Áudio Imersivo

Por Sérgio Eduardo Di Santoro Bruzetti

Foto: Globo/Divulgação

Análise e estado da arte de uma nova forma de sonorização

Da mesma maneira que o vídeo tem evoluído com a adoção de mais pixels através do aumento de resolução (4K, 8K) e melhores pixels (com a adoção do HDR, já descrito anteriormente), o áudio também tem progredido de forma a oferecer ao telespectador uma experiência cada vez mais imersiva, como se estivesse inserido na cena apresentada em sua tela.

Após a captação do áudio em um único canal (sistema monoaural), a primeira tentativa de reproduzir áudio imersivo provavelmente começou com o áudio estéreo. O que faz todo o sentido, uma vez que nosso sistema auditivo se compõe de dois ouvidos, e, portanto, dois alto-falantes fariam com que nos sentíssemos imersos nos sons reproduzidos. Ou também a técnica binaural para fones de ouvido.

A ideia com o áudio binaural é gravar a fonte de áudio com microfones posicionados nos ouvidos de uma cabeça humana real ou artificial, de modo que o mecanismo de localização natural da cabeça (resumido com o termo ***Head Related Transfer Function*** ou Função de Transferência Relacionada à Cabeça, numa tradução livre) é codificado no áudio gravado.

Fonte: Dummy Head KU 100 da Neumann
https://en-de.neumann.com/ku-100

Alguns exemplos de captação binaural podem ser vistos nos seguintes links:
https://www.youtube.com/watch?v=pxYNdvwPkbI
https://www.youtube.com/watch?v=IUDTlvagjJA
https://www.youtube.com/watch?v=Yd5i7TlpzCk

Mas, afinal, o que é áudio imersivo?

O áudio imersivo entrega ao usuário uma experiência muito mais próxima da realidade, permitindo ao telespectador ouvir sons vindos de todas as direções, o que soa muito mais natural para o cérebro humano. Isto é feito pela adição de uma terceira dimensão de sons que vem de todas as direções e que pode envolver pelo menos três partes: canais, ambisonics e objetos de áudio. (ref.: *Immersive Audio: A Look Inside the Next Generation of Sound* – **Erminia Fiorino**

Em um áudio 2.1, 5.1 ou 7.1 convencionais, o profissional que está mixando o conteúdo está limitado à quantidade de canais disponíveis em seu sistema de produção, e ele tem que pensar a cena sonora limitado as eles. Exemplo: caso o sonoplasta queira um som de um tiro que ocorreu fora do campo visual do vídeo, ele pode escolher direcionar o som desse tiro para a caixa traseira esquerda ou direita ou, utilizar recursos de fase e "enganar" o telespectador, direcionando esse som entre a caixa traseira direita e esquerda, sem muita precisão de localização.

Já nas tecnologias de áudio imersivo encontradas no mercado, a mentalidade de quem está mixando ou produzindo muda radicalmente, pois, apesar do conceito de canal ainda estar presente, esse é utilizado apenas como base na mixagem. A introdução do novo conceito de objeto de áudio, permite ao sonoplasta ter uma liberdade artística incrível, levando mais realismo ao telespectador.

Abaixo você poderá conhecer um pouco mais sobre esses recursos

Canais

As caixas de som representam os canais disponíveis. Por exemplo: um sistema estéreo tem dois canais: caixa de som esquerdo e direito. Já em um *home theater* 5.1 surround temos seis canais, sendo um frontal esquerdo, um frontal direito, traseiro direito, traseiro esquerdo e o *subwoofer*.

Os sistemas de áudio imersivos atuais podem chegar até 11.1.10. Dessa forma, cada canal de transmissão é associado tradicionalmente com uma localização alvo fixa precisamente definida das caixas de som com relação ao ouvinte. Uma experiência de áudio imersiva é usualmente criada melhorando as configurações tradicionais 5.1 ou 7.1 com caixas de som posicionadas no alto. Tipicamente, quatro caixas de som "altas" são adicionadas numa camada superior à camada do meio (isto é, no plano horizontal do ouvinte) num ambiente doméstico.

Ambisonics (ou áudio 3D)

Entrega um cenário de áudio de 3600 que é responsivo ao campo visual e que é "agnóstico à caixa de som". Quando você move a sua cabeça numa direção ou outra, o áudio muda para refletir o movimento. É uma característica desejável que agora é parte de como o áudio imersivo é entregue em quase todo o padrão ou formato. O áudio é codificado ou armazenado de forma a que ele possa ser reproduzido ou adaptado para qualquer número de caixas de som em qualquer arranjo particular. Um dos grandes avanços é que as opções necessárias de entrega são substancialmente reduzidas, uma vez que pode ser decodificado por qualquer sistema. Utilizado principalmente para vídeos em 360º para o YouTube e Facebook, e em aplicações de realidade aumentada e de realidade virtual.

