Building The World’s Most Complex TV Network (A Test Bed for Broadcasting Immersive and Interactive Audio[1])

Robert Bleidt

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

Herbert Thoma

Fraunhofer IIS, Erlangen, Germany, herbert.thoma@iis.fraunhofer.de

Wolfgang Fiesel

Fraunhofer IIS, Erlangen, Germany, wolfgang.fiesel@iis.fraunhofer.de

Stefan Kraegeloh

Fraunhofer IIS, Erlangen, Germany, stefan.kraegeloh@iis.fraunhofer.de

Harald Fuchs

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

Rinat Zeh

Fraunhofer IIS, Erlangen, Germany, rinat.zeh@iis.fraunhofer.de

Jim De Filippis

TMS Consulting Inc., Pacific Palisades, CA, jimdtv12@gmail.com.

S. Merrill Weiss

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

Written for SMPTE Journal©

Abstract. *Fraunhofer and its partners have developed a TV audio system based on the new MPEG-H Audio standard, now part of the ATSC 3.0 A/342 standard adopted for Korean broadcasts in 2017. Given its complexity, a complete broadcast plant was built to test the features envisioned. At NAB 2015 we demonstrated "The MPEG Network" on the show floor. It was perhaps the most complex*

[1] Some readers may take issue with “world’s most complex”. As far as we know, this was the first TV plant that carried interactive and immersive audio and the first to show how such audio could be distributed to local affiliates.

combination of broadcast audio content ever made in a single plant, involving 13 different formats. The network was designed to handle immersive audio in both channel and HOA-based formats, with each using audio objects for interactivity. Live mixing at a simulated sports remote was contributed to a network operating center, with distribution to affiliates, and then emission to a consumer living room, all using the MPEG-H audio system. Both system and equipment design are presented, including an Audio Monitoring and Authoring Unit to mix signals using existing consoles.

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

Introduction – The Challenges of Next-Generation TV Audio Systems for TV Facilities

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 “next-generation” audio systems (NGA) to provide additional features or performance beyond those offered today. One of these systems is the MPEG-H TV Audio System developed by Fraunhofer and its partners which is now part of the ATSC 3.0 Candidate Standard. MPEG-H will be the first next-generation audio system used in over-the-air broadcasting when Korea begins ATSC 3.0 broadcasts in Spring 2017.

Early in the development of MPEG-H, we realized broadcasters would face several challenges we would have to help them overcome to fully transition to next-generation audio:

Immersive audio may be produced in channel-based or Higher-Order Ambisonics formats of varying complexity and will be interspersed with legacy content in stereo or 5.1 channels. It is likely that for full NGA production in the future, a complex mixture of formats will be present in a broadcast facility and existing techniques such as fixed channel rundowns may have to be abandoned.

The interactive features of NGA, which are enabled by audio objects transmitted separately from a channel bed or HOA components, are likely to be different from one program to the next. A SDI audio channel carrying an “away team announcer” in a football game might carry a “pit crew radio” at an auto race.

Thus, metadata for each program needs to label the objects for the viewer, set any broadcaster-authored limits on viewer adjustments, and establish any “preset” mixes for the casual viewer. This metadata can be used to identify which audio channels of a SDI signal or media file contain the objects and the overall channel format or HOA configuration used.

In previous papers, we outlined a four-stage transition process for broadcasters to move to immersive audio program production that could involve automation-triggered encoder settings or simple XML files to accomplish this in the initial stages. However, to fully utilize the system in its most elaborate form in our final stage, it is necessary to carry frame-by-frame metadata for the position of dynamically panned objects, synchronized with the audio program itself. For live broadcasts, this requires a reliable data path for metadata from the remote truck through the entire contribution and distribution signal path to the viewer’s TV.

