electronic design

print | close

MPEG-H Audio Brings New Features to TV and Streaming Sound

Electronic Design
Robert Bleidt
Fri, 2015-07-10 12:29

With the introduction of digital TV and DVDs at the end of the 20th century, consumers were able to experience surround sound at home for the first time. Blu-ray discs brought 7.1 surround and lossless audio to living rooms, where audio-video receivers and home-theater-in-a-box systems became popular.

In this decade, it might be argued TV sound quality has deteriorated. When mainstream consumers purchase a flat-screen TV, it's usually supplemented with a stereo sound bar instead of surround speakers. Or, instead, today's consumer may prefer watching on a tablet computer with its one or two half-inch speakers. On the other hand, many consumers now purchase full-size headphones for mobile listening, at price levels once reserved for professional or audiophile users. So, consumers still appreciate sound quality—if it's available in a convenient and fashionable way *(see "How is TV Sound Mixed These Days?")*.

Over the next few years, TV and Internet streaming viewers will experience a resurgence in sound quality, with the opportunity to enjoy previously unavailable features. This will be buttressed by the introduction of TV standards that include new audio systems now under development.

Related

Advancing the Audio Interface for Mobile and Mobile-Influenced Designs

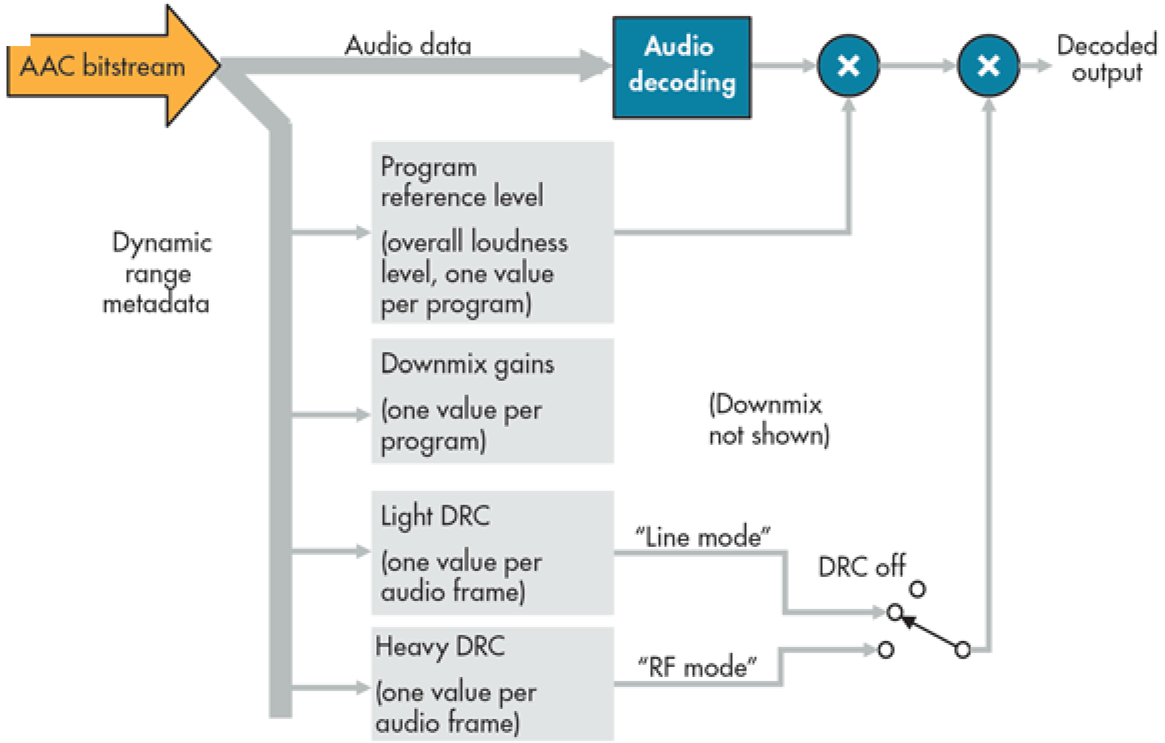
Understanding MPEG Audio Codecs From mp3 To xHE-AAC

Ray Dolby: A Breaker Of Sound Barriers

One of these systems is based on the recently approved MPEG-H Audio standard from the MPEG group of ISO/IEC (ISO/IEC 23008-3), which is being commercialized by the MPEG-H Audio Alliance of Fraunhofer, Technicolor, and Qualcomm. This system will offer the consumer three primary new features: interactive audio, immersive sound, and universal delivery. To explain them, let's first examine today's modes of sound transmission.