Sistemas utilizados em Broadcast - MPEG-H

Há um bom tutorial a respeito no seguinte link:
https://www.youtube.com/playlist?list=PL1wHeEmBdcWS2QsG-SzHD2PkLaaTyZ71V
(ref.: https://training.npr.org/2018/11/27/360-audio/)
(ref.:https://www.thebroadcastbridge.com/content/entry/13324/immersive-audio-primer-part-1)
(ref.: https://en.wikipedia.org/wiki/Ambisonics)

Objetos de áudio

Um objeto de áudio consiste de uma fonte de áudio acompanhada por metadados que descrevem a localização espacial de um som específico, através de coordenadas tridimensionais, permitindo a reprodução do som numa dada direção, e adaptando-se a quaisquer condições de reprodução, seja uma TV estéreo, barras de som de alta qualidade, ou um cinema com múltiplas caixas de som. (ref.: *Immersive Audio: A Look Inside the Next Generation of Sound* – **Erminia Fiorino**

A principal diferença entre objetos e canais é que a posição espacial de um objeto de áudio pode variar com o tempo e a informação de posicionamento é conduzida como informação secundária entre outros metadados usados para descrever o objeto totalmente. Os metadados associados habilitam o decoder a processar o objeto para a configuração final do alto-falante no lado do receptor.

Este padrão foi desenvolvido para trazer uma nova experiência ao usuário nas aplicações de broadcast e de streaming. É também frequentemente referenciado como Next Generation Audio (NGA) ou Áudio de Nova Geração em uma tradução livre. Ele introduz novos recursos tais como áudio imersivo e interativo, explorando novos conceitos de áudio baseado em objetos e baseados em cena, complementando os avanços de codificação de vídeo para displays Ultra-HD (UHD) com resolução de 4K ou 8K.

Principais recursos

I - Áudio imersivo

Distinguindo-se do som *sourround* pela expansão da imagem de som na dimensão vertical (isto é, o som pode vir de todas as direções, incluindo acima ou abaixo da cabeça do ouvinte), oferecendo uma experiência mais envolvente e realista.

O padrão MPEG-H é um sistema que pode nativamente entregar o áudio imersivo usando qualquer combinação dos três formatos bem estabelecidos descritos acima, ou seja, canais, ambisonics e objetos de áudio. O caso mais comum em aplicações *broadcast* é usar uma mistura de uma "cama" de canais imersiva fixa (isto é, 7.1+4H) e muitos objetos de áudio adicionais (muitas línguas e serviços de descrição de vídeo ou efeitos espaciais aéreos para conteúdo cinematográfico).

II – Personalização e interatividade

Através dos metadados, o MPEG-H permite aos telespectadores manipularem os objetos de áudio, atenuando ou aumentando seus níveis, desabilitando-os, ou mudando suas posições no espaço tridimensional.

O metade do está sob total controle do radiodifusor ou do criador do conteúdo e contém toda a informação necessária para habilitar ou desabilitar recursos específicos, bem como os limites nos quais o usuário pode interagir com o conteúdo.

A Figura 1

A mostra um exemplo de uma interface simples ao usuário, que habilita aos consumidores em casa selecionar entre diferentes presets representando múltiplas versões do conteúdo. Por exemplo, neste caso o usuário poderia escolher entre uma versão do conteúdo sem o comentarista ("*Venue*") de forma a experimentar a sensação de estar presente no estádio, ou uma versão com diálogo melhorado, para um melhor entendimento ("*Dialog+*"). Adicionalmente, um ajuste diferente é usado para habilitar o serviço de audiodescrição ("*AD English*")

Fig. 1-A: Exemplo da interface do usuário usando presets

Adicionalmente, o sistema permite uma interface avançada ao usuário, como mostrado na *Figura 1-B*. Baseado nos metadados criados durante a produção, os consumidores podem interagir com o conteúdo ainda mais, ajustando os níveis de cada competidor ou mover os objetos no espaço 3D. Por exemplo, durante eventos esportivos, é possível direcionar o áudio do comentarista para a camada dos falantes de teto, ao invés de tê-los no falante central.

