INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2015/N15820
October 2015, Geneva, Switzerland

Title **Text of White Paper on MPEG Technology: Spatial Audio Object Coding**
Source: **Audio Subgroup**
Status **Approved**

Introduction

This output document proposes the text for a White Paper on MPEG Technology: Spatial Audio Object Coding (SAOC).

SUMMARY

What it does:
It allows highly efficient storage and transport of individual audio objects (e.g. voices, instruments, ambience, etc.) in an audio mix, while preserving the possibility for the listener to adjust the mix based on his personal taste. That includes changing the rendering configuration of the audio scene from mono or stereo over surround to even binaural reproduction.

What it is for:
Interactive audio mixing (e.g. Karaoke or personalized audio channels in broadcasting) or highly flexible and efficient teleconferencing solutions are the most prominent applications for this technology.

Where:
ISO/IEC 23003-2, ISO/IEC 23003-2:2010/Amd.3, ISO/IEC 23008-3

Technology

Introduction

MPEG-D Spatial Audio Object Coding (SAOC) is an audio coding algorithm which allows highly efficient storage and transport of individual audio objects (e.g. voices, instruments, ambience, etc.) in an audio mix, while preserving the possibility for the listener to adjust the mix based on his personal taste. That includes changing the rendering configuration of the audio scene from mono or stereo over surround to even binaural reproduction.

MPEG-D SAOC [1] was standardized between 2007 and 2010, following the standardization of MPEG-D MPEG Surround [2].

Motivation

MPEG Surround technology supports very efficient parametric coding of multi-channel audio signals by transmitting a backward compatible downmix of the *multi-channel signal* together with some small amount of side information characterizing the spatial sound image. The downmix signal can be encoded by known perceptual audio coding techniques such as AAC (Advanced Audio Coding). In contrast, the idea of MPEG-D SAOC is to apply similar basic assumptions together with a similar parameter representation for very efficient parametric coding of *individual audio objects* (tracks). Additionally, a rendering functionality is included to interactively render the audio objects into an acoustical scene for several types of reproduction systems (mono, stereo, or 5.1 for loudspeakers or binaural for headphones).

Overview

MPEG-D SAOC is designed to transmit a number of audio objects in a joint mono or stereo downmix signal to later allow a reproduction of the objects in an interactively rendered audio scene. For this purpose, in addition to creating the downmix signal, an SAOC encoder creates a parametric description of the perceptually relevant properties of the audio objects using Object Level Differences (OLD) and Inter Object Correlation (IOC), and describes the downmixing process with Downmix Channel Level Differences (DCLD) and Downmix Gains (DMG). These descriptions are encoded into a parameter bitstream. In contrast to MPEG Surround, this parametrization is independent from the output loudspeaker configuration at the decoder. Both the downmix signal and the SAOC parameter stream are transmitted to the decoder. The downmix signal may be represented efficiently using perceptual audio coding techniques. The SAOC decoder uses the parametric data for allowing interactive manipulation of the output audio scene for the user. For monophonic, stereo or binaural reproduction, the SAOC decoder directly produces the desired output (“decoding mode”). For multi-channel reproduction, the SAOC decoder converts the SAOC parameter representation into an MPEG Surround parameter representation (“transcoding mode”), which is then decoded together with the downmix signal by an MPEG Surround decoder to produce the desired audio scene. In all cases, the user interactively controls this process to alter the representation of the audio objects in the resulting audio scene. This typically entails control over the level and the spatial position of each object, but also other changes are possible.

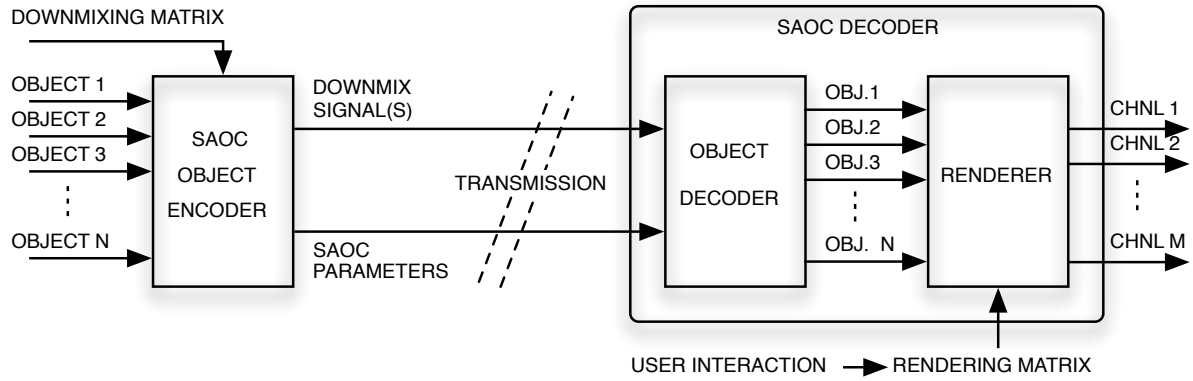

Figure 1: Basic principle of SAOC ("decoding mode"). Conceptually, the decoder first attempts to reconstruct the original objects and then to render them into an output scene. In practice, both aspects are combined into a single efficient processing step.