Current TV Sound Offers 5.1 Channels

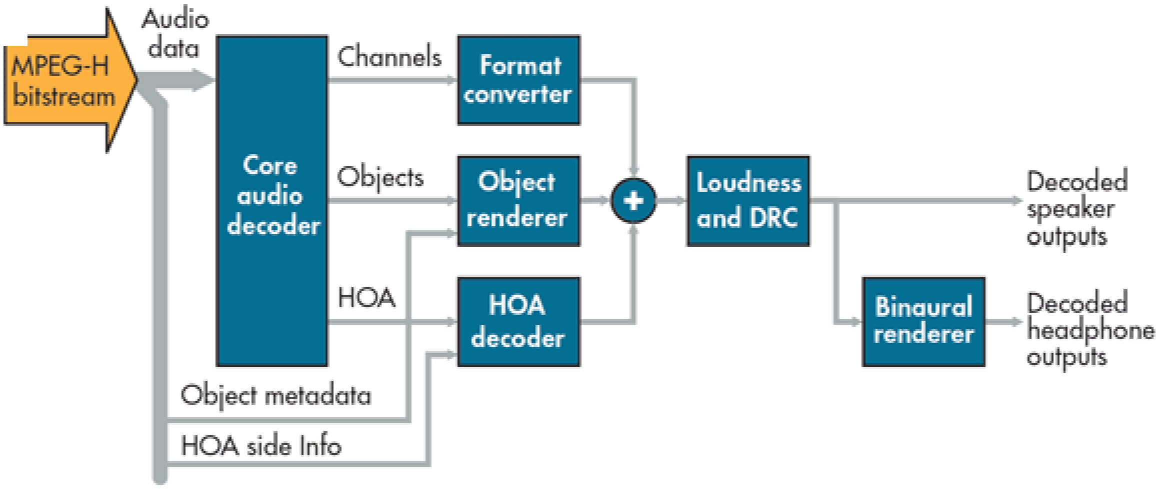
In current TV broadcasts, audio is transmitted to the home using either Dolby's AC-3 codec in countries using the ATSC standard, or MPEG's Advanced Audio Coding (AAC) codec in countries using the ISDB standard. Europe and several other countries use the DVB standard, which includes both AC-3 and AAC. In the U.S. ATSC system, most major sports and dramatic programs are carried in 5.1 surround. Although Android, iOS, Windows, and MacOS have supported AAC surround sound for several years now, operating just as it does for TV, most audio on web-based content today is stereo.



TV audio today is transmitted using a separate channel for each 5.1 loudspeaker. For stereo or mono listening, the TV's audio decoder downmixes the five channels to two or one using standardized gains *(Fig. 1)*. Both codecs also send extra gain coefficients to scale the audio envelope for less dynamic range if preferred by the viewer. Loudness metadata is sent so that the TV's decoder can adjust the overall gain of each item. However, few broadcasters use this feature today, opting to normalize the loudness of each item before transmission.

Controlling the Audio Mix via Objects

In next-generation TV audio systems such as the MPEG-H Alliance system, a more advanced renderer replaces the downmix process *(Fig. 2)*. These new systems include the ability to send audio objects—signals independent of a loudspeaker—to the TV's decoder as well. Unlike channels, objects have a specified 3D position and gain. Audio scene-description metadata controls whether the user can change the gain or position, or if the object is a part of one or more presentations. The metadata can also update the object's position over time, enabling the sound source to move for artistic effect or to track visual action.



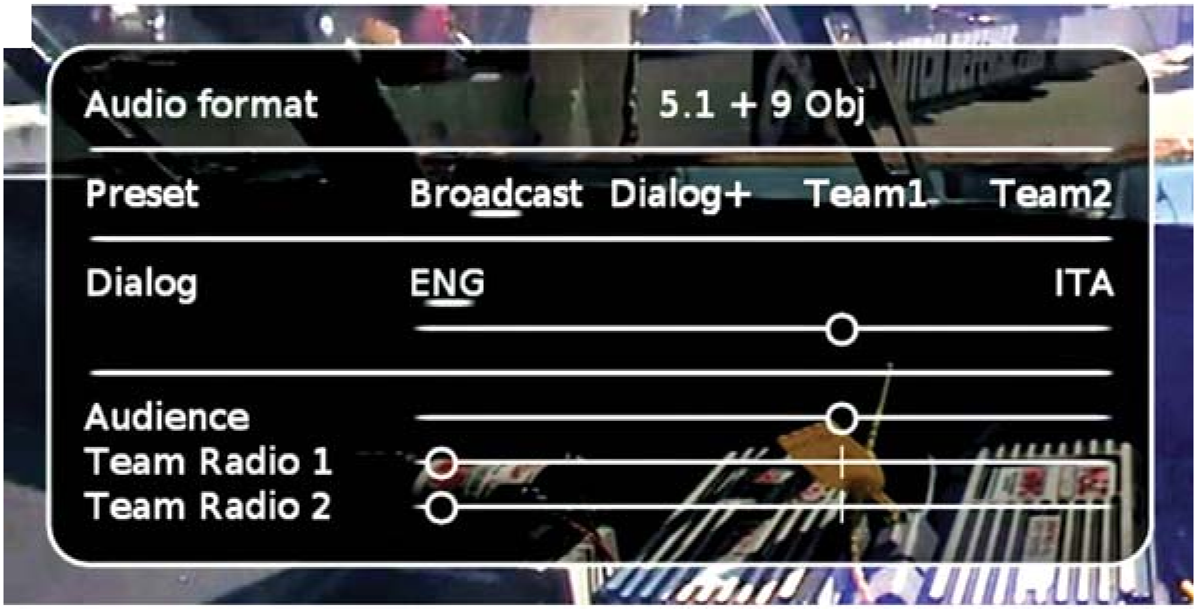
Objects are mixed into loudspeaker channels by a renderer that considers the object's position in azimuth and elevation to determine signal amplitude in each speaker. The most common rendering algorithm is "vector-based amplitude panning," which uses the three nearest speakers to an object's position to render it.