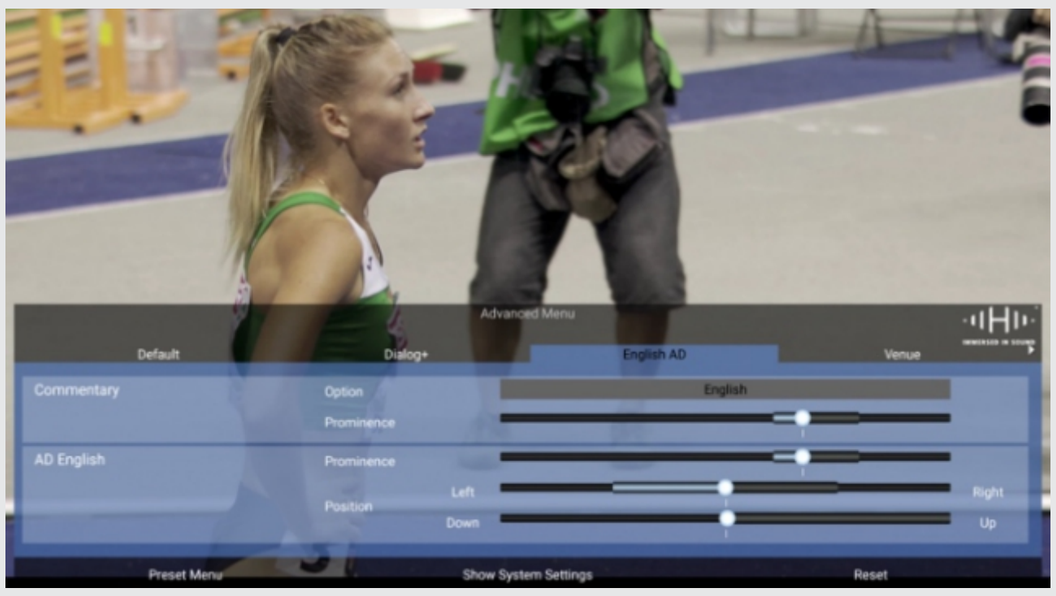


Fig. 1-B: Exemplo da interface de usuáriousando opções avançadas de interatividade.

III – Melhoria do Diálogo

Uma das reclamações mais comuns recebidas pelos *broadcasters* hoje é relacionada à inteligibilidade do diálogo. Engenheiros de som sempre tentam criar uma mixagem como um "compromisso" entre o nível do diálogo e o som de fundo (isto é, música, torcida em um estádio etc.). Existem estudos e experimentos que mostram que mesmo para boas mixagens, os telespectadores podem preferir uma mixagem diferente.

Por exemplo, pessoas com deficiência auditiva podem se beneficiar de um nível do som do diálogo mais alto, ou, os telespectadores podem consumir o conteúdo numa outra língua, que não a sua nativa, de maneira a entender melhor o diálogo. Como o consumo do conteúdo hoje também se dá em vários ambientes, devido ao acesso aos dispositivos móveis, e que, nesses ambientes, o diálogo compete com o ruído ambiente, o recurso de melhoria do diálogo dá à audiência a opção de ajustar o nível à sua própria preferência pessoal, melhorando a experiência de audição.

IV – Acessibilidade e Serviços Multi-idiomas

Com os áudio *codecs* existentes, programas multi-idiomas são transmitidos como mixagem completas em separado. Usando um *stream* para cada mixagem requer um alto *bitrate*, diretamente proporcional ao número de idiomas adicionais oferecidos. Mais do que isso, serviços de audiodescrição têm que ser fornecidos como mixagens completas adicionais, que requerem que a largura de banda aumente ainda mais.

O padrão MPEG-H habilita um uso muito mais eficiente na oferta de serviços de acessibilidade e de multi-idiomas, fazendo uso de objetos de áudio. Por exemplo, como mostrado na *Figura 2*, um programa em 5.1 é entregue em cinco linguagens diferentes num único stream, usando um objeto de áudio para cada idioma. Um sistema legado necessitaria transportar seis mixagens completas em cinco *streams* diferentes.