Similarly to MPEG Surround, the SAOC downmix signal can be transmitted as audio signal over existing mono or stereo distribution channels. Adding the SAOC side information in a backward compatible way enables SAOC-equipped decoders to interactively influence the output sound scene, while legacy decoders continue to work (but do not provide interactivity).

Target applications

Among the numerous conceivable applications for SAOC, a few typical scenarios are listed here.

Interactive Mixing: Consumers can create personal interactive remixes using a virtual mixing desk. For example, certain instruments can be attenuated for playing along (similar to Karaoke), the original mix can be modified to suit personal taste, etc.

Interactive Gaming: For interactive gaming, SAOC is a storage and computationally efficient way of reproducing sound tracks. Moving around in the virtual scene is reflected by an adaptation of the object rendering parameters. Networked multi-player games benefit from the transmission efficiency using one SAOC stream to represent all sound objects that are external to a certain player's terminal.

Enhanced Telecommunication: Current telecommunication infrastructure is monophonic and can be extended easily in its functionality. Terminals equipped with an SAOC extension pick up several sound sources (objects) and produce a monophonic downmix signal, which is transmitted using the existing (speech) coders. The side information can be conveyed in an embedded, backward compatible way. Legacy terminals will continue to produce monophonic output while SAOC-enabled ones can render an acoustic scene and thus increase intelligibility by spatially separating the different talkers ("cocktail party effect"). For the use in telecommunication applications, Low Delay SAOC (SAOC-LD) technology is available.

Dialogue Enhancement: In the context of broadcasting, amplifying the dialogue relative to the background sound has been recognized as an important feature in order to provide a clean dialog service for hearing-impaired listeners or for broadcast reproduction in noisy environments. Alternatively, it may be desirable to, e.g., boost sports stadium atmosphere relative to the commentator to enhance the listener's feeling of being at the stadium. These use cases are covered by an amendment of the MPEG-D SAOC specification [3] called SAOC-DE (SAOC for Dialogue Enhancement) [4]. While the underlying concepts are the same as in regular SAOC, SAOC-DE is adapted to the specific scenario: the processing is done in an "in-place" fashion,

that is the SAOC-DE downmix and output have the same number of audio channels and correspond directly to one another. Only gain changes of foreground (e.g., dialogue) and background objects are supported (i.e. no change of spatial position) which allows a very efficient implementation. Additionally, the number of natively supported downmix and output channels is increased, so no MPEG Surround processor is required even for processing surround sound.

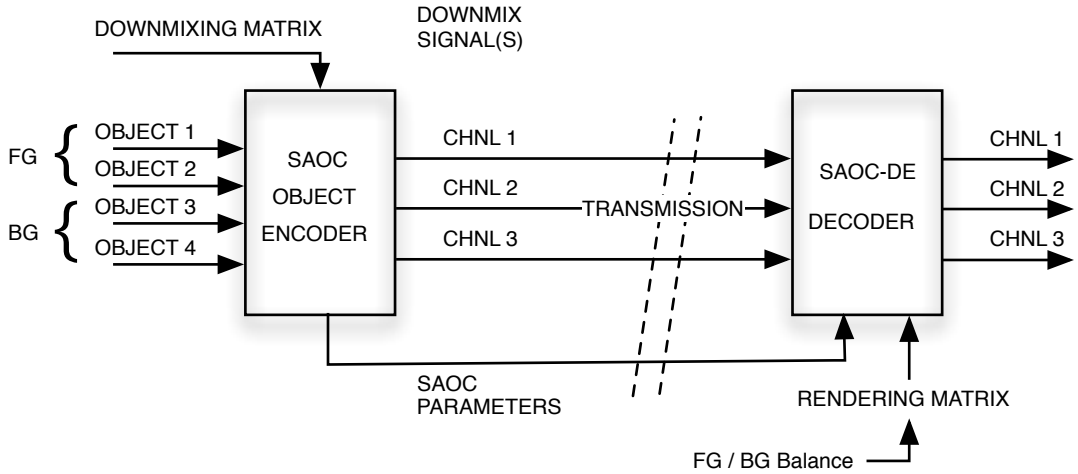

Figure 2: Basic principle of SAOC-DE processing (example: two foreground (FG) objects, two background (BG) objects, three audio channels). The user interaction controls the foreground / background balance.

