Object-Based Audio: Opportunities for Improved Listening Experience and Increased Listener Involvement

Robert Bleidt

Fraunhofer USA Digital Media Technologies, San Jose, CA, codecs@dmt.fraunhofer.org

Arne Borsum

Fraunhofer IIS, Erlangen, Germany, arne.borsum@iis-extern.fraunhofer.de

Harald Fuchs

Fraunhofer IIS, Erlangen, Germany, harald.fuchs@iis.fraunhofer.de

S. Merrill Weiss

Merrill Weiss Group LLC, Metuchen, NJ, merrill@mwgrp.com.

Written for SMPTE Journal©

Abstract. *A new TV audio system based on the MPEG-H Audio standard is being designed and tested to offer interactive and immersive sound, employing the standard's audio objects, height channels, and Higher-Order Ambisonics features. Object-based interactive audio offers users the ability to personalize their listening experience, setting their preferred language and dialogue level, or selecting elements to "hear their home team" or listen to their favorite race driver's radio. A four-stage process is introduced for implementing the complete system in TV networks. Additionally, the plant design, creative, and operational implications of producing content are discussed, based on the design and field testing of the system. Consumer reproduction implications are also presented, such as a "3D Soundbar" prototype, the control of loudness in the system, and rendering for playback on both traditional and new media devices.*

Keywords. MPEG-H Audio, Immersive Audio, Interactive Audio, ATSC 3.0, Dialogue Enhancement, Loudness, Audio Objects, Object Audio, 3D Audio, Barrier-free

Introduction – A Vision Of Future TV Audio Systems

The television industry and related standards bodies around the world are preparing for the delivery of UHD video through new standards such as ATSC 3.0, Super Hi-Vision, and the components of DVB-UHDTV Phase 2. All of these standards include or are considering new audio systems to provide additional features or performance beyond those offered today.

Several factors are driving the need for improved TV audio. One is the traditional desire to maintain parity with other media, such as cinema or Blu-ray, that are also in the process of deploying improved sound. Another factor is a change in the consumption model and expectations of consumers. The 20th century model of producing "one size fits all" audio primarily consumed by "couch potatoes" in the living room is no longer relevant in the 21st century world of new media that plays on many devices of varying capabilities in diverse environments according to the consumer's schedule and preferences. The proposed standards also reflect the possibility that future distribution of content or elements may happen over alternate distribution mechanisms such as mobile data or Internet as well as traditional broadcast distribution channels. Finally, there is the technical requirement that a new audio delivery system complement the realistic viewer experience possible with UHD video with similar audio fidelity.

In this paper, we will consider the needs of new TV audio systems and discuss the requirements and features of the system based on the MPEG-H Audio[1,2] standard being jointly developed by Fraunhofer, Qualcomm and Technicolor, along with its implications for TV plant and network design, consumer reproduction, and our experience using it in initial field tests. This system is being proposed to or considered by ATSC and DVB for use in future standards.

Requirements Of TV Are Different Than The Cinema

The cinema industry is currently transitioning to immersive sound, primarily employing the Dolby Atmos and Barco/Auro 3D systems. The question might be asked: Why not adopt or adapt these systems for television?

In the past, film and TV audio have shared many techniques and creative practices, but today's cinema sound systems have been designed to address different issues than are faced by the television viewer.

In the cinema, the viewer is in an acoustically large room, mainly experiencing sound in the reverberant field of the loudspeakers, while a TV viewer listens in a living room or home theater with much more direct sound. [3]The home listener does not have the expectation of correct reproduction of the sound image at the edge of his listening room, while the paying cinema listener may have to take a seat on the side aisle.

Further, the primary need for audio objects in the cinema is the desire to offer spatial accuracy (or some sense of it in dynamic cases) to a distributed audience served by a large number of loudspeakers arrayed on the cinema walls and ceiling. As will be explained below, the home listening environment can be adequately served by budget and décor-friendly speaker arrays with a smaller number of speakers. This reduces the need to use objects for mixing and transmission efficiency for content to TV viewers. However, as shown in the next section, objects have very powerful but different uses in TV content.

Another distinction is the assumptions on reproduction environment. In cinema sound, the audience is served by loudspeakers located in established positions following the standard house curve frequency response and playback level. In the home, we have a spectrum of

reproduction quality, from elaborate home theaters, to the mainstream consumer with a stereo soundbar, to the viewer watching on a tabletop “kitchen TV” with internal 5 cm speakers, and now the tablet or mobile phone viewer with 1 cm speakers.

The cinema is intended as an engaging but passive long-run experience, while the home viewer has full use of a remote control with buttons to change channels, fast-forward through uninteresting content, and adjust the volume while watching with a varying interest level.

Cinema viewers normally watch feature films carefully scored and mixed in a post production environment. While television viewers may watch similar films or scripted dramas, they also consume news, sports, documentary, and reality shows where the audio production is not as elaborate nor are there normally time and resources to post-produce the audio program, as well as remote or live events, where the sound may be mixed in the confines of a remote truck or OB van.

Thus, while our system will support the ingest and transmission of cinema content, including dynamic objects, a new TV audio system must consider the needs of a diverse set of production requirements and consumption patterns, and offer features attractive for delivery to the home and mobile devices.

Elements Considered For Future TV Audio

Interactive audio elements and consumer choice

Today all audio elements are mixed to one single signal for emission. Offering an additional version of the audio mix with e.g. dialogue in a different language requires sending an additional complete mix and effectively doubling the required audio bitrate.

An object-based approach enables broadcasters to offer a broader range of options for personalization and consumer choice. Not all audio elements need to be delivered as separate audio objects to reach that goal. Only those audio elements that are beneficial for personalization are delivered as separate audio objects. All other elements are mixed into a “channel bed” that contains the main ambience or music and effects.

One element that should be delivered as a separate object is dialogue. If several dialogue elements are offered, consumers can select e.g. one language out of a number of available languages or in a sports scenario choose either the home or the away team commentary. In a second step, the consumer may select additional effects objects or dialogue objects, such as narration for the visually impaired or pit crew radios in car races.

Figure 1. Interactive audio elements controlled by viewer

Objects may be mono or multichannel audio elements. Dialogue is usually a mono audio element whereas an additional effects object may contain a number of encoded signals.

In the latter case, the encoded signals that form the object are part of one “group”, so that only the complete group can be selected and manipulated and not single elements of that group.

Another important concept for object-based audio is the “switch group”. A number of objects are member of a switch group and only one of those objects can be selected at the same time. A switch group avoids that e.g. two dialogue objects with different languages are played back simultaneously.

In addition to selecting different objects, the relative sound level of objects can be adjusted for a personalized listening experience. An important use case is e.g. raising the dialogue or commentary level over the background sound. Objects can also be controlled individually in terms of their dynamic range, which ensures audibility in all dynamic range compression modes.