In the latest cinema sound systems, objects allow sounds to accurately track the sound mixer's intended presentation, despite the varied speaker arrangements in cinemas around the world. Most cinemas are retrofitted with extra speakers over the audience to provide sounds from above for greater realism, in addition to the speakers along the walls and rear of the theater. They formally reproduce the same signal, but can now be addressed separately.

A typical film consists of a "channel bed" of seven full-bandwidth channels plus the low-frequency effects (LFE) channel, which is then supplemented with objects. Building a mix around a bed simplifies the sound design for releasing the film with a sound track for legacy theaters or DVDs.

Home listening doesn't present the same acoustic problems like those in a theater. Moreover, TV broadcasting is envisioned to use objects primarily for interactivity, although dynamic (moving over time) objects do offer the potential for creative special effects.

Since objects can be controlled in gain or position, broadcasts can be made interactive. The most fundamental use of objects is to allow consumers to change the level of dialogue in a program without affecting the loudness of the other program elements *(Fig. 3)*. This feature has been extremely well received by consumers in pioneering trials Fraunhofer conducted with the BBC.



Audio format 5.1 + 9 Obj

Preset Broadcast Dialog+ Team1 Team2

Dialog ENG ITA

Audience

Team Radio 1

Team Radio 2

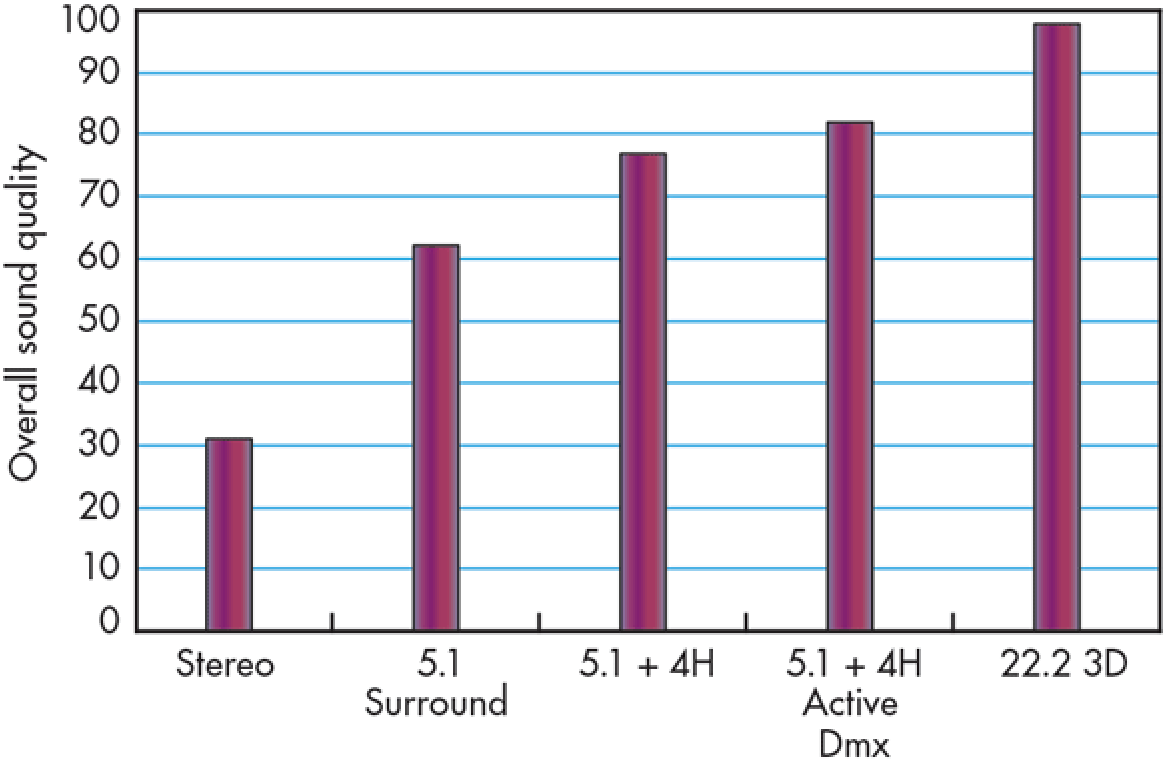
In one experiment that took place in 2011, a panel of about 2000 consumers listening to the Wimbledon tennis matches were able to adjust the volume of the commentators relative to the natural sound from the court, players, and audience. The distribution of consumer preferences was bimodal. Consumers preferred a level several decibels above or below that used in the broadcast, though the mean matched the broadcast level, validating it was the best compromise for the legacy broadcast system.