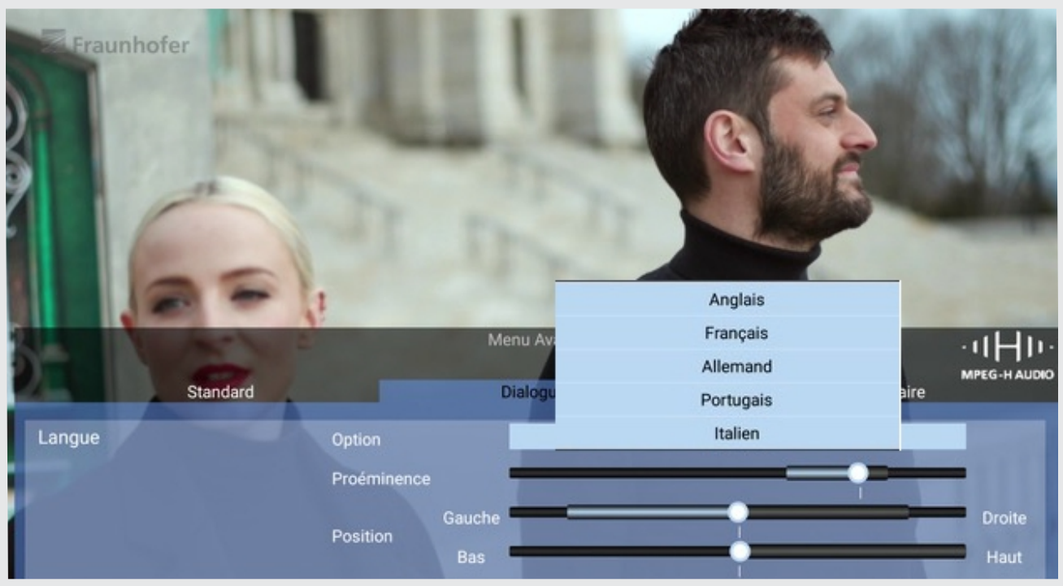


Figura 2: Exemplo de serviço multi-idiomas usando o MPEG-H.

Assumindo os *bitrates* típicos para um som *surround* 5.1 legado, sendo transmitido usando HE-AAC, como mostrado na *Tabela 1,* a abordagem utilizada pelo MPEG-H proporciona mais de 50% de economia na taxa de *bits*.

5.1 Multi-channel Surround in 5 Languages			
Bitrate using MPEG-H Audio		Bitrate using HE-AAC	
5.1 Bed	128 kbps	5.1 in Language 1	160 kbps
Language 1	32 kbps	5.1 in Language 2	160 kbps
Language 2	32 kbps	5.1 in Language 3	160 kbps
Language 3	32 kbps	5.1 in Language 4	160 kbps
Language 4	32 kbps	5.1 in Language 5	160 kbps
Language 5	32 kbps		
Total	**288 kbps**	**Total**	**800 kbps**

Tabela 1: Exemplo de comparação de taxa de bits com codecs legados para serviços de multi-idiomas.

Da mesma forma, no caso de uso de múltiplos idiomas, a abordagem baseada em objeto permite aos criadores de conteúdo explorar novas opções na produção. No caso de eventos esportivos, por exemplo, eles podem permitir aos telespectadores selecionar entre comentaristas diferentes e escutar o comentarista de seu time favorito ou escolher ouvir somente a torcida do seu time.

V – Entrega Universal

A forma como a mídia é consumida tem mudado dramaticamente. Enquanto o conteúdo é entregue em muitos canais diferentes (isto é, TV linear, internet, plataformas móveis), as opções de dispositivos de reprodução têm se tornado as mais diversas, variando entre home theaters de alta qualidade até telefones celulares com fones de ouvido baratos. Além disso, o conteúdo não é mais consumido predominantemente em casa, mas em ambientes variados.

Neste contexto, o padrão MPEG-H fornece não somente um *codec* de áudio, mas uma solução de áudio integrada para entregar a melhor experiência de áudio, independentemente do sistema de reprodução. Isto inclui processamento e funcionalidade de *downmixing*, como também *loudness e Dynamic Range Control* (DRC) avançados. O módulo de normalização de loudness assegura loudness consistente entre programas e canais para diferentes ajustes e configurações de reprodução.

Adicionalmente, o MPEG-H inclui um componente de compensação de *loudness*, responsável pelo ajuste do nível de *loudness* após a interação do usuário. Por exemplo, se o usuário aumenta o nível de diálogo, o nível geral de loudness aumenta. Neste caso, o decoder MPEG-H automaticamente diminuirá o nível da mixagem total após a interação do usuário de tal forma que o nível de *loudness* total permaneça constante, como mostrado na Figura 3.