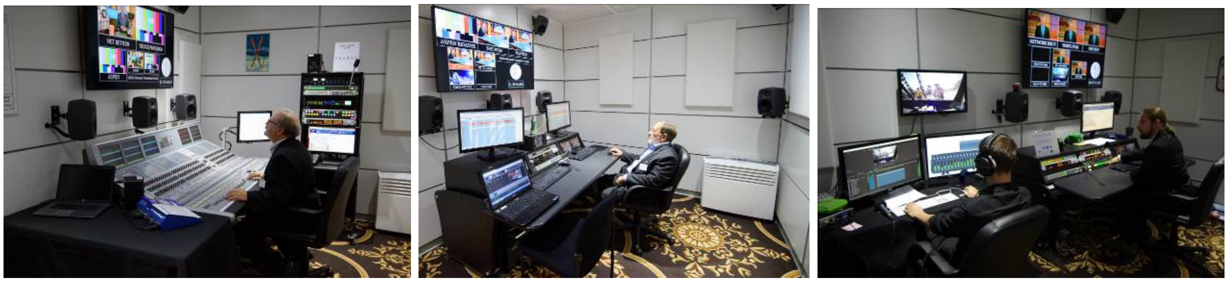

Figure 1. Professional rooms of the test bed: Remote Truck, Network Operations Center, Local Affiliate (l to r)

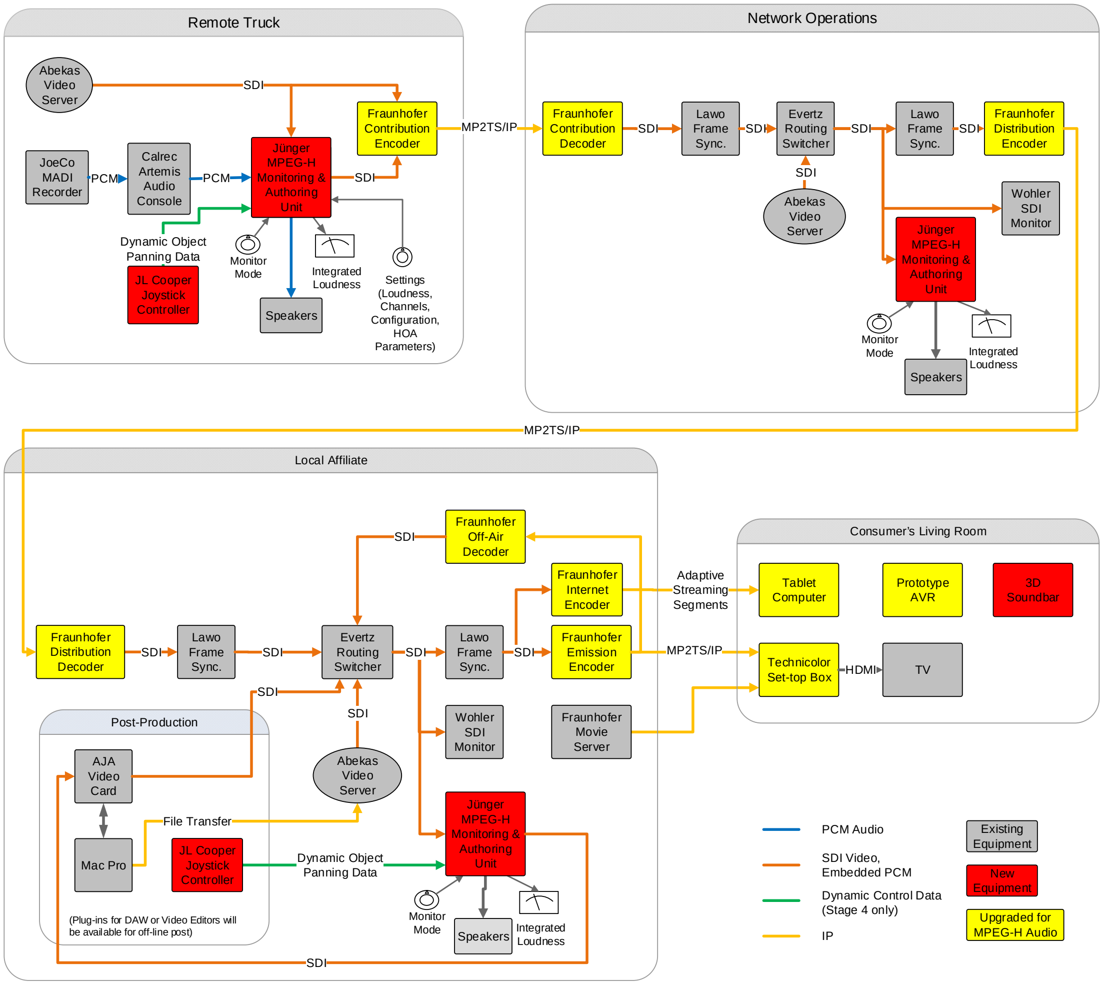

Figure 2. Block Diagram of the Test Bed

While the implementation of IP-based video routing in TV plants would make transmission of metadata easier, we could not assume this would be completed throughout the industry before immersive audio production was deployed. Thus, our design philosophy was to plan on using the existing SDI plant infrastructure, audio consoles, recorders/servers, and post-production equipment that broadcasters were likely to already have.