Objects also enable the transmission of additional languages. Current TV audio systems only allow for one additional language track, and many countries now require this track to be used for descriptions for the visually impaired, thus limiting broadcasts to one language. Each additional language may be sent using a 20- to 40-kb/s mono audio object.

Furthermore, objects can be used to enhance a broadcast by adding elements that would not be possible to put in the default mix. For example, Fraunhofer has proposed using objects to carry commentary favoring each team at a sports event, or to listen to athlete or official microphones, or listen to pit crew to driver radio traffic at auto races.

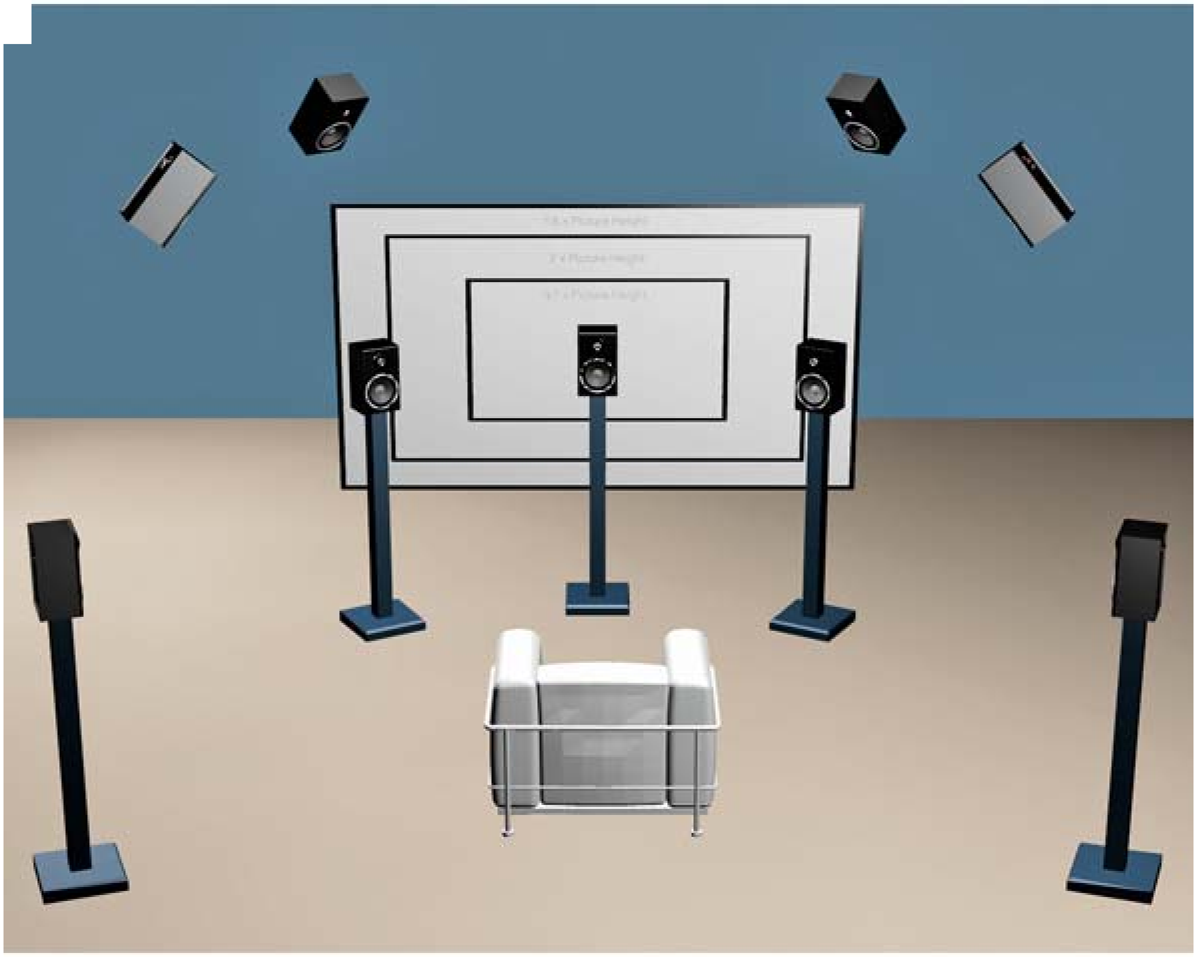
Become Part of the Audience with Immersive Sound

A second feature of new TV audio systems is immersive sound. This improves the realism of reproduced sound by adding sound from above the listener. Listening tests at Fraunhofer have shown this can substantially improve the perceived realism of sound, almost as great as the step from stereo to 5.1 surround *(Fig. 4)*.