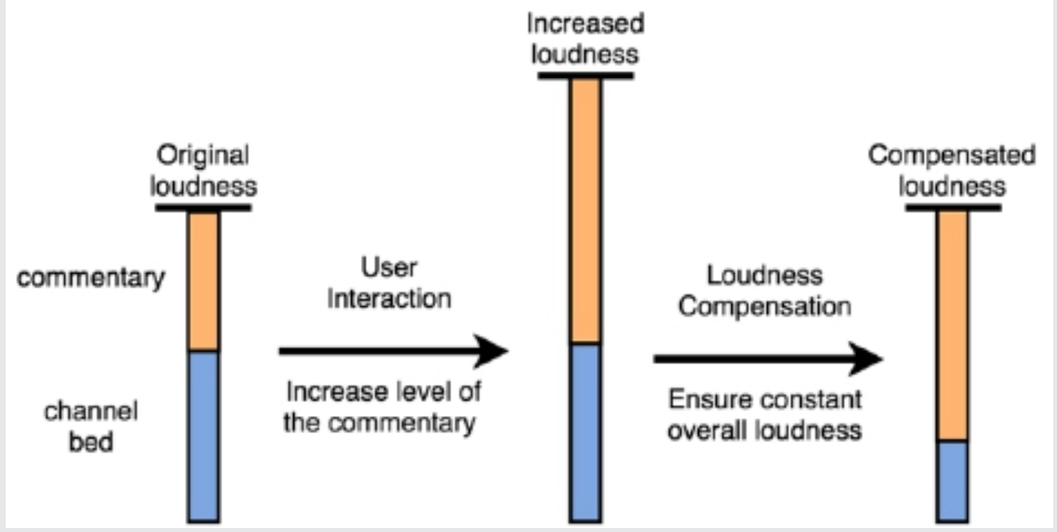


Figura 3: Compensação do loudness após a interação do usuário no MPEG-H.

Estes recursos avançados de controle de *loudness* e DRC, juntos com um amplo conjunto de metadados, permitem aos criadores de conteúdo produzir uma única versão para qualquer plataforma, proporcionando ao usuário obter sempre a melhor experiência sonora, independentemente do dispositivo utilizado. Por exemplo, os telespectadores em casa poderiam experimentar o conteúdo sobre múltiplas plataformas de reprodução, como receptores de TV com estéreo DMX, processador binaural para fones de ouvido, reprodução de *home theater* usando caixas de som ou soundbar imersivos.

VI – Apresentação dos serviços do MPEG-H

O padrão permite incluir metadados descrevendo os ajustes/configurações (“rótulos ou *labels*”) em múltiplas línguas. O produtor do conteúdo pode decidir baseado nas regiões onde o conteúdo é distribuído, criar todos os rótulos em uma ou mais línguas. Baseado no idioma preferido selecionado no receptor, os ajustes serão mostrados ao telespectador. Por exemplo, a *Figura 4* mostra os rótulos criados durante uma transmissão ao vivo em um teste ao vivo em duas línguas: inglês e francês.

Figura 4: Exemplo de rótulos multi-idiomas (inglês na parte superior e francês na inferior)

