

Meeting the Requirements of Next-Generation Broadcast Television Audio

Robert Bleidt (Fraunhofer USA DMT), Harald Fuchs,
Stefan Meltzer, Stephan Schreiner (Fraunhofer IIS),
Skip Pizzi (Consultant)

Abstract

The requirements for the audio codec of the next-generation broadcast television system are multifaceted. The audio codec needs to support multichannel audio, 3D audio extension, parallel transmission of different language tracks, additional channels for the hearing and visually impaired, and the transport of metadata containing, for instance, dynamic range or downmix parameters. To transmit all this data at the bit rates available for broadcast transmissions, the audio codec has to be very efficient. It should also be as flexible as possible to enable future extensions of the audio subsystem. And it would be desirable to use the same audio codec for fixed and mobile services.

The first part of the paper will discuss the different approaches to fulfill the previously listed requirements. The different formats for the 3D audio extension currently under discussion will be presented and analyzed with respect to the resulting requirements for the audio codec system. The support of special services for hearing and visually impaired people plays a more prominent role today due to legislative initiatives. While the additional descriptive audio channel for visually impaired persons has already become a standard feature of today's broadcast, the solution for the support of hearing impaired people through improving speech intelligibility is still under discussion. Different approaches will be described, and the pros and cons discussed.

The second part will describe how the requirements can be handled by the MPEG-4 HE-AAC audio coding system. Based on independent tests, the performance and efficiency of the codec will be demonstrated, along with an estimation of the bit rates required to deliver all the new features.

Introduction

The next-generation TV system shall be capable of handling all requirements for the future TV broadcast as they are known or anticipated today, as well as being prepared to the extent possible for new requirements arising in the future. To fulfill this demand, certain flexibility needs to be built into the system. It is the task of system designers to find the right balance between necessary flexibility and resulting additional complexity of the system.

As a basis for such decisions, a thorough analysis of the possible requirements is critical. Some aspects like the demographic changes of the population are well known. Others like the changes in user behavior are monitored closely, but it is impossible to define a long-term trend, since recent history has shown many unexpected changes. Another aspect is how TV broadcasting will fit into the future media distribution landscape and conform to user expectations. Although the transmission component plays a significant role in the overall system, final user experience is mainly dominated by the capabilities of the audio and video subsystems. This paper focuses on the audio part.

Regarding demographics, one trend clearly defines the need for the audio subsystem to improve the

situation for hearing-impaired users. To a certain extent, this is also true for visually impaired users. Therefore the audio subsystem needs to offer improvements for both groups.

In terms of user experience, the biggest recent trend in video is 3D-TV. One could argue that with its 5.1 surround sound capability, DTV has already offered a “3D” aural experience for some time. Nevertheless, it may be expected by users that the audio experience will be improved to match any new 3D video experience. Other improvements in this area involve greater interactive user-manipulation of TV audio. Putting these elements together might lead to a completely new way of representing and transmitting audio, leaving the existing channel-based audio formats behind.

To summarize, the requirements for next-generation TV audio are as follows:

- Improved service for hearing and visually impaired people
- 3D audio extension
- Multilingual support
- Enhanced user experience

Improved services for hearing and visually impaired users

Support for visually impaired users is currently realized by an additional descriptive audio channel, which is mixed at the receiver with main program audio at the user’s option. This solution currently works well enough, although one improvement might involve additional spatial guidance to locate the descriptive narration closer in the sound field to the event which is described at the moment. The effectiveness of this capability for visually impaired users requires further investigation, however.

Other improvements in this regard could allow users to simply place the descriptive narration at a fixed point of their preference within the sound field, or to route the narration channel through a separate output, such that the visually impaired user might hear the description via an earpiece, for example, while others viewing the program in the same room do not hear the descriptions.¹

In the case of hearing impairment, one needs to distinguish between users with complete loss of hearing versus those with gradually increasing hearing loss. The second group is much larger and is growing rapidly with the increasing age of the population. The only solutions for profoundly deaf users are the already well-established closed captioning techniques. On the other hand, for users with partial or increasing hearing loss, the improvement of dialog intelligibility is a more appropriate solution, and one that could benefit from technical improvements in next-generation systems. One method to address this is raising the relative volume of the center channel while reducing all other channels’ relative volume. Generally, this approach is only successful when the program’s audio is mixed with all its dialog elements (and no other sound elements) in the center channel. While this is the case for most cinematic productions, it is not always the approach followed for TV programs.