The MPEG-H-based TV audio system offers three methods for transmitting immersive sound: traditional speaker channels, audio objects (as used in cinema sound), or a scene-based coding technique based on Ambisonics. Typically, two or four new speaker channels are added above the corner speakers to reproduce sound from above. These may be driven directly by channel signals or audio objects may be rendered to them. The higher-order Ambisonics method doesn't use channels, but instead sends a spatially compressed set of signals thatdescribe a sound source's direction by means of their relative amplitudes and polarities *(see "Higher-Order Ambisonics")*.

A typical immersive broadcast would send the standard 5.1 surround channels plus four additional height channels, otherwise referred to as 5.1 + 4H *(Fig. 5)*. It's also possible to use only two front height channels, or use the 7.1 Blu-ray surround configurations with two or four height speakers.



While ceiling-mounted loudspeakers offer the ultimate immersive sound quality, the consumer trend toward flat-screen displays have led them to abandon discrete speakers for soundbars. This has actually reduced the achievable sound quality, as almost all soundbars are stereo devices.

Fraunhofer has developed a potential solution to this problem with 3D soundbars that have additional speakers and acoustic processing, allowing an array of speakers surrounding the TV to produce realistic immersive sound *(Fig. 6)*. This offers the mainstream consumer a way to experience good immersive sound without any additional wiring or configuration—a décor-friendly, hang-it and hear-it experience.

Universal Delivery Maximizes Sound on All Devices

In the early days of HDTV, consumers watched TV broadcasts on television sets while streaming video was consumed on PCs. Today, viewers may watch streaming video on their smart TV, and in some limited cases, watch broadcast TV on mobile devices. While broadcasters, phone manufacturers, and mobile carriers in most regions have yet to agree on a true over-the-air delivery of broadcast signals to phones, broadcast signals conveyed or retransmitted over the Internet are becoming commonplace.

Today's audio codecs were designed to offer only limited adaption to the different listening environments faced by consumers through the use of dynamic-range control gain values. These are coefficients sent to each audio frame to scale the audio envelope for less dynamic range. A "light" set is used to reduce the dynamic range of content intended for home-theater presentation to a range suitable for casual listening, and a "heavy" set, sometimes seen as "night mode" on consumer menus, handles loud sounds that might disturb neighbors.

These dynamic ranges were appropriate for the era of listening in living rooms with 40- to 45-dBA ambient-noise levels, but consumers now commonly listen on the street or in flight at 75- to 85-dBA noise levels. Such levels tend to overwhelm the speakers in mobile devices, and earphones or headphones are often used, offering anything from 0-dB acoustic isolation to perhaps 35 dB on the most tightly sealed in-ear models. Not all consumers are willing to adapt to earplug-style listening, or carry full over-the-ear headphones, so we must adapt the audio signal to this noisy environment without damaging their hearing due to a too-loud program.

Mobile devices also are used to consume music, and unlike the video industries, music producers have continually tried to make their songs louder than any others. What results is music recorded at a level that's 10 to 15 dB louder than video programming. A compromise solution to this difference is to boost the level of video programming on mobile devices to match that of music.

h of these issues—reducing dynamic range and matching video to music loudness—are addressed in the EG-H system through additional sets of dynamic-range control coefficients sent in the bitstream. The system also offers special processing of the dialogue to increase the intelligibility in very noisy environments.

In addition, since earphones and stereo speakers are the most common modes of listening on mobile devices, the system includes binaural rendering to create a realistic immersive listening experience on headphones through psychoacoustic techniques. It's also designed to work with products such as Fraunhofer Cingo, which extends this virtual listening experience to tablet speakers as well.