To enable easier choice for the consumer and better control for the broadcaster, the broadcaster may offer different presets. A “preset” is a selection of objects for a different mix. In a sports scenario, examples could be a “commentary preset” with a pronounced commentary and only moderate ambience, or a “live preset” without commentary but with additional ambience objects.

Figure 1 shows an example of an MPEG-H Audio Scene with five different groups and one switch group. The switch group contains three commentaries to choose from, two different English commentaries and one foreign language. Each mono dialogue element can be encoded in MPEG-H Audio with a bitrate typically between 20 and 40 kbit/s. Additionally the user may select the “sound effects” object. In this case the sound effects object is not a single mono source but a multi-channel object with pre-rendered content.

A “commentary preset” for this Audio Scene could contain the groups “1”, “2” and “5”. The groups “1” and “2” are “on”, group “5” is “off”. A “live preset” would contain the groups “1”, “3” and “5” and all those groups are “on”.

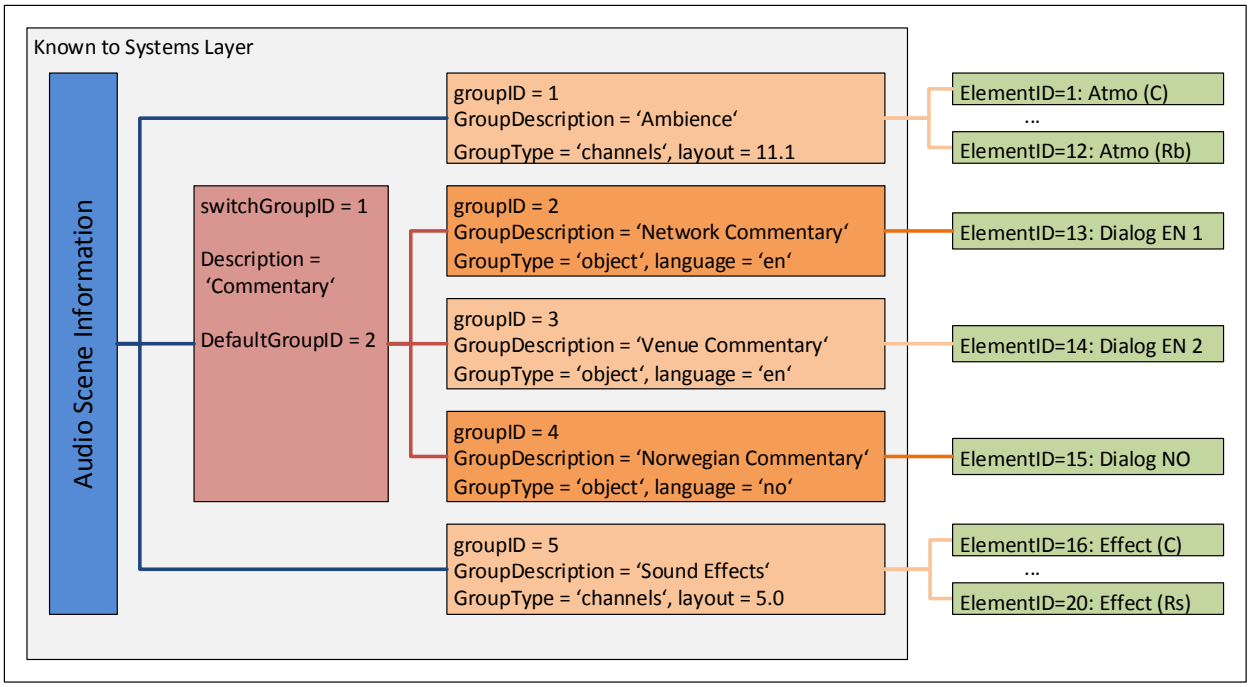

Figure 2. Example of an MPEG-H Audio scene for object-based audio

Immersive sound

Today's 5.1 surround sound broadcasts offer a good listening experience that is superior to stereo, but lack the height element of realism to convince a viewer he is actually in the scene being portrayed. While the 7.1 speaker configuration of Blu-ray offers improved spatial accuracy behind the viewer, an element that is missing in these systems is sound from above. Outdoors there are rain, birds, and aircraft providing direct sound sources from above, as well as elevated reflections from buildings, terrain, and clouds. Indoors there are similar reflections from ceilings and walls as well as direct sound sources. These sounds emanating above the horizontal listening plane offer important auditory cues that are missing from today's stereo, 5.1, and 7.1 speaker configurations. Thus, immersive (to avoid confusion with stereoscopic TV, we prefer this term to "3D") sound systems have been developed to reproduce a spatial volume of sound instead of planar sources.

As the recommended viewing distance decreases with the increasing resolution of UHD TV systems, it becomes interesting to follow action vertically as well as horizontally and to provide realistic auditory presentation for sources in motion on or off the screen. Humans are able to perceive angular differences in audio signals of 1-3º horizontally, and perhaps 4-17º vertically[4]. In a UHD-1 TV system, viewed at a conservative 2 times picture height, the angle occupied by the display is 47º horizontally and 28º vertically, thus it becomes useful to pan sounds vertically as well as horizontally.

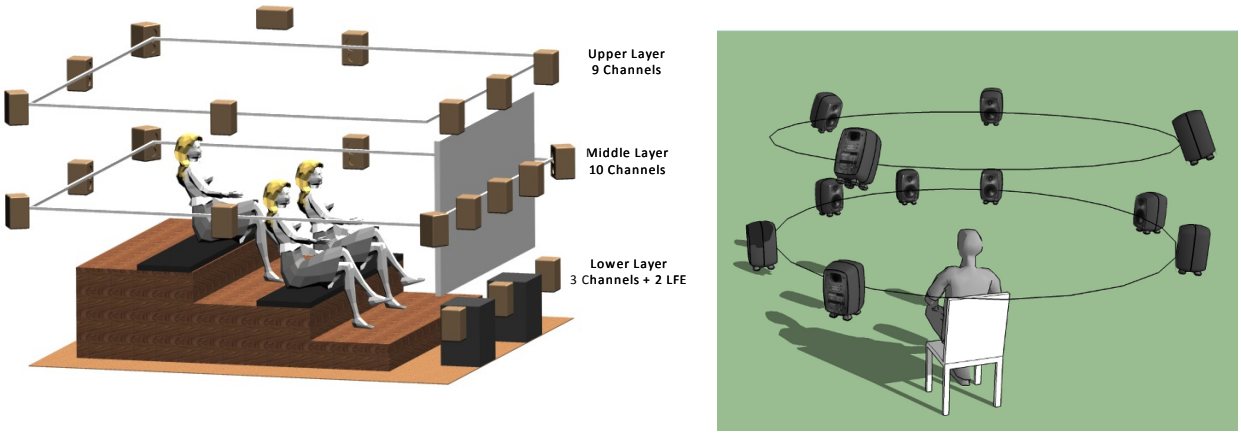

Figure 3. 22.2 and 7.1 + 4H Speaker Configurations

A system proposed by forward-thinking broadcaster NHK uses 22.2 channels which provides excellent realism and is a de facto benchmark for spatial reproduction[5]. Fraunhofer has assisted NHK in the initial development of AAC transmission coding for this system and a 22.2 transmission may be carried with transparent quality at a bitrate of 1200 kb/s. However, a bitrate in this range may begin to encroach on the video bitrate intended for a TV system, and most current TV plants are built on SDI architecture with 16 channels of audio available. 22.2 audio may certainly be accommodated by using advanced SDI signals, as will be required for 4K video, but then there are also the operational issues of creating 22.2 channel programs and reproducing 22.2 channels in the viewer's home.