Another issue is that, except for film consoles, no equipment existed to prepare immersive audio mixes. Existing post-production tools outside of theatrical cinema stages could accommodate a maximum of 5.1 or 7.1 audio. No immersive loudness monitoring or way to support the monitoring of different consumer presets existed nor was there a way to transmit immersive audio in contribution or distribution.

Building a Test Bed - “The MPEG Network”

Concurrent with our development of MPEG-H, the ATSC and other standards organizations were considering the future implications of NGA and potential selection of NGA systems. Thus, we decided to build a test bed that we could use to investigate the issues, prove our proposed solutions, and demonstrate a working system to the industry during this period. The test bed was planned and constructed in 2014-15 and was first shown at the NAB 2015 convention and then in a special event organized by the ATSC in August 2015.

The test bed was based on the operation of a hypothetical broadcast network, “The MPEG Network” with local affiliate station WMPG-TV. A separate division of the MPEG Network operated a cable movie channel.

We organized the demonstration into four rooms:

- A simulated remote truck where pre-recorded microphone signals from an extreme sports event were mixed and panned live on an unmodified console equipped with a MPEG-H Audio Monitoring and Authoring Unit, explained below.
- A Network Operations Center where the live truck signal was switched under automation control with recorded programming from a video server.
- A local affiliate, WMPG-TV, where local commercials were produced and inserted into the network feed.
- A consumer living room where the content was played back on a set-top box and prototype 3D sound bar. Movies mixed in Dolby Atmos format could also be received from the MPEG Cable Movie Channel on the set-top box.

The test bed was designed to make use of existing commercially available equipment where possible and to be operated by American broadcast professionals. Demonstrations were run continuously on a fixed schedule by playout automation in the NOC and WMPG-TV.

In order to illustrate the different modes and possibilities of the MPEG-H system, we prepared content in 13 different formats, as shown in the table below.

Table 1. Program Log Showing Audio Formats Used in the Test Bed (H indicates overhead Height channels, statO indicates static objects, dynO indicates dynamic (moving) objects). The schedule repeated every twenty minutes during the demonstrations.

MPEG-H Audio Alliance - ATSC 2015 Live Broadcast Demonstration - Combined Program Log/Rundown

	Format	Presentations	Language/Dialog	Program Log at each demo location			Feature Shown	Seconds	Content ID
				Remote Truck	MPEG Network	WMPG			
				Aspen	New York	Birmingham			
Intro	2.0	Broadcast	MOS		Opening Title			20	A 180
	7.1 + 4H	Broadcast	ENG		Introducing Demonstrations		How to use system, immersive studio production (speaker chimes)	196	A 160
	2.0	Broadcast	MOS		Show Title			14	A 181
	5.1 + 4H + 2dynO	Broadcast	Music Only		Network ID Long		Dynamic Objects, immersive production	12	B 141
Network Show	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - opening segment		Dialog level of guest and host adjustable seperately	202	C 142
	5.1	Broadcast	ENG		PB: Big Air - Host Mix		Comparison to ATSC 1.0	51	D 161
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - setup of H mix			9	E 162
	HOA + 1statO	Broadcast, Dialog+, Live	ENG		PB: Big Air - MPEG-H Version		HOA for sports	51	F 144
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - half-pipe setup			82	G 145
	5.1	Broadcast	ENG		PB: Half-pipe - Host Mix			41	H 168
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - setup of half-pipe live H mix			21	I 167
	5.1+4H + 3statO + 1dynO	Broadcast, Dialog+, Live	ENG(Network), ENG(Venue), NOR	Half-pipe (live)	Cut to Aspen - live mix of half-pipe		Channels+Objects Immersive for sports: Dynamic Objects + 3 Languages mixed live in the truck	41	live
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - throw to commerical			27	K 147
Local Break	5.1	Broadcast	ENG		National Spot - AAA			60	L 149
	5.1+4H	Broadcast	ENG		Network Cover: Technicolor Promo	WMPG ID - WeatherCenter 84	Immersive sound in affiliate production	5	
	3 x 2.0statO	Broadcast, Dialog+	ENG, SPA, CHI			Local spot #1 - Crown Nissan	Loudness Control, Preferred Language	30	N 150
	3 x 2.0statO	Broadcast, Dialog+	ENG, SPA, CHI			Local spot #2 - airbag lawyer	Extend reach with additional voiceover, Preferred Language	30	O 151
	5.1+4H + 2dynO	Broadcast	Music Only		Network ID Short		Dynamic Objects, immersive production	7	P 152
Network Show	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - setup NASCAR			76	Q 153
	5.1 + 5.0statO + 4 x 1.0statO	Broadcast	ENG, ITA		PB: Nascar		multi-channel objects, team radio, Preferred Language	58	R 154
	2.0 + 2statO	Broadcast, Dialog+	ENG		SportsTech Show - close			68	S 155
Break	HOA + 2statO	Broadcast	ENG, CHI		National Spot - Qualcomm: SnapDragon		HOA for commercials, Preferred Language	60	T 148