ATSC 3.0: The First TV Standard with Immersive and Interactive Sound

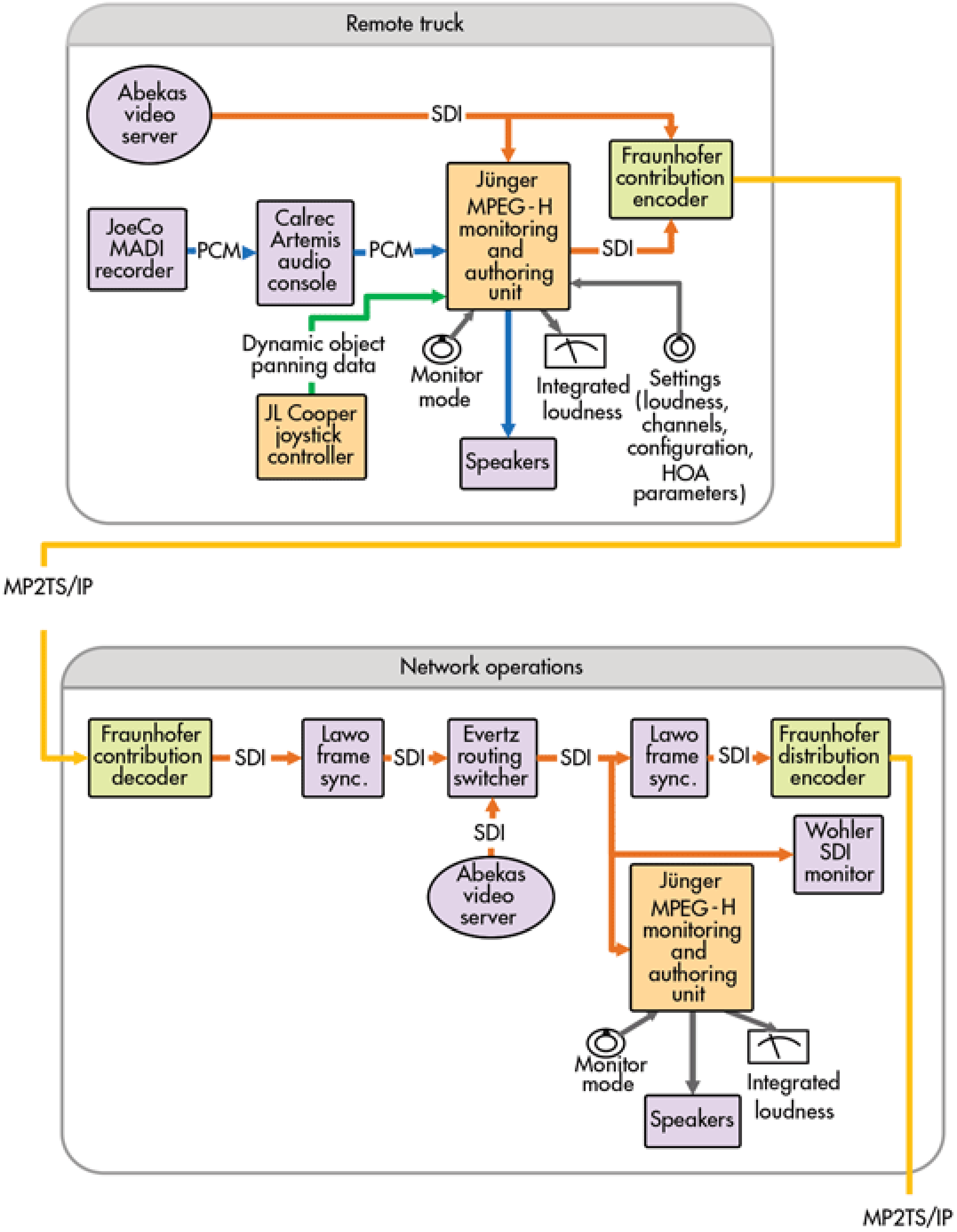
All three of the world's TV standards families have immersive sound in their roadmap. ISDB plans on using 22.2 channel non-interactive audio based on AAC coding, while DVB is currently studying the requirements for both interactive and immersive sound. ATSC is perhaps the most advanced in this process, currently considering proposals from the MPEG-H Audio Alliance and Dolby for the ATSC 3.0 standard, which is expected to be released this year.