Another problem with this approach arises from the occasional increase in ambience that may occur when the center-channel audio level is increased. This effect can actually reduce intelligibility. An alternative approach is to transmit the “dry” dialog separately from any other elements in the complete mix, so the receiver could have access to pure dialog content separately, and recreate the full mix via adding the dialog at the receiver. This would allow users to adjust the level of the dialog (only) exactly according to their needs. (The receiver default setting would be a unity-gain setting

¹These improvements don’t require any changes in the transmission standard, but can simply be added to the receiver. They are mentioned here for completeness.

that restores the program’s original mix.) This method could also increase the transmission bandwidth requirement by adding from one to as many of five full-range channels of audio to the program.²

An alternative approach is the use of Spatial Audio Object Coding (SAOC). This technology is part of the MPEG-D standard.³ The basic idea of SAOC is to treat the different components of an audio signal (for example, the vocals and different instruments in a song recording) as separate “objects”. During the production process, prior to the creation of a final program audio mix, certain perceptually relevant properties of each individual audio object are captured and represented by a compact set of parameters. These parameters are transmitted in a low data-rate channel alongside the program’s full audio mix. The end user can then apply the parametric signal to the audio mix, thereby synthesizing a flexibly rendered audio scene, independent of the system’s audio coding or multichannel configuration. The data rate required for each individual audio object is in the range of 2-3 kbps. In the accessibility application discussed here, it is only necessary to transmit a single audio object representing the dialog element of the audio mix (in whatever channel or spatial orientation that dialog appears, since the object can include steering information). SAOC-capable receivers could then use the additional data channel to extract the pure dialog signal from the audio mix, and manipulate it as desired (e.g., increase its volume or change its equalization). This approach would be backward compatible with legacy receivers, since the full audio mix is transmitted in normal fashion. The use of SAOC is a good solution to this problem, since it offers full flexibility to adapt the pure dialog audio to users’ needs, with only marginal increase to the transmitted data rate, and without adding untenable complexity to the production process.

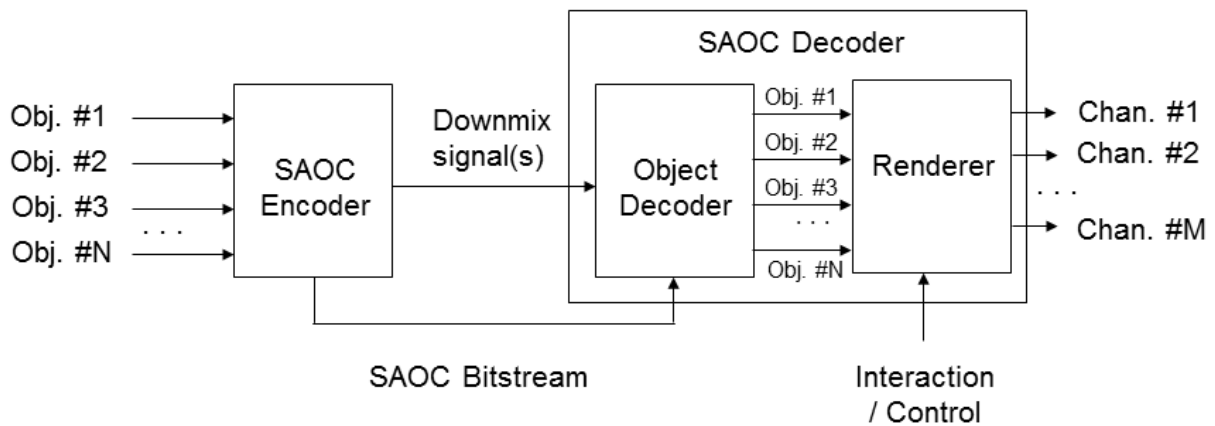


Figure 1: SAOC Conceptual Overview

² Dialog is often contained in only the center channel, but it may also appear in any other channel (other than LFE). To fully implement this approach, a worst-case scenario would add five more “dry” dialog channels to be mixed in the receiver. This could create untenable production and transmission requirements.

³ [ISO/IEC 23003-2:2010](https://www.iso.org/obp/ui/#iso:code:38100:23003-2:2010), MPEG-D Part 2 – Spatial Audio Object Coding

