서정일
한국전자통신연구원(ETRI) 전문위원 / 선임연구원

TECH & TREND

UHDTV를 위한 MPEG-H 3D Audio 표준화 현황

+ **서정일** 한국전자통신연구원(ETRI) 전문위원 / 선임연구원

새로운 방식의 방송서비스를 도입하기 위해서는 프로그램 제작에 필요한 카메라부터 시작하여 전송시스템 및 단말까지 모두 교체하여야 하므로 많은 시간이 소요되는 것이 사실이다. 디지털 방송의 경우에도 관련 기반기술들은 1990년도 초반에 개발이 완료되었으나 본격적인 디지털 방송은 2000년대가 되어서 시작되었다. 그러나 지상파에 한정되었던 전송매체가 케이블, 광대역망, 이동통신망 등으로 다변화되고, 전송시스템을 제외한 비디오 및 오디오 코덱들은 소프트웨어 기반으로 개발되어 손쉽게 변경 또는 추가가 가능하게 됨으로써 새로운 방송시스템을 도입하는 데 대한 저항은 예전에 비해 많이 약화된 것이 사실이다.
3DTV와 UHDTV(Ultra High Definition TeleVision)로 대변되는 차세대 실감방송 서비스는 대화면 고해상도(4K 또는 8K) 비디오와 다채널(9.1채널, 10.2채널, 22.2채널 등) 오디오를 이용하여 기존의 HD(high definition television) 비디오와 5.1채널 오디오가 표현할 수 있는 시공간적인 해상도를 향상시켜 실감(realistic)나고 몰입감(immersive) 있는 방송을 제공하는 것이 목표이다. 본 글에서는 UHDTV로 대표되는 차세대 실감방송 서비스를 제공하는 데 필요한 여러 가지 기술들 가운데 오디오 기술에 대한 표준화 동향을 논하고자 하며, 특히 2013년부터 표준화가 시작된 MPEG-H 3D Audio에 대해 중점적으로 다루고자 한다.

3차원 음향공간

UHDTV에서 오디오의 역할은 대화면 고해상도 실감 비디오와 부합하면서 디스플레이가 표현할 수 없는 공간에서도 오디오를 통해 3차원 음향공간을 표현하는 데 있다. 따라서 기존의 스테레오 및 5.1채널 서라운드 오디오가 표현할 수 있는 2차원 음향공간을 3차원으로 확장하여 시청자를 중심으로 3차원 모든 방향에 둘러싸는 음장감을 제공하여야 하며, 대화면 디스플레이에서 표현되는 비디오 객체와 동일한 공간상의 위치에서 음상이 정확하게 표현되어야 한다. 상기와 같은 목적을 위하여 일본 NHK에서는 [그림 1]과 같은 22.2채널 오디오 시스템을 개발하여 일본의 UHDTV 방송 표준 시스템 SHV(Super Hi-Vision)에 적용하였다. NHK 22.2채널 오디오 시스템은 상층, 중층 및 하층의 3개 계층으로 22개의 스피커들을 3차원 공간상에 배치한다. 특히 대화면 디스플레이가 위치하는 전방에는 스피커들을 촘촘히 배치하여(총 11채널) 대화면 UHD 영상과 공간상의 동일한 위치에 음원들이 재현되도록 하였다. 또한 2개의 서브우퍼 채널을 이용하여 좌우가 분리된 풍부한 저음역대를 재생할 수 있다.