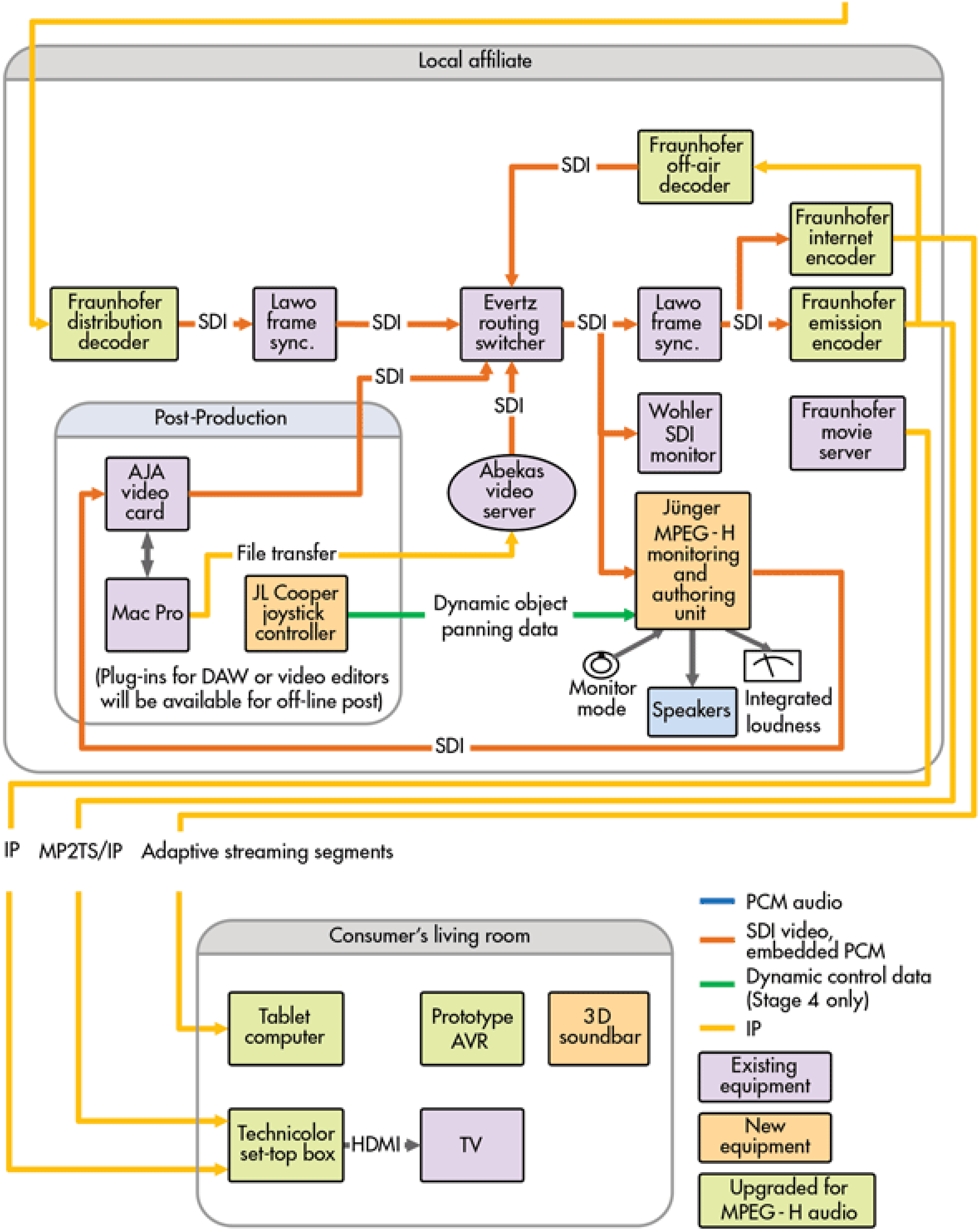
The design requirements for ATSC 3.0 include both immersive sound and interactive elements. Tests consider both 5.1 + 4H and 7.1 + 4H channel formats, as well as the MPEG-H higher-order Ambisonics audio scene capture mode.

ATSC 3.0 also incorporates both static and dynamic (moving) objects. On top of that, there's a requirement to send some objects over Internet or broadband channels and combine them in the TV with other audio elements received through broadcast channels. Other requirements involve flexible rendering to correct for misplaced or non-standard speaker arrangements.

On the Air Live at the 2015 NAB Show

In April 2015, the Alliance demonstrated a complete implementation of an MPEG-H Audio-based broadcast TV system at the National Association of Broadcasters (NAB) show in Las Vegas *(Fig. 7)*. The demonstration traced the production of a live sports event, from mixing the sound in a simulated remote truck, switching it through a TV network operations center, distributing the network signal to a local affiliate station, and then transmitting it to a viewer's home—just like a normal broadcast practice.


Remote truck
Abekas video server
SDI
Fraunhofer contribution encoder
JoeCo MADI recorder
PCM
Calrec Artemis audio console
PCM
Jünger MPEG-H monitoring and authoring unit
SDI
Dynamic object panning data
JL Cooper joystick controller
Monitor mode
Speakers
Integrated loudness
Settings (loudness, channels, configuration, HOA parameters)
MP2TS/IP
Network operations
Fraunhofer contribution decoder
SDI
Lawo frame sync.
SDI
Evertz routing switcher
SDI
Lawo frame sync.
SDI
Fraunhofer distribution encoder
SDI
Abekas video server
Wohler SDI monitor
Jünger MPEG-H monitoring and authoring unit
Monitor mode
Speakers
Integrated loudness
MP2TS/IP



The remote truck used an existing 5.1 channel Calrec console adapted for mixing immersive and interactive programming by a new Junger Audio MPEG-H Audio Monitoring and Authoring Unit. The monitoring unit accepted the audio objects as separate signals, which are not mixed into the console's main 5.1 output, and accepted the height channels as a separate bus output from the console. The monitoring unit included a joystick-controlled three-dimensional panner to allow for dynamic panning of one of the objects.

One of the challenges of sports production is creating an audio mix during the furor of a live broadcast. The audio mixer is expected to create good sound while simultaneously listening to the broadcast's producer and director over the intercom, coordinating with staff on the field to track action and troubleshoot problems, as l as mixing in music, sound effects, and replays. On top of these jobs, he must also maintain the loudness of program to meet CALM Act legal requirements. This task is complicated by the future need to monitor loudness on several presentations at once. The MPEG-H monitoring unit helps simplify this process by providing guide meters that offer a quick visual reference to each presentation's loudness level.

Another challenge involves monitoring how the program will sound on various receiving devices. Viewers may listen to the program in 12-speaker media rooms, or on a tablet computer with one speaker. With the monitoring unit, the mixer can hear exactly how the program will sound with any downmix or dynamic-range control applied.