SAOC-DE became a part of the DVB specification TS101154 (v2.2.1) as an optional tool for "Advanced Clean Audio" [5].

3D Audio: In addition to the previously described use cases, SAOC technology is also employed to serve as a technology component within the ISO/MPEG-H 3D Audio standard [6] to facilitate efficient parametric coding of object and channel signals. To this end, the original Spatial Audio Object Coding codec has been enhanced into "SAOC 3D" [7] with the following extensions:

- While MPEG-D SAOC supports only up to 3 downmix channels, MPEG-H SAOC 3D supports arbitrary number of downmix channels.
- While rendering to multi-channel output has been possible with MPEG-D SAOC only by using MPEG Surround as a rendering engine, SAOC 3D performs direct decoding/rendering to multichannel/3D output with arbitrary output speaker setups.
- Some MPEG-D SAOC tools that have been found unnecessary within the MPEG-H 3D Audio system have been excluded.

Additionally, SAOC 3D specifies different decoding modes for efficient processing of a large number of channels and dynamic objects.

Legacy compatibility

A key aspect of MPEG-D SAOC technology is that the transmitted object downmix (e.g. stereo) is a high-quality audio signal, which can be sent using the existing distribution services. In this way, existing receivers/decoders will continue to provide a high-quality output. Transport of additional SAOC side information can be done in a way that it is hidden from legacy receivers. This allows SAOC capable receivers/decoders to provide interactive sound output for advanced services. In case an artistic downmix signal is available at the encoder side, it may be the

preferred presentation of the downmix signal to be transmitted instead of the encoder-created downmix signal. Hence, considering the quality of the default output, the legacy receivers are in no way disadvantaged relative to MPEG-D SAOC capable receivers.

SAOC performance

To illustrate SAOC performance, MPEG conducted two formal verification tests according to MUSHRA test methodology for relevant application areas of SAOC, namely teleconferencing and interactive remix cases. Listening tests for both application cases included several items representing typical audio content for the considered scenarios. Approximately 30 expert listeners participated in the subjective tests in which headphone and stereo loudspeaker reproduction setups were used.

The teleconferencing use-case test simulated a real-world audio conferencing application scenario with several active participants and a varying degree of double-talk, as illustrated in Figure 3.

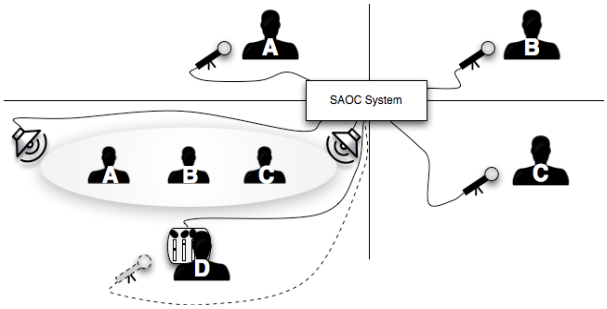

Figure 3: Conceptual scheme for the teleconferencing use-case with an SAOC system allowing participant D to control spatial position and level of the voices of the remote participants A, B, and C.

To satisfy realistic delay and bitrate requirements, the SAOC-LD ("low delay") mode with a mono AAC-ELD core coder has been tested against AAC-ELD based legacy technology, which separately encodes each of the four mono objects. It can be seen from the listening test results shown in Figure 4, that the SAOC technology delivered audio quality comparable to legacy technology while requiring less than a half of the bitrate and providing additional functionality (interactivity).

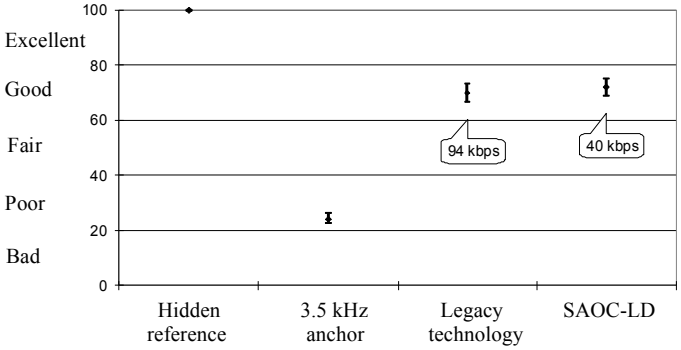

Figure 4: MUSHRA listening test results (mean values and 95% confidence intervals) for the teleconferencing scenario.