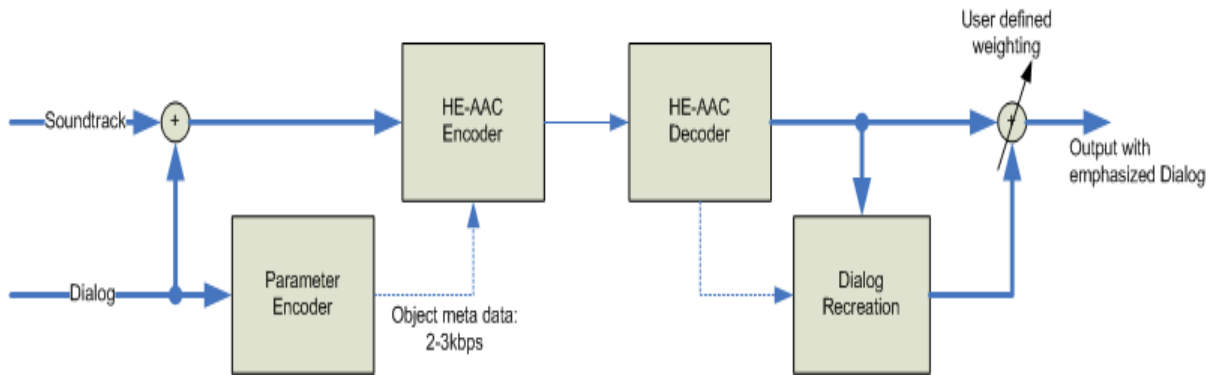


Figure 2: Improved Dialog Reproduction using Spatial Audio Object Coding

3D Audio

3D is currently a hot topic in TV, but relates nearly exclusively to the video side of the industry. Nevertheless, some associated discussions are currently taking place on the audio side. Several proposals are on the table with varying numbers of channels and locations of speakers. The number of channels being considered ranges from 6.1 to 22.2. For a television broadcast standard, it is likely that only one channel configuration would be preferred. One could argue that a configuration should be selected that allows derivation of other configurations from it, just as today's situation in which 5.1 multichannel is downmixed to stereo. Although this might be possible to a certain extent in an envisioned next-generation system with more audio channels, some configurations might be too different from one another to transfer audio signals between them without intolerable loss of audio quality. Another problem is user acceptance of an increased number of speakers. From the introduction of 5.1 multichannel audio it is known that many users haven't converted from stereo to multichannel audio because of the space requirements for the additional speakers and the cabling effort, as well as for aesthetic reasons.

A completely different approach is to leave the concept of channel-based audio production behind, and use an object-based approach instead. In this solution, a number of audio objects are described, along with a scene description, and these parameters are transmitted to the receiver, which renders the audio signal in accordance with the physical audio reproduction setup in the user's environment. This solution will provide for all possible setups, including perhaps those not yet envisioned at the time of the system's introduction. The optimal solution would support reproduction techniques like wave field synthesis or ambisonics. Of course this approach requires a new production technology since the audio is no longer mixed to one target configuration, but instead each audio object is managed separately. Early production tools for this kind of authoring are already available. It might be even worthwhile for producers to consider the use of such tools today for "future-proofing" archival purposes, since an object-based scene description allows the creation of a sound mix for any possible spatial configuration. One question in this context is the number of audio objects required to describe complex scenes. The experience in using production method for wave field synthesis shows that it is sufficient to use a maximum of 16 audio objects simultaneously. This means that the bit rate requirement is roughly equivalent to a 16 channel audio signal.

Multilingual Support

Multilingual support is already a must in today's TV broadcasts. It is currently solved by transmitting the complete sound track in all required languages in parallel. Using an object-based approach, it would be possible to send a single, full audio mix (with or without a default dialog language included), along with dialog objects for all available languages for the program. The user can then select which language to use, and the complete soundtrack can be rendered by the receiver.

Enhanced User Experience

Enhancing the user experience has a necessarily broad scope. Besides canonical audio enhancements like multichannel surround, new features such as interactive audio are possible. Interactive audio could allow the user to manipulate the audio not only in terms of volume and equalization, but also to actively remix the audio. For example, this allows the user to select a preferred listening position in the audience during a concert broadcast, or to change the audio mix in a way that it is more appealing to children's or other users' preferences. This can be achieved by adding metadata to the audio objects, which steers the rendering process based on filters set by the user. Another dimension of audio interactivity could emerge if the video representation also becomes object based.

MPEG-4 HE-AAC

From the above discussion, requirements for a next-generation audio codec can be derived. The codec needs to support a large number of independent audio channels, and has to provide the flexibility to carry additional metadata. Because the number of audio channels will likely increase, coding efficiency also plays an important role. Ideally, the codec should be able to be used for all TV transmission scenarios – satellite, terrestrial, cable, IP-based and mobile.

MPEG-4 High Efficiency Advanced Audio Coding (HE-AAC) fulfills all these requirements. It is already widely used for these purposes. In Europe and South America, DTV services using HE-AAC stereo and multichannel audio are on air, and nearly all the world's mobile TV standards are based on HE-AAC. Due to its efficiency, an increasing number of IP-based services are also switching to HE-AAC.

HE-AAC is today's most efficient high-quality audio codec. Independent tests conducted by the European Broadcasting Union in 2007 have shown that HE-AAC can produce an excellent audio quality for a 5.1 surround sound signal at a bit rate of 160 kbps. Figure 3 shows the summary of the test results.

