standard specification. Consequently, in the presence of residual signals in the SAOC bitstream, a standards compliant SAOC decoder/transcoder automatically activates EAO processing. If no residual information is provided by the SAOC encoder, regular SAOC decoding/transcoding is performed.

6. REFERENCES

[1] J. Breebaart, and C. Faller, *Spatial Audio Processing*, J. Wiley & Sons, 2007.

[2] J. Engdegard, 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, Preprint 7377.

[3] C. Faller, F. Baumgarte, “Efficient Representation of Spatial Audio Using Perceptual Parameterization”, *Proc. WASPAA*, New Paltz, New York, Oct. 2001, pp. 199-202.

[4] F. Baumgarte, and C. Faller, “Binaural Cue Coding – part I: Psychoacoustic Fundamentals and Design Principles”, *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 509-519, Nov. 2003.

[5] C. Faller, and F. Baumgarte, “Binaural Cue Coding – part II: Schemes and applications”, *IEEE Trans. on Speech and Audio Proc.*, vol. 11, no. 6, pp. 520-531, Nov. 2003.

[6] H. Purnhagen, “Low complexity parametric stereo coding in MPEG-4”, *7th Int. Conf. on Audio Effects (DAFX-04)*, Naples, Italy, Oct. 2004.

[7] E. Schuijers, J. Breebaart, H. Purnhagen, and J. Engdegård, “Low complexity parametric stereo coding”, *116th AES Conv.*, Berlin, Germany, May 2004, Preprint 6073.

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

[9] J. Herre, and S. Disch, “New Concepts in Parametric Coding of Spatial Audio: from SAC to SAOC”, *Proc. IEEE Int. Conf. on Multimedia and Expo*, Beijing, China, 2007.

[10] ISO/IEC, *MPEG audio technologies – Part 1: MPEG Surround*, ISO/IEC Int. Standard 23003-1:2007, 2007.

[11] ISO/IEC, *MPEG audio technologies – Part 2: Spatial Audio Object Coding*, ISO/IEC Int. Standard 23003-2.2:2010, 2010.

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

[13] M. Dietz, L. Liljeryd, K. Kjörling, O. Kunz, “Spectral Band Replication, a novel approach in audio coding”, *112th AES Conv.*, Munich, Germany, May 2002, Preprint 5553.

[14] M. Wolters, K. Kjörling, D. Homm, and H. Purnhagen, “A closer look into MPEG-4 High Efficiency AAC”, *115th AES Conv.*, New York, USA, October 2003, Preprint 5871.

[15] G. Hotho, L. Villemoes, J. Breebaart, “A Backward-Compatible Multichannel Audio Codec”, *IEEE Trans. on Audio, Signal and Language Proc.*, vol. 16, no. 1, pp. 83-93, Jan. 2008.

[16] 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”, *J. Audio Eng. Soc.*, vol. 45, no 10., pp. 789-814, Oct. 1997.

[17] Int. Telecommunication Union, *Methods For The Subjective Assessment Of Small Impairments In Audio Systems Including Multichannel Sound Systems*, ITU-R BS. 1116-1, 1994-1997.

[18] EBU Technical recommendation, *MUSHRA-EBU Method for Subjective Listening Tests of Intermediate Audio Quality*, Doc. B/AIM022, October 1999.