Bitrates in kb/s for:	Good	Recommended	Transparent
22.2 Channels	256	512	1200
7.1 + 4 Height Channels + 4 Objects	200	384	800
5.1 Channels	96	160	256
2.0 Channels	32	56	160

Table 1. Typical bitrates for common channel configurations.[6]

Thus, Silzle et al. undertook a study[7] to determine the relative overall perceived sound quality of several speaker configurations, to determine if a practical compromise was possible between the sound quality provided by a 22.2 system and today's 5.1 and 2.0 formats. As shown in Figure 4, the perceived quality improvement from 5.1 surround to 22.2 is greater than that from stereo to 5.1. However, ignoring the LFE channels, an upgrade from stereo to 5.1 requires 3 new speakers, while an upgrade from 5.1 to 22.2 requires 17 new speakers. Our tests using an active downmix method show that most of the perceived improvement of upgrading 5.1 to a 22.2 system can be obtained with four additional height speakers .Our TV audio system offers support for up to 128 channels or objects, and has channel configurations defined up to 22.2, but we recommend for initial deployment a 5.1 + 4H (4 Height speakers) or 7.1 + 4H configuration.

Of course, any channel configuration must include the possibility of reproduction on a system with greater or fewer channels. Traditionally, the TV audio decoder includes the possibility of down-mixing a 5.1 program to stereo or mono using a fixed equation, perhaps with gains for the center channel and surround channels sent in the bitstream.

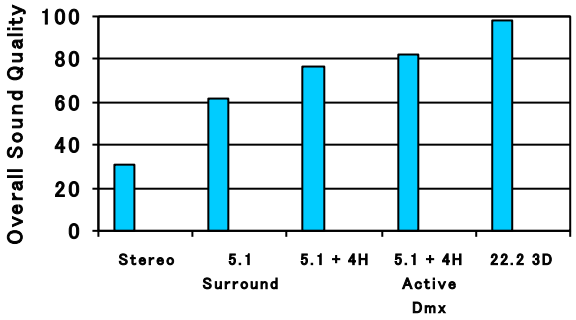

Figure 4. Overall sound quality improvement with expansion of reproduction system to surround and immersive / 3D formats compared to 22.2 channel reference signal

In our system, the concept of flexible rendering has been introduced. This allows any channel configuration to be reproduced (with varying degrees of success depending on the arrangement) on any arbitrary arrangement of speakers. This allows for the possibility of correctly reproducing the sound image in the viewer's home even if architecture or furniture make it impossible to place the speakers in the ideal locations. Active downmixing for rendering is carried out using a frequency-domain, energy-preserving downmix that avoids the phase cancellation effects that sometimes occur with traditional downmixes.

In addition to channel-based audio, our system also includes the option of using Higher-Order Ambisonics[8] components to represent the sound image instead of channels. Ambisonics components, sometimes referred to as B-format signals, are not related to speaker positions but instead describe a sound source's direction by means of their relative amplitudes and polarities.

One way of thinking of these components is to consider them as coincident microphone signals of increasing complexity. Figure 5 shows the shape or "polar pattern" of these components for orders zero to three. The zero-order component can be thought of as an omnidirectional microphone, and the first-order components as figure-of-eight microphones (such as a side-address vocal condenser microphone) oriented along the x, y, and z axes in space. The second and higher order components have patterns that are not common in traditional microphones, but combine with the lower order components in a manner similar to a Taylor or Fourier series to more accurately describe the direction of a sound or sound field. Indeed, these components are termed spherical harmonics.

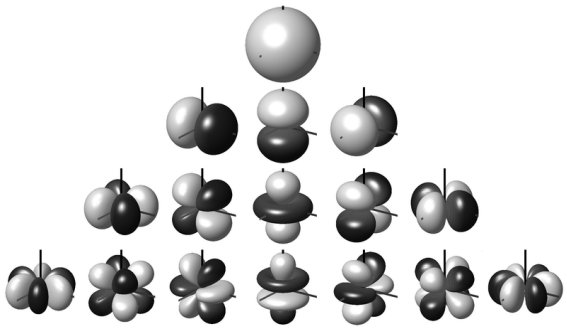

Figure 5. Visual representation of the Ambisonic B-format components up to third order.[9]

The practical implications of using Ambisonic components is that they are independent of any speaker layout and may be used to capture a sound field using a single complex "ball" microphone. It is also possible to create Ambisonic signals artificially using a panner which

samples the x,y,z values of each component. This may be used to add effects, music, or spot microphones into an Ambisonic signal.

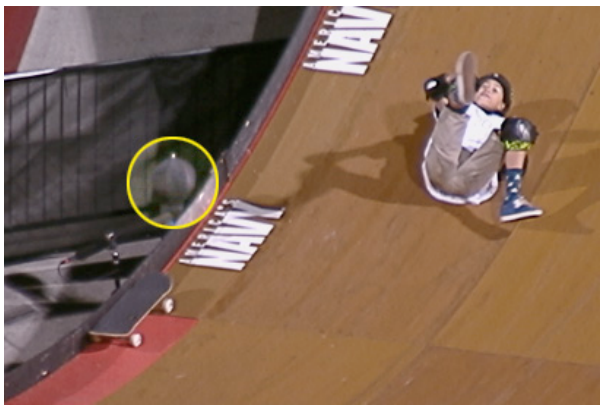

Figure 6. Ambisonics microphone next to host broadcaster shotgun microphone on skateboard ramp

Depending on the order N, of the desired Ambisonic program, $(N+1)^2$ total components are needed. Thus, the classical first-order Ambisonic recordings used four components, W, X, Y, and Z to represent the sound field. Today's Higher-Order Ambisonics typically employs third or fourth-order signals with 16 or 25 components. Just as for signals in a time-domain series expansion, it is possible to delete higher-order components of an HOA signal to reduce complexity, the tradeoff being reduced spatial accuracy. HOA signals may also be spatially compressed during encoding to reduce the number of components input to the audio codec and transmitted.