Using all of these formats would require metadata to be carried with the content to identify each format, a method of getting dynamic panning metadata from a live event to the home was needed, a way to encode the audio for contribution and distribution was required, and there would need to be a frame-accurate transition between each content item. (Each item was played from a different server channel to simulate in-plant switching between sources)

This required several innovations to be developed in our system:

Adapting Live TV Audio Consoles – The Audio Monitoring and Authoring Unit

Today, no TV audio consoles support audio beyond 5.1 channels[2]. Audio consoles for remote production are also surprisingly complex, given the mixing and routing needed in sports productions, and are expensive to replace. Thus, we developed an accessory monitoring unit product in collaboration with Junger Audio to transform an existing 5.1 console into one suitable for MPEG-H production. The monitoring unit replaced the monitor controller of the console with one that would drive 12 speakers and provide downmixes of immersive sound formats.

In addition to immersive sound, the MPEG-H system also offers interactivity by sending some audio objects on separate channels for viewer mixing. It is envisioned that broadcasters may prepare some "preset" mixes in addition to the default mix for the casual viewer to quickly choose from. An enthusiast viewer can use advanced controls to adjust elements to make his own mix, within limits set by the broadcaster. Thus, we included separate guide metering for rapid checking of the loudness on each preset, as well as switching to allow monitoring each preset mix.

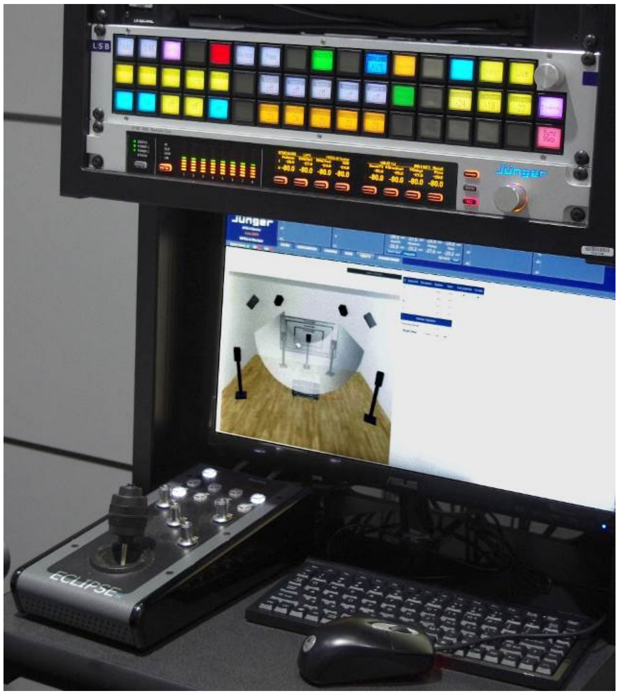

Figure 3. Junger Audio MPEG-H Audio Monitoring and Authoring Unit

The parameters of preset mixes, as well as labels describing them and the related audio objects, have to be created. This involves the authoring function of the monitoring unit, which

[2] Except for special consoles built for experimental use with the NHK 22.2 system.

has a web interface to enter this information, as well as the channel configuration, range limits, and other parameters into Audio Scene Information for the program. The scene information is constantly encoded in the Control Track along with the loudness and dynamic range metadata for each preset.