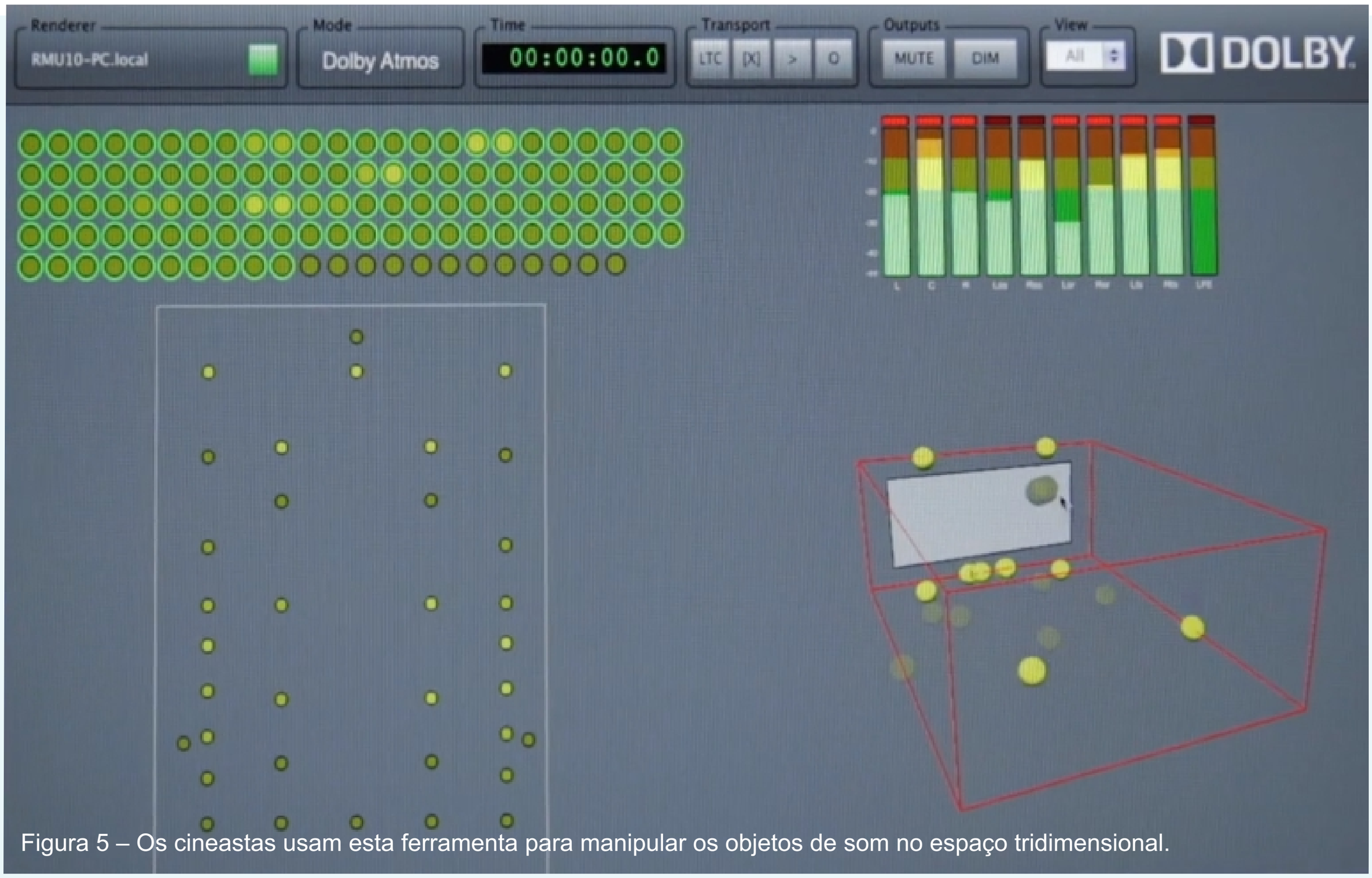


Figura 5 – Os cineastas usam esta ferramenta para manipular os objetos de som no espaço tridimensional.

Usando os metadados do MPEG-H, os criadores de conteúdo ou radiodifusores podem assegurar que suas intenções artísticas e os vários recursos que eles queiram habilitar são corretamente apresentados ao usuário. Desta forma, os broadcasters sempre terão o controle de seus conteúdos e os usuários experimentarão o conteúdo da mesma forma em todos os devices.
(ref.: The Use of MPEG-H Audio in Broadcast - Adrian Murtaza, Stefan Meltzer - Fraunhofer Institute for Integrated Circuits (IIS) - Erlangen, Germany, adrian.murtaza@iis.fraunhofer.de stefan.meltzer@iis.fraunhofer.de)

Um vídeo de apresentação do padrão pode ser acessado em: https://www.youtube.com/watch?v=oErdKT20S6U

Dolby Atmos

Este padrão teve a sua primeira instalação no *Dolby Theater* em Los Angeles, para a estreia do filme *Brave* (lançado com o título de Valente, em português) em junho de 2012.

É uma tecnologia que permite até 128 canais de áudio mais metadados descritores de áudio espacial associado (mais notadamente localização ou automação de dados de *pan*, sendo este último responsável pela distribuição de um sinal de som em um novo estéreo ou campo de som multicanal). Cada trilha de áudio pode ser designada para um canal de áudio ou para um objeto de áudio. Este padrão tem por *default* uma "cama" de 10 canais 7.1.2 para eixos ambientes ou diálogo central, deixando 118 trilhas para objetos de áudio.

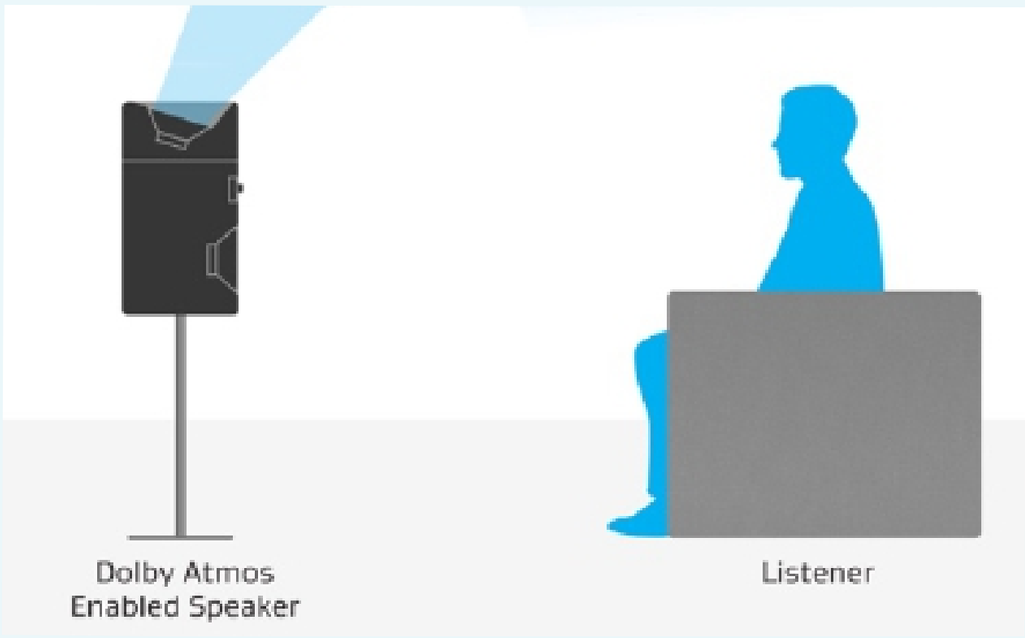
Uma vez implantado nos cinemas, a *Dolby* desenvolveu a tecnologia necessária para transportar a experiência do Dolby Atmos dos cinemas para os home theaters, através do AC4 e o Enhanced AC-3 (Dolby Digital Plus). Neste caso, todo o som da mixagem é representado como um objeto de áudio.