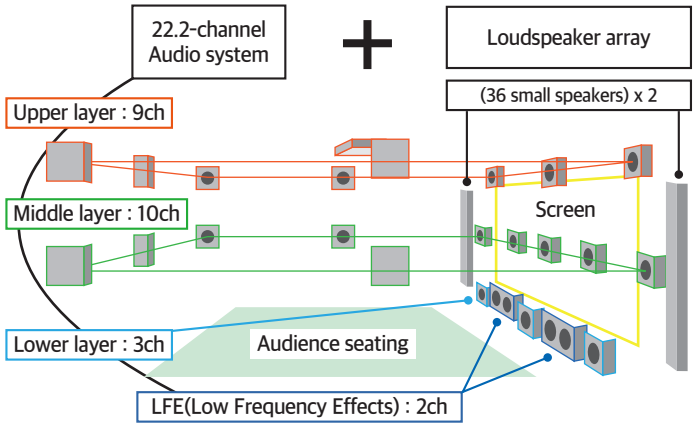

그림 1. NHK 22.2채널 오디오 시스템

NHK뿐만 아니라 Dolby, DTS, AURO 3D, THX, ETRI/삼성전자 등에서도 청취자를 둘러싸는 3차원 스피커 포맷을 개발하여 디지털시네마 서비스에 적용하고 있다. [그림 2]는 AURO 3D에서 제안한 9.1채널 스피커 포맷이며, 기존의 5.1채널 스피커 배치에 센터채널을 제외한 4개의 채널 위쪽에 스피커들을 추가하여 배치함으로써 음의 고저감을 표현한다.

그림 2. AURO 3D 9.1채널 포맷

또한 UHDTV가 제공하는 3차원 오디오를 구성하는 음원(sound source) 중 대사(dialogue)나 이동음원과 같이 청취자에게 중요한 정보를 전달하거나 3차원 음장감을 제공하는 데 주요하게 작용하는 음원들은 다양한 수신환경에서 가능한 한 정확하고 또렷하게 전달되어야 한다. 오디오 객체를 활용하는 가장 대표적인 예인 Dolby ATMOSTM는 영화음향을 방향성 없이 분위기만을 전달하는 9.1채널 배경음(Beds)과 대사와 같이 청감상 주요한 정보를 가지고 있는 오디오 객체들로 분리하여 제작하고, 극장마다 상이한 스피커 설치환경에서도 사운드 엔지니어가 의도한 위치에 객체들을 표현할 수 있게 하는 영화음향 시스템이다[그림 3 참조]. 따라서 별도의 분리된 오디오 객체(audio object)를 수신기에서 제어할 수 있게 함으로써 청취자가 갖춘 음향재생 환경에 최적으로 3차원 오디오를 표현할 수 있으며, 다양한 대화형 서비스도 제공할 수 있다.

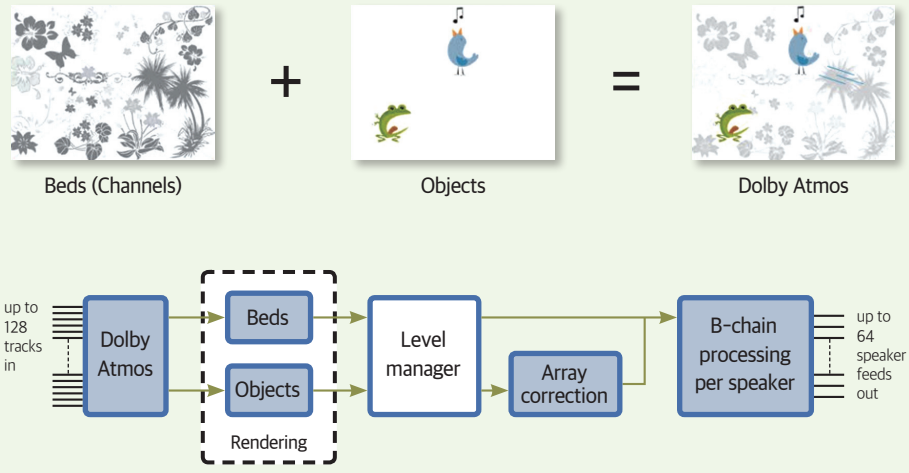

그림 3. Dolby ATMOS에서의 Beds와 객체의 개념도(위)와 재생과정 블록도(아래) / 출처. www.dolby.com

서정일
한국전자통신연구원(ETRI) 전문위원 / 선임연구원

또한, 미래의 방송에서 22.2채널 오디오를 전송하고 수신기에서 출력할 수 있다 하더라도 모든 가정에서 22.2채널 스피커를 갖추어 청취하는 것을 기대하기는 어렵다. 따라서 전송된 다채널 오디오 신호를 시청자가 갖추고 있는 스피커 배치환경에 최적으로 변환하여 재생하는 기술도 필요하다.

지금까지 설명한 UHDTV와 같은 실감방송을 위하여 새롭게 제시되는 오디오 서비스에 대한 요구사항은 아래와 같이 3가지 정도로 요약될 수 있다.