Another feature needed is the ability to pan objects in three dimensions as needed to track action on screen. The monitoring unit includes a joystick interface for this purpose, which we used in the demonstration by panning sound effects to follow a snowboarder’s movement at an extreme sports event.

Carrying Metadata in the TV Plant – The Control Track Concept

Carrying audio metadata in traditional SDI-based TV facilities has been a challenge. Although standards exist for carrying metadata in the vertical or horizontal interval of the SDI format, much existing equipment does not support them. The other approach is to dedicate an audio pair for carrying a metadata data stream or typically, a complete compressed audio stream such as the Dolby E format. This creates an operational burden as each piece of equipment has to be programmed and maintained to treat the channel pair as a “data mode” signal, where no crossfades, filtering, resampling, or gain changes can be performed that would cause the audio channel to not be stored and transmitted in a bit-exact manner.

In our design, we developed an alternative approach that does not suffer from these limitations. Our observation was that existing technologies for data communications over analog channels could be applied to this issue, resulting in what we internally termed the “metadata modem”. In our system, metadata for the audio signal is collected into packets synchronized with the video signal and modulated with communications modem techniques into a Control Track signal that fits in the audio channel bandwidth. This signal is unaffected by typical filtering, or scaling operations in the audio sections of broadcast equipment. Importantly, resampling of the audio channels, as is common in frame synchronizers and other terminal equipment, does not corrupt the Control Track.

Since the signal includes a guard interval around vertical sync, frame accurate transitions in the audio are automatic as the metadata switches simultaneously with the audio without corruption. It is also possible to use the Control Track in an untimed plant or with crossfades in a routing switcher with 1-2 frame accuracy.

As the Control Track is just a timecode-like audio signal, it can also be carried as another audio track in audio or video editing systems. In our demonstrations, we showed creation of a highlight reel of content in several audio formats in a video editor just by selecting clips and dragging them to a new timeline. One clip might be a broadcast in channel-based immersive sound, and another in higher-order ambisonics, interspersed with a 5.1 broadcast with several objects. The highlight reel plays seamlessly since the control track from each clip is cut with the audio and indicates the format and assignment of each audio channel in the content.[3]

[3] The Control Track is one of several options for storing immersive audio that works well with existing equipment. In the future, we expect post-production software to support immersive audio file formats such as BWF-ADM.

Switching Formats During Broadcast – Stream Splicing

Since the in-plant audio in this demonstration is always PCM, it can be cut at any audio sample. However, if there is a change in channel format, an audio format transition has to happen at a video frame boundary, since that is the transition point between data packets in the Control Track that indicate the channel assignment and format. Cuts at video frame boundaries are desirable since the audio and video programs will transition simultaneously and this is the normal behavior of routing and switching equipment.

The SDI audio signals thus have seamless transitions during switches or cuts, but this creates a concern for the following MPEG-H encoder. Like all perceptual audio codecs, MPEG-H performs time/frequency transforms on a frame basis, with a frame typically being a multiple of 1024 audio samples. The relations between audio and video sampling clocks, raster sizes, video frame rates, and audio codec frame lengths result in there typically being several seconds between instances of exact frame alignment between audio and video frames. This means that a given codec audio frame would likely contain two different channel formats if there is a cut at a video frame boundary during the audio frame. Unfortunately, the audio encoder can't deal with two channel formats in one audio frame, and this situation has to be resolved.

One approach is to change the audio frame length to a shorter value that aligns with a given video frame rate, so that an audio frame can only have one channel configuration. This works, but leads to reduced audio quality for a given bitrate, as the frequency resolution of the time/frequency transforms is reduced.

Instead, in our system, we encode a frame consisting of the final audio samples before a cut, and another frame consisting of the initial audio samples following a cut, and we send information in the bitstream indicating the appropriate point to transition between them. This requires encoding and sending an additional audio frame in the bitstream and thus causes a temporary peak in the instantaneous bitrate, but this peak is absorbed by the decoder's input buffer, just as peaks from difficult-to-encode audio frames are. This approach offers the ability to cut a program at any video frame boundary while using standard-length audio codec frames for the best coding efficiency.