O AC-4 foi projetado para proporcionar experiências de áudio imersivas e inclui diversas ferramentas de codificação eficientes para processar áudio baseado em objetos. Os decodificadores AC-4 possuem um renderizador de objetos sonoros que mapeia o áudio baseado em objetos para saídas via alto-falante, com base em metadados que variam conforme o tempo e definem a posição de cada objeto na cena sonora.

A compatibilidade do áudio baseado em objetos no Enhanced AC-3 é obtida utilizando o algoritmo de codificação JOC (EAC-3 JOC). Um encoder EAC-3 JOC efetua o downmix dos objetos sonoros para uma representação baseada em canais (5.1 ou 7.1 canais) e gera os metadados utilizados por um decodificador de Enhanced AC-3, operando em modo de decodificação JOC, para reconstruir objetos sonoros. Bitstreams de Enhanced AC-3 JOC são compatíveis com decodificadores Enhanced AC-3 que não suportam decodificação baseada em objetos.

A chave para a reprodução do campo sonoro tridimensional é a criação de uma camada de som sobre o ouvinte. As trilhas sonoras em *Dolby Atmos* podem ser apreciadas em *Blu-ray Disc™*, *Ultra HD Blu-ray™*, radiodifusão, e mesmo em fontes de streaming que suportam a reprodução em 4K e Dolby Vision™. Para tanto, foram desenvolvidas as caixas de som para a camada de som superior, que direcionam o som para o teto da sala, onde é refletido, produzindo um som aéreo incrivelmente realista.

Dessa forma, há a possibilidade de ter-se home theater em 7.1 e 5.1, para a reprodução do som imersivo.



O Dolby Atmos permite também a personalização do programa de áudio, onde o consumidor pode escolher entre diferentes apresentações do mesmo. Por exemplo, comentários diferentes sobre um evento esportivo podem ser fornecidos, assim como diferentes sons de fundo, definindo a cena de áudio, como se fosse ouvida em uma determinada área do estádio. Documentários podem ser produzidos com diferentes programas de áudios para consumidores com diferentes níveis de especialização (**Veja reportagem da edição 196 da Revista da SET).**
Link: https://www.flipsnack.com/tmade/revista-da-set-196.html

Exemplos da tecnologia Dolby Atmos podem ser acessados no seguinte link: *https://vimeo.com/search?q=dolby+atmos*

Tanto o MPEG-H quanto o *Dolby Atmos* foram inseridos no ISDB-T através do trabalho do Fórum Brasileiro de TV Digital. Ref.: https://forumsbtvd.org.br/nova-adaptacao-de-audio-imersivo-e-deliberada/
(Nota: O *Dolby Atmos* é parte tanto do *Dolby AC-4* e do *Dolby Enhanced AC-3)*.