- **고차 다채널 오디오(High-order Multichannel Audio)** : NHK 22.2채널과 같이 상층 채널(Height Channel)을 포함하여 스피커들이 3차원으로 배치되는 다채널 오디오 재생 시스템을 지원하여야 한다.
- **객체기반 오디오(Object-based Audio)** : 다채널 오디오 배경음과 분리되어 있는 오디오 객체 음원들을 전송하고 단말에서 제어할 수 있어야 한다.
- **자유로운 다운믹싱 및 재현(Flexible Downmixing and Rendering)** : 입력된 다채널 오디오를 단말에 구비된 스피커 배치환경에 최적으로 다운믹싱 하거나 채널 포맷을 변환하여 재생할 수 있어야 한다.

오디오 기술 표준화

이러한 요구사항을 충족시키기 위하여 MPEG(Moving Picture Experts Group), ITU-R(International Telecommunication Union Radiocommunication sector), ATSC(Advanced Television System Committee), DVB(Digital Video Broadcasting), SMPTE(Society of Motion Picture and Television Engineers) 등과 같은 국제 표준화 단체에서는 차세대 실감방송을 위한 오디오 기술에 대한 표준화를 진행 중이며, 국내에서도 TTA(Telecommunications Technology Association)를 통하여 UHDTV를 위한 오디오 표준화를 진행하고 있다.

주로 비디오 및 오디오 신호에 대한 압축 부호화 표준화를 담당하는 MPEG에서는 1997년 AAC(Advanced Audio Coding) 오디오 부호화 표준을 제정한 이후 오디오 부호화 코어(core)에 대한 표준화를 진행하지 않았으나, 2008년부터 높은 비트율에서 AAC 보다 우수한 음질을 나타내고, 낮은 비트율에서는 오디오 신호뿐만 아니라 음성신호에 대해서도 우수한 음질을 나타내는 USAC(Unified Speech & Audio Coding) 기술을 표준화하였다. USAC는 MPEG에서 최근까지 개발된 다양한 오디오 부호화 툴들과 음성신호에 대하여 강점을 가지는 CELP(Code Excitation Linear Prediction) 계열의 AMR-WB+(Extended Adaptive Multi-Rate - Wideband) 기술을 효과적으로 접목하여 다양한 비트율에서 세계 최고 수준의 부호화 성능을 발휘한다. 표준제정 초기에는 USAC가 이동방송, 오디오북, 음악 플레이어 등과 같은 스테레오 오디오 신호를 이용하는 서비스로 응용분야를 한정하였기 때문에 스테레오 오디오 신호에 대한 압축방법만 표준에 명시되어 있지만, 인코더와 디코더를 스테레오 쌍(stereo channel pair)으로 확장하면 UHDTV와 같은 다채널 오디오 서비스에 대해서도 적용이 가능하다.

UHDTV를 위한 22.2채널 오디오 시스템을 개발한 NHK는 자사의 다채널 3차원 오디오 포맷을 MPEG 표준에 채택시키기 위하여 다채널 오디오 신호에 대한 새로운 부호화 및 재생방식에 대한 표준화를 MPEG에서 수행해야 한다고 2010년부터 주장하였으며, 비슷한 시기에 진행된 고품질 비디오 부호화 표준인 HEVC(High Efficiency Video Coding)에 버금가는 오디오 품질을 제공할 수 있는 실감 오디오 기술에 대한 표준화 필요성이 부각되었다. 이러한 배경으로 MPEG에서는 HEVC와 동일한 표준 프로

젝트(MPEG-H) 내에서 3D Audio란 이름으로 다채널 고음질 오디오 부호화 및 재생방법에 대한 표준화를 진행하기로 결정하고 2013년 1월에 기술을 제안받기 위한 CfP(Call for Proposal)를 발표하였다.