The renderer in our system is designed to render channel-based, HOA-based, and object-based audio simultaneously. In a typical use, a channel-based ambience "bed", perhaps mixed as 7.1 + 4H channels would be combined with independent audio objects for dialogue or other elements in the renderer. Alternatively, a third or fourth-order HOA sound field representation would be similarly combined with objects. All of these elements would then be rendered for the actual speaker configuration (and precise speaker location if available) the viewer is using.

Improved loudness management

Loudness control in multi-platform distribution

The traditional TV broadcast environment uses a well-defined end-to-end solution to deliver media content to the end user. Accordingly, it has been a good compromise to define a particular target loudness and dynamic range for this particular delivery channel and well-known type of sound reproduction system.

However, new types of delivery platforms have become significant and there are signs that these two types of delivery platforms have begun to converge. In a multi-platform environment, the same content is delivered through different distribution networks (broadcast, Internet, mobile networks, etc.) and is consumed on a variety of devices (TV set, tablet, mobile phone, etc.) in different environments.

From the consumer's point of view, the characteristics of the audio should fit the individual listening conditions irrespective of the origin and distribution channel of the content. For

example, when watching movies, sports or TV shows on his TV set in the home, the consumer may use different distribution sources such as terrestrial broadcast or Internet streaming to receive the content. Variations in loudness when switching between these sources may cause annoyance and lower consumer satisfaction.

Also, the playback of TV content on new media devices such as smartphones or tablet computers should ideally match the loudness of existing media such as music stored on the device, which typically has a perceived loudness of -10 to -14 dB LKFS. Thus TV content on these devices must be adjusted for similar loudness by dynamic range processing in the decoder.

Tailoring loudness and dynamic range for different listening environments

For classic TV broadcasting, typical listening scenarios include using built-in stereo speakers of TV sets and 5.1 home theater reproduction. Similar program loudness and dynamic characteristics are desirable for both types of reproduction[10].

Therefore, typical U.S. TV program sound is well-normalized at -24 LKFS and the dynamic range is typically adequate for living room reproduction, as the home listening environment is reasonably quiet. The listener will adjust the volume according to their preferences – perhaps by finding a compromise between intelligibility and not disturbing others.

Since TV content is increasingly consumed with mobile devices in various places, the need for adjusting the audio characteristics to the playback scenario has gained importance. Depending on the connected earphones, mobile devices can produce sound pressure levels beyond 100 dB(A). However, the maximum output levels of such devices may be limited to prevent users from hearing loss motivated by product liability or country-specific regulation. Tablet speakers, however, are physically limited to much lower output levels, though signal processing can improve this in many cases.

Background noise is one of the determining factors of appropriate loudness and dynamic range characteristics. For example, many people enjoy listening to music or watching movies using mobile devices in airports or on public transport, where the sound pressure level of background noise can reach more than 80 dB(A). In these cases, high dynamic range content cannot be reproduced appropriately on typical earphones which have no acoustic isolation. In contrast, background noise of typically 40 dB(A) in a living room would allow for a reproduction with more dynamic range.

As a consequence of the wide variability of listening environments, further adaptation of the audio is required to avoid user annoyance in many cases. The features of our system allow the loudness and dynamic range of a single broadcast bitstream to be tailored to the viewer's listening environment and playback device.

Loudness and Dynamic Range Control functionality of MPEG-H Audio

To ensure consistent loudness across channels and within a program, MPEG-H provides an improved loudness and dynamic range control (DRC) scheme. During standardization, requirements for interactive and immersive audio have been taken into account as well as quality improvements compared to existing DRC solutions of legacy audio codecs.[1] MPEG-H includes capabilities to convey long-term loudness and peak information in the bitstream as given or measured at the encoder side. These parameters are used in the decoder for normalization to a given target loudness. Further, associated decoder processing blocks in interaction with Dynamic Range Control (DRC) data in the bitstream allow for adaptation to different listening conditions.[1]

The DRC processing of MPEG-H is controlled by the encoder, thus assuring broadcasters that the final sound reproduction quality in all receivers is acceptable. To improve the resulting sound quality compared to the DRC schemes of legacy audio codecs, MPEG-H DRC offers an improved time resolution of transmitted and applied DRC data. In addition DRC data for multi-band compression can optionally be conveyed. By providing appropriate DRC configurations, the audio decoding can be adapted to different playback scenarios. With this, MPEG-H has improved flexibility, including for reproduction with the higher target levels required for mobile devices.[1]

Deploying immersive and interactive audio in TV broadcasting calls for new concepts and methods to measure loudness. The loudness measurement algorithm in ITU-R 1770-3 has been defined for up to 5.1 channel signals. As immersive audio adds the height dimension to multi-channel audio, existing concepts need to be extended for formats with higher channel counts.

It has been described above that default mixes or presets will be used for a defined reproduction of the audio. The reproduction of object-based audio according to a preset is controllable by the broadcaster. Therefore, MPEG-H provides the mechanism that loudness information and DRC data can refer to such presets for the representation of objects as described in reference 11.

In addition, interactive audio can be a powerful tool for an improved approach of solving loudness issues. Today, loudness control basically means to align the overall levels. By adding objects for personalization, it will become possible to tailor the audio to different playback conditions and for individual capabilities. For example, dialogue intelligibility is an issue often associated with loudness aspects.

Apart from DRC data for a complete preset/representation, MPEG-H can provide DRC data specific for an object. Here, improved functionality for personalization is possible, such as ducking for voice-over or improved intelligibility by separate processing of a dialogue object.[1]

Altogether, the loudness and DRC functionality of MPEG-H offers a reliable way to assist broadcasters in their efforts to control loudness. Moreover, in conjunction with enhanced DRC tools of MPEG-H Audio, the resulting audio can be tailored to different listening conditions while the DRC functionality can further improve user experience of personalized audio.

Extending consumption with barrier-free accessibility

Audio description or narration for the visually impaired is an important accessibility feature that is already offered for many programs today. One common practice is to deliver a second complete mix as an alternative (“broadcast-mix”) that includes audio description. In an object-based scenario, the narration for the visually impaired is added as an additional optional dialogue object to the presentation thus minimizing the additional bitrate required.

Another accessibility feature for a hearing impaired audience that is easily enabled with objects is dialogue enhancement. Delivering the dialogue as a separate object allows the user to adjust the mix between the dialogue and the background signal according to his preference. Users may need different settings based on their individual hearing impairment. Tests have shown that an enhancement of the dialogue by about 6dB offers a substantial improvement in intelligibility for an audience with a typical age-related hearing loss[12].