The system also needs a way to specify whether the program is channel- or scene-based, and determine the position, gain, and adjustment limits of any objects. To give casual viewers an easy choice, the broadcaster may prepare presentations that include default settings for all of this metadata. Typical presentations used in Fraunhofer's field tests include a traditional TV mix with normal commentary, a "live" mix with no commentary, or a mix with the venue PA commentary.

The audio mixer inputs all of these specifications into the monitoring unit's web interface to create control data that's sent to the MPEG-H encoder, either in the video encoder used to transmit the program to the network's operations center, or in the video encoder that transmits the network signal to affiliate stations or distributors. At the NAB show, an H.264 contribution encoder was used to send the signal over an IP connection to the network.

At the network operating center, simulated at the show using standard broadcast video servers, routers, and automation equipment, the signal from the live remote was decoded and combined with stored programming and commercials. The network's output was then encoded with MPEG-H and H.264 for distribution to affiliates. At the show, a local affiliate station received and decoded the network signal back to HD-SDI, switched in local commercials, and re-encoded the signal for transmission to a listening room simulating the viewer's home.

New Audio Decoders Will Need More MHz

Implementing this new audio system is straightforward in professional equipment *(see "Clever Engineering Sends Data Stream Through Audio Paths")*, but offers some challenges for consumer devices. The MPEG-H codec is more complex than today's codecs, such as HE-AAC, and to implement new features like interactive or immersive sound will require the transmission and decoding of more channels.

For the most elaborate programming with immersive and interactive sound, a decoder would need to decode six surround channels, four height channels, and perhaps four mono objects—a total of 14 mono coding channels. Some of the objects, such as for language or dialogue, may be mutually exclusive and not need decoding. For most TV broadcasts today, 15 channels is a practical limit due to the HD-SDI infrastructure used in TV production.

Decoding 14 or 15 channels will take at least three times the computing power of today's 5.1 decoders. However, 5.1 AC-3 or AAC audio was designed for decoding using 20th century technology, namely 80-MHz processors, while today we have 2-GHz multicore embedded CPUs. Fraunhofer is currently working to port MPEG-H decoders to popular digital-signal processors (DSPs) from companies such as Tensilica, Cirrus, Texas Instruments, and Hexagon, as well as ARM and x86 CPUs.

Design Challenge: Frame Accuracy

One of the challenges of designing a TV audio system is the need to allow cuts or transitions in the program at eo-frame boundaries. TV audio operates with a 48-kHz sampling rate, while the video is at multiples of 25 in former PAL countries and 30 (1001/1000) Hz in countries (including the U.S.) formally using the NTSC analog TV standard. At a nominal 30-Hz frame rate, there's a five-frame period between points where audio and video samples align. Audio codecs need to operate with frames of their own, accumulating samples for time-frequency transforms. Typically, an audio codec would use frames of 1024, 2048, or 4096 samples. It can take minutes of a bitstream for audio frames to align with video frames.

Some competitive systems have proposed using audio frames of shorter length to align with video frames, but this limits the coding efficiency (audio quality for a given bit rate) due to the reduced frequency resolution. Further, audio codecs typically shift dynamically between frame sizes to adjust the tradeoff between time and frequency resolution to match the character of the audio signal.

The Alliance system includes a new technique that allows arbitrary audio-frame lengths, but splices the ending audio frame of a program to the first frame of another. This allows for frame-accurate cuts in the audio at any video-frame boundary.

Broadcasting or Streaming Dramatic Content with MPEG-H

Live productions of sporting events are a high priority for over-the-air networks. In turn, the Alliance has developed equipment to demonstrate real-time workflows. Studio productions, such as TV dramas or feature films, are another important type of content that will benefit from MPEG-H's features.

Many of today's A-list movies are already mixed in immersive sound for theatrical or Blu-ray release. The Alliance developed the capability to automatically encode these films, as well as episodic dramas and other studio-produced content, into MPEG-H versions. Since Technicolor does the sound mixing for many Hollywood feature films, it's been possible to test the complete delivery chain, comparing the sound mix in the mixing stage or studio to that reproduced by MPEG-H with consumer equipment. As a result, it ensures the artistic intent of the film has been preserved.

Bringing MPEG-H to All Media and Devices

New TV audio systems such as MPEG-H not only offers consumers more accurate sound, but the ability to adjust the sound to their preferences and hear it as it can best be reproduced on their devices. We are hopeful that MPEG-H will be standardized by the ATSC later this year, for inclusion in TV sets in 2017-2018. Simultaneously, we expect many of today's IP delivery platforms to be upgraded with MPEG-H audio decoders soon, as IP-based content distributors begin to offer content with next-generation sound.

To enable this, Fraunhofer will ensure that good encoder and decoder libraries, as well as test content and verification tools, are made available to developers. We will also work with our partners Technicolor and Qualcomm to ensure that all of the affected industries are supported.

Source URL: http://electronicdesign.com/communications/mpeg-h-audio-brings-new-features-tv-and-streaming-sound