HE-AAC's format allows it to carry up to 48 independent audio channels. The use of ancillary data fields within the bitstream offers the required flexibility to carry additional metadata. This flexibility helps to future-proof the codec, since it allows the addition of new features while maintaining backward compatibility. Existing decoders will ignore information in the ancillary data fields that they cannot decode. Therefore MPEG-4 HE-AAC is an appropriate choice for a next-generation TV system's audio codec.

And although the ATSC's NGBT planning process does not require backward compatibility, HE-AAC's current use as the audio codec for ATSC M/H allows extension to any next-generation system with a measure of compatibility to the most recently added elements of current-generation ATSC broadcasts.

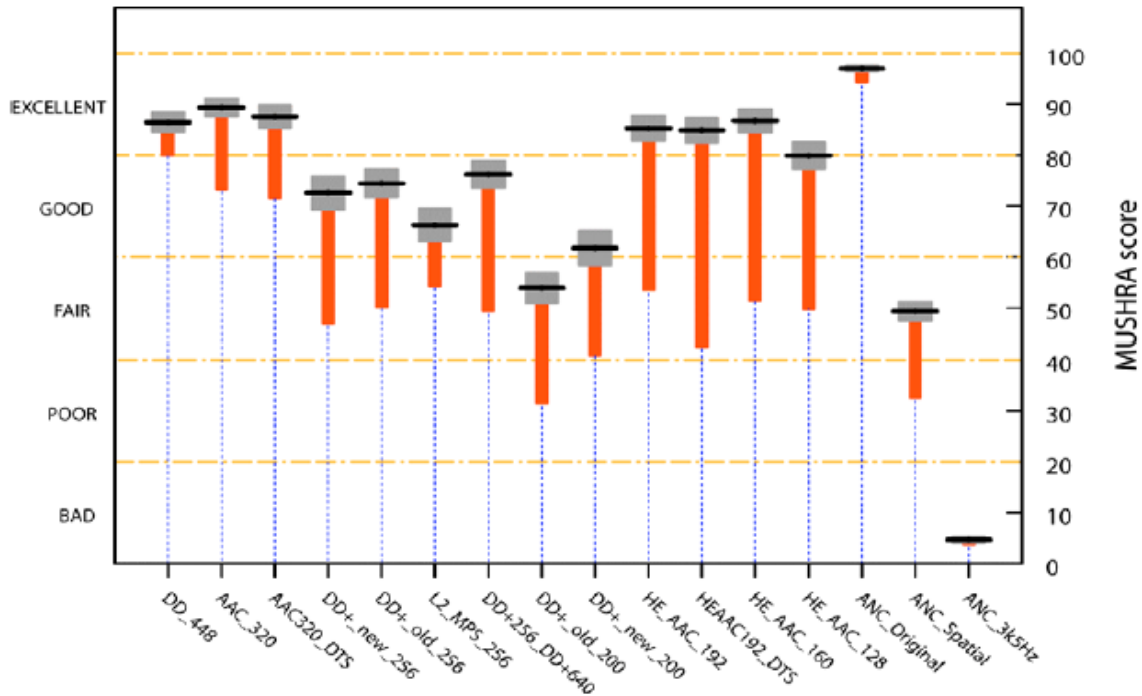


Figure 3: Test Results of EBU's Multichannel Audio Codec Evaluation (2007)

Summary

This paper lists the requirements for the audio part of the next-generation TV system. With the object-based approach a new concept is proposed, which can fulfill all requirements listed. Although it requires changes in the way TV audio is produced today, it is worthwhile to investigate whether it should be introduced. The MPEG SAOC technology offers a solution that introduces some object-oriented aspects into a channel-based audio solution, as a method of providing a number of expected requirements in next-generation television broadcasting. (SAOC could also incrementally improve conditions for hearing-impaired users in current-generation DTV.) Finally, the paper shows that the MPEG-4 HE-AAC codec can satisfy all envisioned audio requirements for the next-generation TV system, whether a channel- or object-based representation of the audio subsystem is selected.

References

- [1] [ISO/IEC 23003-2:2010](#), Information technology -- MPEG audio technologies -- Part 2: Spatial Audio Object Coding (SAOC), MPEG-D.
- [2] [ISO/IEC 14496-3:2009](#), Information technology -- Coding of audio-visual objects -- Part 3: Audio, MPEG-4.
- [3] [EBU-Tech 3324](#), EBU Evaluations of Multichannel Audio Codecs, Geneva, September 2007.
- [4] Jonas Endegard, Barbara Resch, Cornelia Falch, Oliver Hellmuth, Johannes Hilpert, Andreas Hoelzer, Leonid Terentiev, Jeroen Breebaart, Jeroen Koppens, Erik Schuijers, Werner Oomen. [“Spatial Audio Object Coding \(SAOC\) – The Upcoming MPEG Standard on Parametric Object Based Audio Coding.”](#), in AES 124th Convention, Amsterdam, Netherlands, May 2008.