Challenges of Implementing Future TV Audio for Broadcasters

Transitioning to future TV audio systems such as ours will impact creative, operational, and engineering functions of TV production, just as with the change from stereo to 5.1 sound

broadcasts that happened in the HDTV transition era. Some impacts will be similar to the HDTV 5.1 sound transition and some very different.

The MPEG-H decoder supports all the features discussed in this paper as well as the ability to render the audio in the best quality possible on the viewer's device or the audio system connected to it. Broadcasters are thus free to choose when and how to implement new features as they desire based on consumer acceptance, operational maturity, and competitive factors. We propose four stages of deployment in existing TV plants and networks:

The four-stage deployment model

First, a broadcaster may initially use the system just as he does today for stereo or 5.1 content, including existing loudness control techniques, and have no operational or creative impact, only the bitrate savings from improved coding efficiency.

Second, interactive objects may be added alongside 2.0 or 5.1 programs to support additional languages, provide mandated access features, or offer viewers the possibility of adjusting their own mix of dialogue and/or other elements. In this stage, the envisioned "home/away team commentary" or "hear pit crew radio for your favorite driver" scenarios we have presented in field tests become possible. Based on our experience, the existing plant and outside broadcast infrastructure is able to accommodate these static objects with few changes to the production equipment or workflows.

In the third stage, immersive channel-based or HOA-based sound would be added, increasing the realism of the audio image to create the "you're in the scene or at the event" experience. Depending on the channels needed for objects and the level of immersive sound, additional audio signal distribution beyond the existing 16-channel embedded SDI audio may be required.

Finally, in the fourth stage, the ability to add dynamic objects to the program would be introduced. This stage would require new panning tools in production and the ability to carry dynamic panning data through the TV plant and network to the emission encoder.

In all these stages, we envision the audio as being mixed, transported, and stored in the production environment as standard uncompressed PCM audio. The transport of dynamic panning data will be the topic of a future paper, but is expected to use existing AES or SDI (or their IP network counterparts) channels as well. We envision limited-bandwidth contribution, distribution, emission, and re-distribution links using compressed MPEG-H Audio bitstreams, multiplexed with the accompanying video signal, as shown in Figure 7.

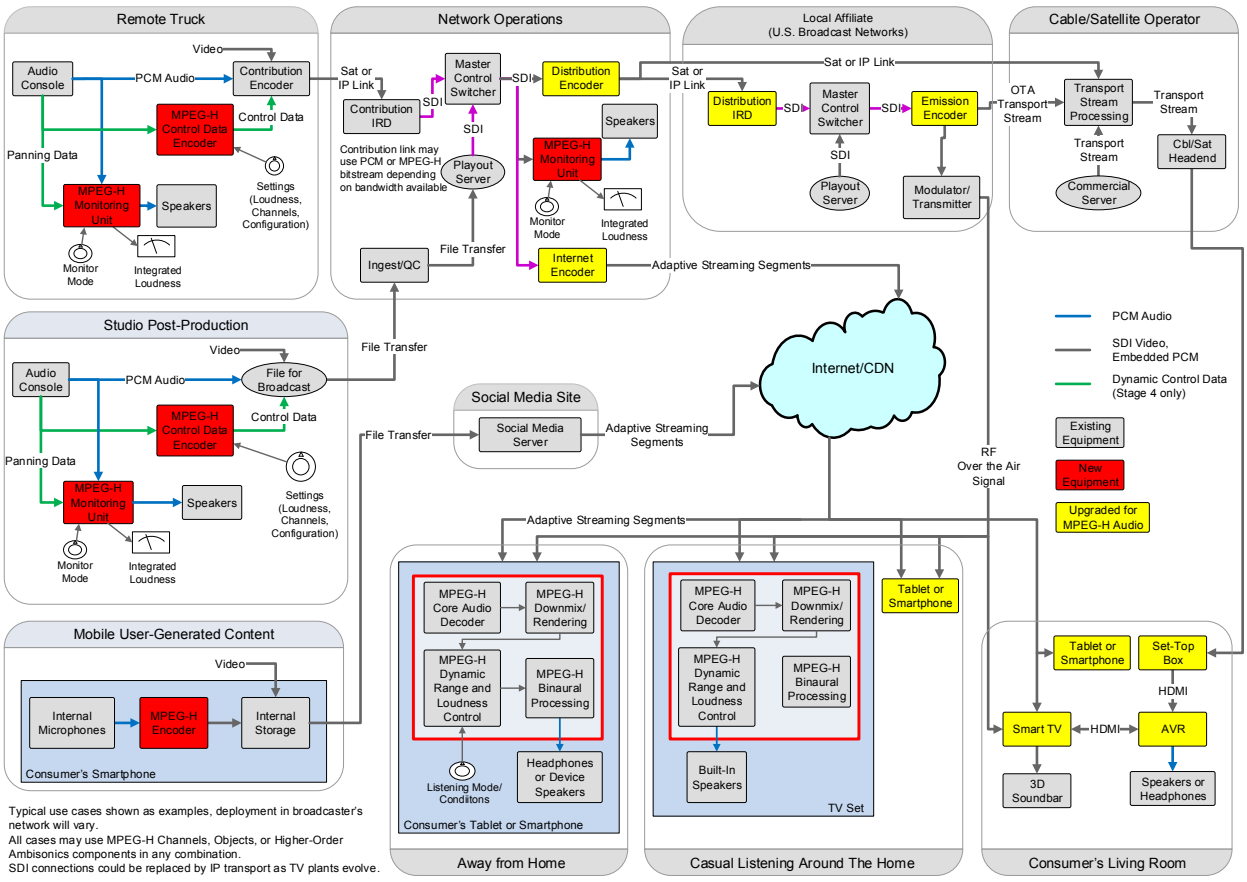

Figure 7. Simplified signal flow for our TV audio system

Audio mixing

As explained below, our field tests at live broadcasts have demonstrated it is very easy to add interactive objects to existing audio broadcasts as these elements already exist as signals on the channel strips of the audio mixer's console. Conceptually, sub-mix or bus faders are just being transported as objects from the audio console to the viewer's playback device. Of course, the mixer is free to put some virtual tape across the fader slot, so to speak, to limit the control range the viewer can exercise, and also set the default gain of the object.

Adding immersive audio not only involves adding new signals, but also extending the panning and mixing functions of the live console or post-production digital audio workstation to handle perhaps 5.1 + 4H or 7.1 + 4H or third or fourth-order HOA signals. There are several operational strategies for producing these signals:

One simple approach is to create an additional 5.1 or 4.0 bus in the console or workstation software to carry the height channels, with the console's existing 2-D panner being used to pan signals horizontally and a bus fader or send gain being used to control the vertical panning indirectly by varying the portion of the signal sent to the "height bus". This does not offer precise vertical panning in terms of x,y,z or r, θ, ϕ coordinates, but may be adjusted by ear, as the mixer is likely to do anyway.