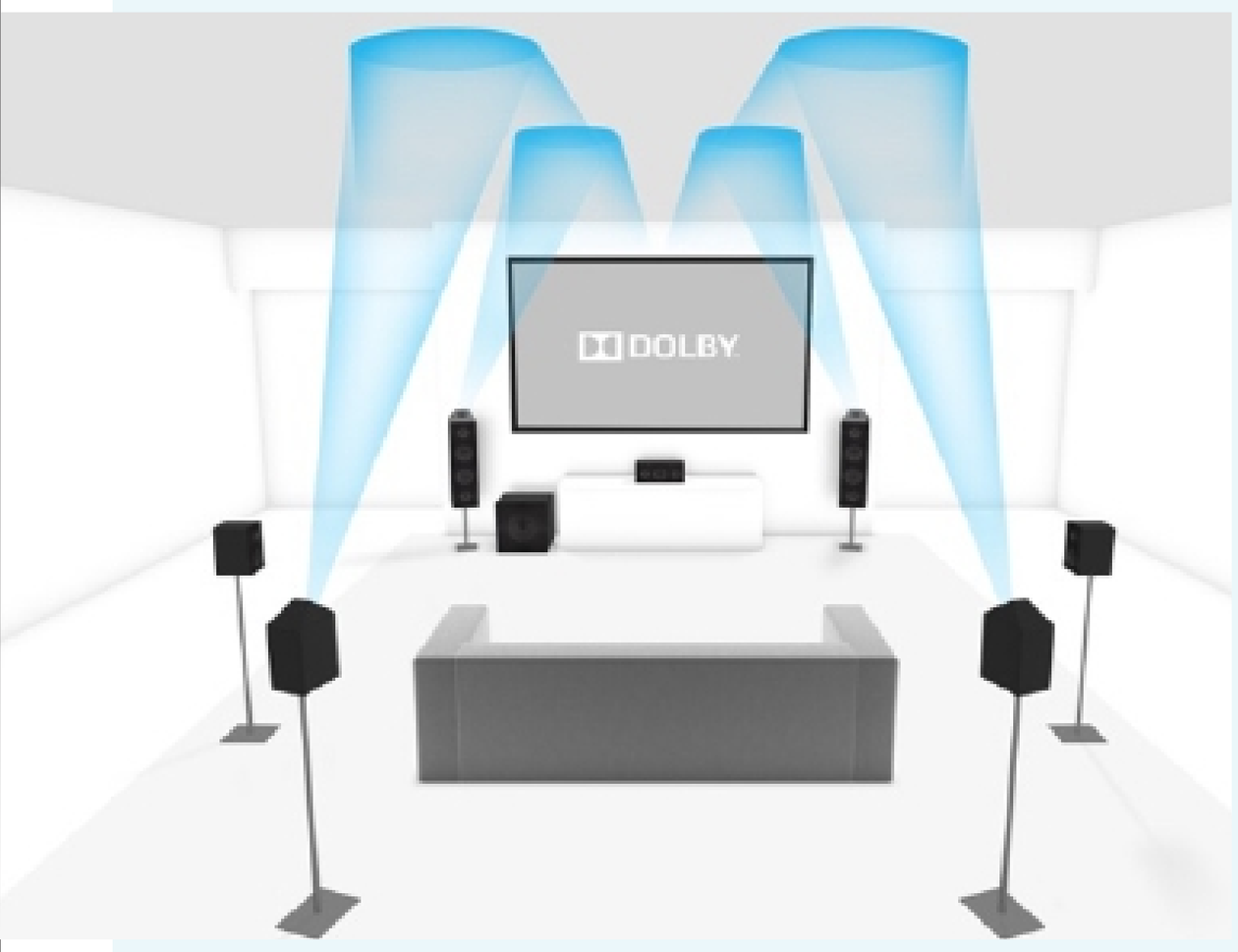


Figura 8: Exemplo de home theater 5.1

Figura 9: Exemplo de soundbar para áudio imersivo

Tem também suporte para transmissão em diversos idiomas no programa de áudio e ainda oferece suporte para a transmissão de programas de áudio suplementares, como por exemplo, um programa de Descrição de Áudio (*Audio Description* - AD) para portadores de deficiência visual ou um comentário do diretor a ser mixado à trilha sonora de um filme.

Há ainda a opção para os dois sistemas (MPEG-H e *Dolby Atmos*) da utilização de um *soundbar* para o áudio imersivo, também com os alto-falantes voltados para o teto.

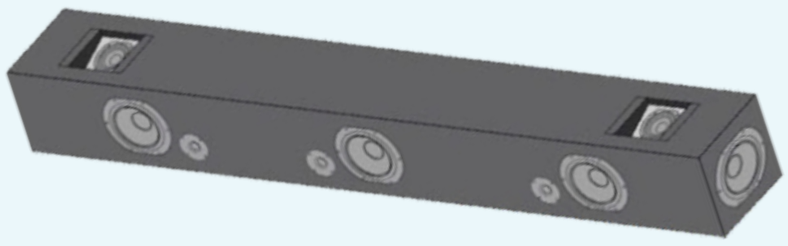


Figura 7: Exemplo de home theater 7.1

Sergio Eduardo Di Santoro Bruzetti é engenheiro graduado em Engenharia Elétrica e Pós-Graduado em Administração Contábil e Financeira, atua no mercado de radiodifusão desde 1977, com passagens por vários cargos nas áreas de engenharia da TV Gazeta de SP, SBT e CNT.

Atualmente na RecordTV, coordenou a implantação dos sistemas de transmissão digitais terrestres de suas principais emissoras e atualmente coordena a transmissão de eventos esportivos internacionais como Olimpíadas de Londres 2012, Jogos de Inverno de Sochi 2014, Jogos Pan-americanos de Toronto 2015 e Olímpiadas do Rio 2016. É membro do Módulo de Mercado do Fórum Brasileiro de TV Digital Terrestre – SBTVD e vice-diretor de TV Aberta da Sociedade de Engenharia de Televisão – SET
Contato: sbruzetti@recordtv.com.br