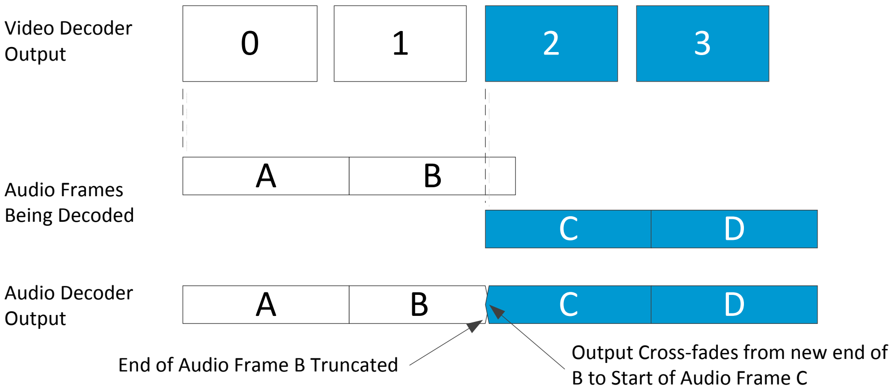
Video Decoder Output
0
1
2
3
Audio Frames Being Decoded
A
B
C
D
Audio Decoder Output
A
B
C
D
End of Audio Frame B Truncated
Output Cross-fades from new end of B to Start of Audio Frame C

Figure 4. Splicing of codec audio frames to accommodate new program with different channel configuration beginning at video frame 2. Audio frame B from the previous program is truncated at the decoder output and the playout of audio frame C begins concurrent with the start of video frame 2.

Connecting Facilities – Fraunhofer Encoders with MPEG-H and AVC

As shown in the block diagram in Figure 2 our setup included 3 A/V links between facilities: A contribution link from the remote truck to the NOC, a distribution link from the NOC to the affiliate, and an emission link from the affiliate to the living room. These links were operated with compressed audio and video, MPEG-H for audio and AVC for video. The compressed audio and video streams were multiplexed into an MPEG-2 transport stream and transmitted over an IP network.

Since no commercial MPEG-H encoders and decoders were available at the time we designed The MPEG Network test bed, we built our own MPEG-H encoders and decoders. All encoders and decoders share the same hardware platform and software framework. The hardware is built from commercial of the shelf components: a standard rack-mounted PC equipped with a Blackmagic Design DeckLink card for SDI I/O. The software is based on the GStreamer framework. GStreamer is a pipeline-based multimedia framework. GStreamer already includes components, called plug-ins, for SDI I/O and AVC video encoding and decoding. So we needed to develop GStreamer plug-ins only for MPEG-H audio encoding and decoding and for MPEG-2 transport stream multiplexing and demultiplexing[4].

Contribution/distribution encoders and decoders and emission encoders and decoders differ only in the configuration of the software framework. Contribution/distribution encoders operate with a fixed channel mapping that encodes 15 audio channels at a higher contribution-quality bitrate. The control track is demodulated and the metadata is transmitted as digital side information in the MPEG-H bitstream. The contribution/distribution decoders decode the 15 audio channels and regenerate the control track from the metadata. The emission encoder encodes the audio to the channel configuration given in the control track and performs configuration changes as required by different content formats. The emission decoder renders the audio to the loudspeaker setup that is connected or to the 3D soundbar.

Playback for Mainstream Consumers – The 3D Soundbar

Although individual loudspeakers provide the highest spatial accuracy and were used in our professional rooms, the expense of installing ceiling-mounted speakers for immersive sound at home likely limits this approach to enthusiasts. Our vision has been to offer a reference design that can be manufactured as a “un-box and listen” product for immersive reproduction in mainstream consumers’ homes. Our initial concept prototype used in the demonstrations was a frame of speakers surrounding the TV.

This has now been reduced to a traditional soundbar form suitable for consumer manufacturing. The soundbar connects to the TV or other program sources over traditional HDMI 1.4 or S/PDIF connections just as stereo soundbars do today and it provides a realistic immersive sound image within a wide listening area.

[4] Although future TV systems with next generation audio will most likely employ different transport formats like MPEG DASH or MMT we choose MPEG-2 TS because DASH and MMT were not sufficiently mature at the time we designed our test bed. Meanwhile broadcast encoders and TV sets implementing MPEG-H audio, HEVC video and DASH/MMT transport are becoming commercially available.