Another is to add a VST or RTAS plug-in to the console or DAW software that implements true 3-D panning (not to be confused with the more prevalent HRTF panning tools for binaural use) using the console's existing buses. Such plug-ins exist today for research or gaming use and

will undoubtedly become more available as immersive audio systems for TV are deployed. Mixes of our initial winter field test broadcasts were done using these two methods.[13]

A third is to connect an outboard hardware unit to the console or DAW that does panning and downmix functions, accepting microphone signals or other sources, and positioning them in the immersive sound volume. Such units are commonly used in the film industry today for mixing immersive sound, and we have successfully used film consoles for creating immersive mixes of recorded broadcasts.

Ultimately, we expect 3D panning and buses to be built-in to new consoles and workstations. Indeed, one console popular for outside broadcasts offers a 22.2 panner option today.

In addition to the channel or HOA bed, immersive productions will likely include objects. If these are statically positioned, such as dialogue to the center channel, no further consideration is needed. If they are dynamic, such as a helicopter sound panned to fly over the listener, or sound effects panned to follow an on-screen source, then the panner positioning data must be carried along with the object's audio signal through the TV plant to the point of encoding into the MPEG-H bitstream. Our plans for doing so will be the subject of a future paper.

Audio monitoring

To produce an immersive audio program, the mixer must necessarily be able to listen to the program rendered in an immersive format, as well as have the ability to spot check the mix in stereo or 5.1 to insure that viewers with only those capabilities will hear a good presentation.

Our experience has been that this can be accomplished by adding four height speakers to an existing 5.1 or 7.1 monitoring system. Our initial field trials were done in an existing rented 5.1 recording truck with four similar but smaller speakers mounted on light stands for the height channels and connected to additional console outputs.

OB vans or remote trucks often have limited space in the "audio closet" and engineers object that there is no room for additional speakers. Here we see three options:

Ceiling Mounted Speakers. One solution is to use smaller speakers and mount them on the ceiling. While there are compromises in output level and bass response, professional monitors intended for desktop use are certainly adequate for monitoring the height signals of immersive ambient sound, though perhaps not the full experience of a helicopter or spaceship flyover. Particularly if proper rendering is used, immersive audio in compact environments down to the size of sports car cockpits has been successfully produced by Fraunhofer. The flexible rendering feature of our TV audio system can also help accommodate the need to place speakers away from their ideal position due to obstacles.

Remote Mixing. Another approach is to adopt the technique of remote mixing now being employed for complex or diverse sports remotes, such as the recent FIFA World Cup. Here stereo audio is mixed in a small audio room; perhaps in an ocean cargo container instead of a section of a custom expanding-wall truck. Audio sources and sub-mixes are also sent over a high-bandwidth IP connection to a remote center with an audio control room provisioned for 5.1 mixing. Such a room could be easily upgraded to 5.1 + 4H or 7.1 + 4H monitoring by adding speakers.

Immersive Headphones. A third option is to use headphones for immersive monitoring. With head tracking, this is an experience very different from conventional stereo headphone listening, employing binaural rendering pre-processing to create a sound image that is very close to what would be heard in loudspeaker reproduction. This

approach is currently being used in film mixing for initial post-production. Open-back headphones would allow audio mixers to continue to hear intercoms and alarms just as they do with speakers today.

Turning to drama and other studio-produced shows: as discussed above, we have successfully used film consoles as well as popular DAW software for post-production. Our experience has been that large film mixing stages do not work well for mixing immersive sound for home delivery, for the reasons indicated earlier. What is needed is to equip the smaller mix rooms used today for Blu-ray, DVD and broadcast delivery mixing with additional height speakers, most likely in 7.1 + 4H configuration.

At a practical level in limited audio post production suites or TV editing rooms, it is a simple matter to adapt them for immersive audio by adding four additional amplified speakers and using an additional "sound card" hardware interface to drive them. Most soundcards intended for DAW use today have ADAT or similar outputs that can be used to drive an additional slave soundcard through the primary's internal DSP mixer.

In our recommended implementations, audio is carried primarily as PCM channels. This aids monitoring and quality checks at points in the signal chain outside creative rooms using standard monitor panels. We also envision the production of "monitoring units" which will accept, render, and optionally downmix PCM audio and control data to allow monitoring of the consumer experience (including action and ranges of object controls) and loudness.

Creative conventions and techniques

With the transition from stereo to 5.1, TV sound designers and mixers could rely to some degree on the experience and techniques of their film counterparts. In transitioning to future TV audio systems, the authors expect less commonality with film production for several reasons:

TV broadcasts are likely to employ object-based interactivity before adopting full immersive sound. Interactivity is unlikely in the cinema, given the large audience sharing the sound where immersive dynamic objects will become common in film mixing. Speaking of dynamic objects, will mixing conventions for TV change to track on-screen sound sources with dynamic objects? How will this be treated when camera angles change? These are all problems diverging from film practice.

Operational concerns

Although monitoring and mixing immersive and interactive audio programming has been shown here to require limited modifications beyond today's 5.1 programming, if a broadcaster chooses to implement the advanced features possible, the programming will become more operationally complex. Today, TV programming tends to fall into a few well-known formats. For example, HD programs may be 720p or 1080i with stereo or 5.1 audio.

In the future, it may be common to have a premier sports event produced in 7.1 + 4H with separate objects for home and away team commentary, second language commentary, and player or official microphones. This may be interrupted by a half-time or update show produced in stereo with no objects, with a mix of stereo, 5.1, and (perhaps someday) immersive commercials inserted that may have second or third language objects. These disparate programs must be recorded, transmitted, and decoded properly according to the viewer's preferences. Thus, future broadcasts have to be adaptive – the current practice of maintaining 5.1 emission and upmixing stereo sources, or the distasteful alternative of downmixing all programming to stereo, will not work in a future era of disparate objects. Either static configuration data must accompany each program or control information must accompany the

audio throughout the TV network to indicate the current configuration of the audio program. Also, the small but important task of labeling objects for display in the user interface of the viewer's device must be considered.

We envision several methods of handling this complexity as the frequency of interactive and immersive broadcasts increases. One method is to use encoder presets triggered by automation system events to change the audio configuration data for each program or segment. Another is to distribute configuration data as XML files or other formats with the audio. Our plans for accomplishing this will be explained in a future paper.

Of course, operational strategies will also depend on the overall transition strategy developed to move to new TV standards. Legacy over-the-air broadcasters will undoubtedly face different challenges than new greenfield over-the-top services.

Content storage

Creating interactive and immersive content in the studio may be done with today's DAW tools and session files. Storing content for transmission will involve either MPEG-H bitstreams or PCM audio with control data. MPEG-H bitstreams, of course, may be stored in the MPEG-4 file format.