As a second test, the interactive remix use-case test simulated adjustments to a given downmix of audio objects. Such adjustments could be applied by a user having freedom to create his personalized remix of a song by using an SAOC rendering interface, see Figure 5.

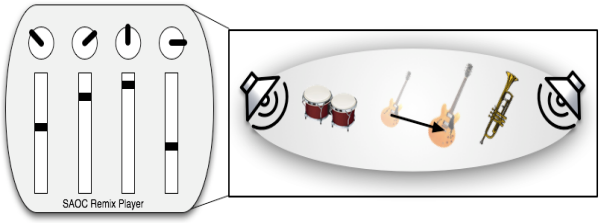

Figure 5: Conceptual scheme for the interactive remix use-case with an SAOC system allowing controlling panning position and level of the individual instruments of the song.

The listening test evaluated performance of both High Quality (i.e. the regular) and Low Power decoding modes with a stereo HE-AAC core coder. The performance of SAOC system has been compared with HE-AAC based legacy technology, which uses separately encoded objects (4 mono or 3 stereo objects) under the same total bitrate constraint. To illustrate the influence of the core coder alone on audio quality, an HE-AAC encoded reference signal was also included in the test. It can be seen from the listening test results shown in Figure 6, that both SAOC decoding modes outperform the legacy technology by approximately 20 MUSHRA points.

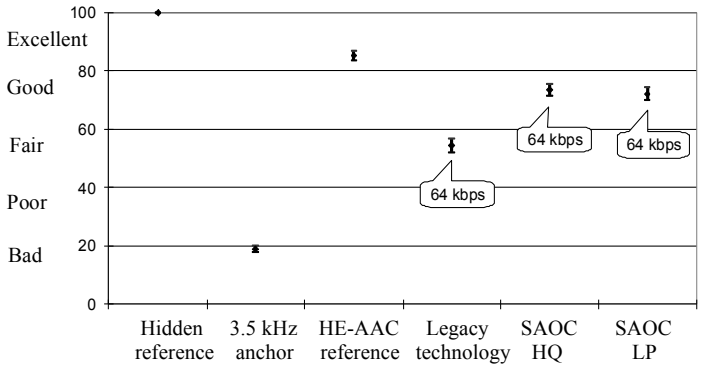

Figure 6: MUSHRA listening test results (mean values and 95% confidence intervals) for the interactive remix scenario.

Conclusions

MPEG Spatial Audio Object Coding (SAOC) is a modern and efficient technology for bitrate efficient and backward compatible representation of audio scenes containing audio objects. Many applications can benefit from this technology, including personalized / interactive sound rendering, advanced telecommunication and networked gaming. Specifically, for the purpose of dialog enhancement in broadcasting, the SAOC-DE specification has been created. SAOC technology is also utilized in as one component (SAOC 3D) for MPEG-H 3D Audio Coding.

References

1. J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilpert, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, M. L. Valero, B. Resch, H. Mundt, and H. Oh: "MPEG Spatial Audio Object Coding – The ISO/MPEG Standard for Efficient Coding of Interactive Audio Scenes", Journal of the AES, Vol. 60, No. 9, September 2012, pp. 655-673.
2. J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén. W. Oomen, K. Linzmeier, K. S. Chong: "MPEG Surround – The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding", Journal of the AES, Vol. 56, No. 11, November 2008, pp. 932-955.
3. ISO/IEC 23003-2:2010/Amd.3 Information technology – MPEG audio technologies – Part 3: Spatial Audio Object Coding (SAOC) Amendment 3: Dialogue Enhancement. 2014.
4. J. Paulus, J. Herre, A. Murtaza, L. Terentiv, H. Fuchs, S. Disch, F. Ridderbusch: "MPEG-D Spatial Audio Object Coding for Dialogue Enhancement (SAOC-DE)", 138th AES Convention, Warsaw, Poland, 2015, Paper 9220.
5. ETSI TS 101 154 v2.2.1, Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream. 2015.
6. ISO/IEC JTC1/SC29/WG11 International Standard ISO/MPEG 23008-3, 3D Audio, Geneva, February 2015.
7. A. Murtaza, J. Herre, J. Paulus, L. Terentiv, H. Fuchs, S. Disch: "ISO/MPEG-H 3D Audio: SAOC 3D Decoding and Rendering", 139th AES Convention, New York, USA, 2015, Paper 9434.