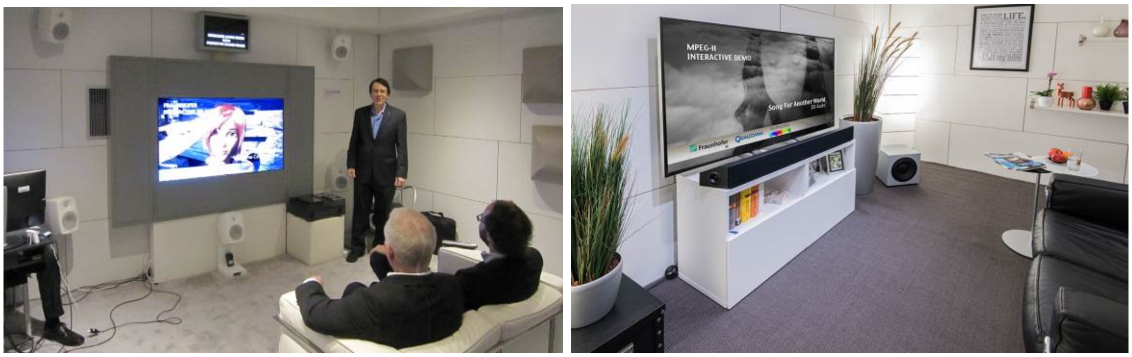

Figure 5. First (left) and second-generation (right) 3D soundbar designs. Wall mounted speakers in the left photo were used for discrete playback to simulate professional monitoring and enthusiast listening.

Conclusions

Construction of this test bed provided us with the opportunity to discover many of the operational issues broadcasters would face with the adoption of next-generation audio systems such as MPEG-H. We were able to test our solutions to these issues on typical broadcast equipment under normal operating conditions to verify that they worked and demonstrate this to the industry. We learned:

- Immersive audio program production can be implemented in traditional broadcast plants with only minor equipment and operational changes. In our test bed, only new video/audio encoders and decoders and the MPEG-H Audio Monitoring and Authoring Units were new equipment. Except for adding additional speakers to control rooms, the standard equipment used for digital broadcasting today was used unchanged.
- Production and delivery of next-generation audio content to the consumer is possible today, as proven by the test bed. Only the RF and transport layers of new TV standards were not considered in our end-to-end tests.
- Evolving use of Next-generation audio will likely eventually lead to a multitude of channel formats in the plant, particularly with object-based interactivity. The Control Track approach demonstrated the ability to transport metadata to transparently manage all of these formats.
- It is possible to do live mixing of next-generation audio, even with dynamic panning of audio objects.[5]
- The AMAU and time code-like Control Track allows audio and video post-production tools to work with NGA audio today, without waiting for full support of NGA to appear in future versions of these products.

[5] Although simulated PL intercoms were reproduced in the truck, the workload of the mixer in dealing with other audio issues was not simulated. Longer or more complex use of NGA for important events may require an A2 or assistant mixer.

Acknowledgements

The authors would like to acknowledge the contributions of Bernd Czelhan, Matteo Agnelli, Nicolas Faecks, and Jim Hilson of Fraunhofer in building and operating the test bed, the support of the Fraunhofer tonmeister group including Hanne Stenzel, Ulli Scuda, and Valentin Schilling in mixing the content, our colleagues Peter Baum, Peter Jax, Claude Gagnon, and Ben Wilkins of Technicolor, and the support of Junger Audio, Calrec, Aveco, Evertz, and Lawo in providing and integrating the test bed equipment.

References

1. “A/342 Part 3: ATSC Candidate Standard – MPEG-H System”, www.atsc.org, May 2016.
2. 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.
3. R. Bleidt, A. Borsum, H. Fuchs, M. Weiss, “Object-Based Audio: Opportunities for Improved Listening Experience and Increased Listener Involvement”, SMPTE Motion Imaging Journal, Volume: 124, Issue: 5, July/August 2015.
4. R. Bleidt, “Installing the MPEG-H Audio Alliance’s New Interactive and Immersive TV Audio System in Professional Production and Distribution Facilities”, Fraunhofer Technical Brief, www.iis.fraunhofer.de/tvaudio, April 2015.
5. 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.
6. MPEG-H Audio Alliance Live Broadcast Demonstration at NAB 2015 (video tour of the test bed), Fraunhofer IIS, https://youtu.be/wnBx9SjOOII