PCM audio with control data may be stored in existing uncompressed audio formats extended to support the associated control data. For example, to support object-based audio in formats such as the Broadcast Wave File (BWF) or MXF, the EBU has developed the Audio Definition Model (ADM)[14]. The control data of MPEG-H Audio is defined so that it supports the main features of ADM. A scene description in ADM can therefore be transferred to the MPEG-H representation.[11]

Challenges of Implementing Future TV Audio for CE Industry and Consumers

Immersive Audio playback

Undoubtedly, videophiles and other enthusiasts will install height speakers to complement their existing 5.1 or 7.1 systems, providing the highest quality reproduction. With the advent of the HDMI 2.0 standard, 32 channels of PCM audio may now be carried, and audio/video receivers with 12-channel amplifiers have been introduced for Blu-ray playback of immersive films.

However, we envision the possibility of mainstream consumers experiencing immersive sound. Fraunhofer has constructed a concept prototype of a "3D Soundbar" which has been demonstrated at industry events, as shown in Figure 8. It provides immersive sound, including the perception of sounds from above and behind the listener, using only speakers surrounding or inside the TV. As previous experiments have suggested[15], consumers could install a wire-free, no-setup immersive sound experience just by taking a productized version out of the box and hanging it on the wall.

For consumers who continue using legacy 5.1 or 7.1 speakers, our system provides two benefits through its flexible rendering feature. One is the ability to use psychoacoustic processing to provide a small amount of height perception from existing mid-level speakers. The other is to compensate for speakers misplaced from their ideal locations due to architectural or décor considerations.

Figure 8. Concept Prototype of 3D Soundbar

Future interconnect standards will allow communication of the actual speaker positions to the decoder. The location of speakers is outside the scope of our system, but can be accomplished through manual measurement, acoustic techniques, or other means. For example, for the past four years, Fraunhofer has used a custom test set to measure the angle, distance, and frequency response of speakers in its listening rooms, as the workload of verifying multiple 22.2 speaker configurations in a room requires an automated solution.[16]

Figure 9. Microphone Array for Automatic Speaker Measurement (including Angle and Distance)

Turning to new media devices, our system will allow for binaural playback through headphones through binaural processing in the renderer. The system is also designed to work with the separate product Fraunhofer Cingo, which provides virtual surround or immersive sound over tablet speakers.

Interactive Audio Playback

It is envisioned that viewers will use an on-screen display (such as in Figure 1) to set the interactive parameters used by the renderer. Several common configurations of home equipment are affected by the need for communications from the remote control held by the viewer to the MPEG-H Audio decoder. For example, consider the case where the program is received over the air by a TV set, which then sends an audio bitstream to an AVR (Audio-Video Receiver). The AVR contains the decoder, while the TV acts as the video source and display. Thus, the AVR cannot provide the OSD, and the viewer may be using the TV's remote control during programs. Work is underway to provide communications for remote control data in home interface standards for this purpose. It may also be desirable to consider equipping new TVs with HDMI 2.0 PCM outputs so that full decoding may take place in the TV instead of an AVR.

Another feature potentially requiring interaction between a consumer device and the MPEG-H Audio decoder is the ability to retain user settings during interstitial items such as commercials. Consider a user who has adjusted his mix of alternate audio objects presented during a broadcast. These settings make no sense for a commercial, which may present different audio objects. Once the commercial break is over, the user's settings must be restored when the program resumes. An additional needed feature is a "reset button" to return settings to the default, not only for user convenience but also for consumer support.

Experiments with the New TV Audio System

An audio codec may be tested with laboratory experiments, but building a complete audio system requires the consideration of operational issues and the discovery of missing features or performance limitations. It is also important that a system design consider creative review of the system's performance and creative opportunities. Thus, Fraunhofer, along with its partners Qualcomm and Technicolor, have conducted field tests during the development of this system.[13]

These tests have involved producing both sports and film-style content. Of most interest from a TV standpoint are the tests done at live sports broadcasts by recording the host broadcaster's audio console signals, as well as supplemental microphones, and then mixing and encoding the signals later in the studio. This was needed as live encoding has only recently become possible as the codec design was finalized. The encoded files were then played on a MPEG-H Audio decoder connected to speakers and a TV, while a TV-style remote control was used to control the object levels, as shown in Figure 10. The files were also loaded onto tablet computers with MPEG-H decoding apps to prototype new media delivery. These tests, exhibited to broadcasters at industry events, included:

- Winter extreme sports competition (skiing, snowboarding, snowmobile racing) carried on major cable network
- Summer extreme sports competition (skateboarding, motorcycle racing) carried on major cable network
- NASCAR race (with pit crew radios) using material from NASCAR
- DTM (European race series) auto race carried on major European sports channels

Production of the interactive objects, such as commentary or sound effects, was easy, as they existed as signals on the A1 mixer's console or could be routed over MADI from a technical operations room. In some cases, sound effects from spot microphones were placed as a separate object to allow testing if viewer adjustment would be useful. The control range of a few objects were limited to experiment with such limits. The amount of additional gain for objects also needs to be considered: too low and the utility of the object may be lost (for dialogue enhancement, for example), while too high and excessive limiting may occur or additional headroom must be provided in the decoder.

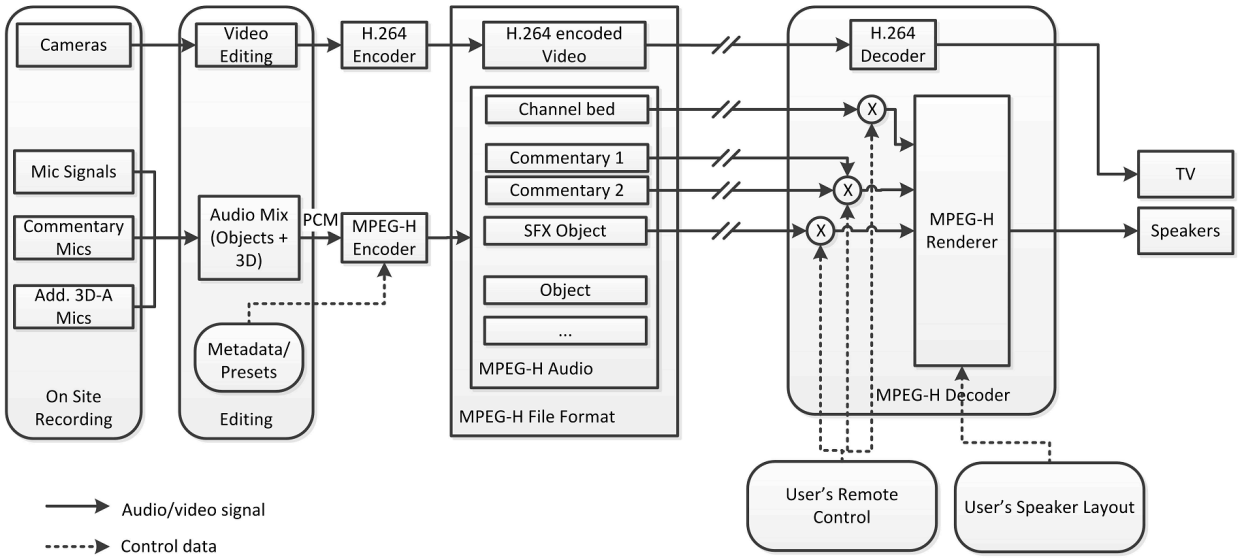

Figure 10. Signal flow of field tests[13]

Adding immersive sound elements was more challenging. Immersive sound is new to sports and will need experience to use well. Since these tests were conducted through the courtesy of host broadcasters, only limited accommodations for pickup of immersive sound could be made in the day or two before broadcast. Two techniques were used for capturing immersive sound: channel-based microphone trees or shotgun arrays and HOA "ball" sound field microphones.

There is no simple, small, *and* inexpensive 3D microphone today. One approach is to construct microphone trees from eight ordinary cardioid microphones on stands and arms, pointing towards each corner of a 60-90 cm cube, to capture the exact sound image of an event. Using the HOA technique, an array of 16 to 32 microphones mounted in a typically 8 cm metal sphere may be similarly used to capture a sound image. Of course, broadcasters may find existing microphone arrangements for surround may be also adapted for 3D sound by supplementing them with additional microphones for the height layer.

A sound image capture approach may work well for some events, where a 3D microphone can be placed appropriately to capture crowd ambience and other natural sound. For example, 3D microphones worked well to capture the audience sound in half-pipe snowboarding and skateboard ramp events. In other sports, such as downhill skiing, they were impractical for capturing action sounds due to the length of the course. (It is easier to cover a long course with mono shotgun spot mics, than with 3D mics at eight times the quantity or for ball mics, fifteen times the cost)

One common issue in our tests was bleed of the public address sound into the ambient sound, which can be challenging to control with 3D microphones. Another issue is change in perspective with camera cuts – a stationary 3D sound image may not work well with changes in perspective. Thus, the theoretical concept of capturing the sound with a 3D microphone array works well in some cases, such as for audience and ambience capture, and not in others, such as for action sounds.

TV sound is as much about creating an engaging experience through sound design as it is realistic capture of the live sound, and we have employed spot microphones, such as practically arranged pairs or groups of shotgun microphones, to augment or replace 3D microphones. Existing spot microphones of the host broadcaster were also panned (in one case dynamically) to help create a detailed and immersive sound image.

Fraunhofer and its partners are currently conducting additional tests and experiments to determine some potential guidelines of 3D microphone technique and sound design for various sports.

The MPEG-H Audio Alliance's New Immersive and Interactive TV Audio System

Our TV audio system, based on the MPEG-H Audio standard, is being jointly developed by Fraunhofer, Qualcomm and Technicolor, who have formed the MPEG-H Audio Alliance. In our listening rooms at industry events, we have shown real-time encoding hardware and the 3D sound bar prototype from Fraunhofer, decoding on a set-top box from Technicolor, and decoding on tablet computers using Qualcomm technology. More information on our system and future developments may be obtained at www.iis.fraunhofer.de/tvaudio and at www.mpeg-haa.com.

Conclusion

Broadcasters now have the opportunity to retain audiences with the engaging new features of immersive and interactive TV sound, while minimizing operational issues and plant impact through the four-stage deployment process we have proposed. We look forward to further tests of our system in cooperation with broadcasters and in consideration of our system for future ATSC and DVB standards, and will present additional details of our system in future papers.

References

1. J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio - The New Standard for Universal Spatial / 3D Audio Coding", AES 137th Convention, Los Angeles, California, October 9-12, 2014.
2. S. Meltzer, M. Neuendorf, D. Sen, "MPEG-H 3D Audio – The Next Generation Audio System", International Broadcasting Convention, 2014.
3. Taken the average directivity of a loudspeaker and the head into account, plus the average reverberation time of a listening room, the home listener sits for lower frequencies (<~3kHz) outside and for higher frequencies inside the critical distance, the border between the direct field and reverberant field. Personal communication A. Silzle, 2014.
4. J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*, MIT Press: Cambridge, MA, 1997 table 2.1 and figures 2.1, 2.2., 2.5.
5. K. Hamasaki et al, "A 22.2 Multichannel Sound System for Ultrahigh-Definition TV (UHDTV)", SMPTE Motion Imaging Journal, April 2008.
6. Neuendorf et al, "Immersive Audio With MPEG 3D Audio – Status And Outlook", NAB Convention, Las Vegas, Nevada, April 2014.
7. Silzle et al, "Investigation on the Quality of 3D Sound Reproduction", International Conference on Spatial Audio, 2011, Detmold, Germany.
8. M. A. Poletti, "Three-dimensional surround sound systems based on Spherical Harmonics," Journal of the Audio Engineering Society, Volume 53, Issue 11 pp. 1004-1025; November 2005.

9. F. Zotter, (Dr Franz Zotter <zotter@iem.at>) [CC-BY-SA-3.0 (http://creativecommons.org/licenses/by-sa/3.0) or GFDL (http://www.gnu.org/copyleft/fdl.html)], via Wikimedia Commons
10. Travaglini, A.; Alemanno, A.; Lantini, F.; "Defining the Listening Comfort Zone in Broadcasting through the analysis of the Maximum Loudness Levels"; AES 132nd Convention, Budapest, Hungary, 2012 April 26–29
11. S. Füg, et al, “Design, Coding and Processing of Metadata for Object-Based Interactive Audio”, AES 137th Convention, Los Angeles, California, October 9-12, 2014.
12. H. Fuchs, D. Oetting, “Advanced Clean Audio Solution: Dialogue Enhancement”, SMPTE Motion Imaging Journal; July-August 2014.
13. H. Stenzel, U. Scuda, “Producing Interactive Immersive Sound for MPEG-H: A Field Test for Sports Broadcasting”, AES 137th Convention, Los Angeles, California, October 9-12, 2014.
14. Tech 3364, “Audio Definition Model – Metadata Definition”, European Broadcasting Union, 2014
15. K. Matsui, A. Ando, “Binaural Reproduction of 22.2 Multichannel Sound with Loudspeaker Array Frame,” Convention Paper 8954, AES 135th Convention, New York, NY, October 17-20, 2013.
16. A. Silzle, et al, “Acoustic Measurement System for 3D Loudspeaker Set-ups”, AES 40th International Conference, Tokyo, Japan. October 8-10, 2010.