MPEG-H 3D Audio는 전통적인 채널 기반의 오디오 신호뿐만 아니라 오디오 객체 신호와 3차원 음장을 직접 녹음하고 표현하는 고차 앰비소닉스(HOA: High Order Ambisonics) 신호도 처리하는 것을 목표로 하였다. 또한 부호화 기술뿐만 아니라 단말의 재생환경에 맞게 디코더 출력 신호를 변환하는 다운믹스(downmix) 기능 및 자유 렌더링(flexible rendering)1) 기능도 표준화의 필수항목에 포함하였다. 다만 고차 앰비소닉스 신호는 공간 하모닉스(spatial harmonics)를 해석하고 합성하는 별도의 처리 과정이 존재하기 때문에 독립적으로 표준화를 진행하기로 결정하였다. 이러한 배경으로 2013년 5월 말 ETRI, Fraunhofer IIS, Fraunhofer IDMT, Sony는 채널과 객체신호 분야에 부호화 및 재생기술을 제안하였고, Orange Labs, Technicolor, Qualcomm은 고차 앰비소닉스 분야에 자사의 기술을 제안하였다. 2개월 동안 청취평가를 수행한 후 2013년 7월에 열린 제105차 MPEG 회의에서 채널과 객체신호 분야에서는 Fraunhofer IIS 기술이, 고차 앰비소닉스 분야에서는 Orange Labs와 Technicolor에서 공동 제안한 기술이 참조모델(Reference Model)2)로 선정되었다.

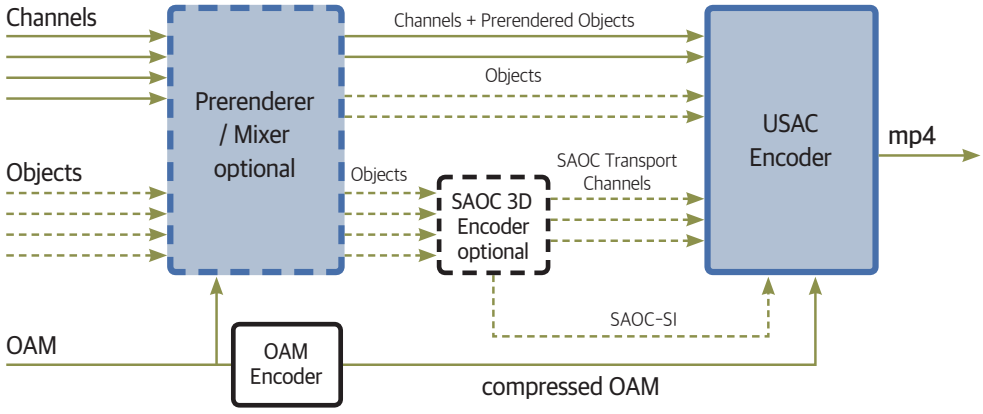

그림 4. MPEG-H 3D Audio 채널과 객체분야 참조모델 부호화기 블록도
(출처: N14060, Working Draft Text of MPEG-H 3D Audio CO RM0, Geneva, Swiss, November 2013)

[그림 4]는 채널과 객체신호 분야에서 선정된 참조모델의 부호화기 블록도이다. [그림 4]에서 확인할 수 있듯이 채널과 객체분야 참조모델은 USAC와 SAOC(Spatial Audio Object Coding)를 핵심 부호화툴로 이용하며, 오디오 객체신호는 채널신호에 미리 믹싱하는 방법과 모노채널 신호로 가정하여 채널신호와 같이 부호화하는 방법, SAOC를 이용하여 고효율로 압축하는 방법 중 선택하여 처리한다. 디코더에서 오디오 객체신호를 출력 채널 신호로 합성하기 위해서는 객체신호의 3차원 공간상의 재생위치를 알아야 하므로, MPEG-H 3D Audio 인코더에서는 오디오 객체신호의 재생위치를 OAM(Object Audio Metadata)이라는 메타데이터 포맷으로 정의하고, 인코딩 과정 중에서 MPEG-H 3D Audio 비트스트림 내에 다중화한다.

1) 다운믹스는 입력되는 다채널 오디오 신호의 채널수를 줄이는 과정(예: 22.2 채널 신호를 5.1채널로)을 지칭한다. 그러나 채널수가 줄어드는 출력환경(예: 5.1채널) 사전에 정의되어 있다. 따라서 다운믹싱되는 채널의 수와 각 채널들에 대한 스피커 위치가 고정되어 있다. 그러나 자유 렌더링은 단말에 구비되어 있는 임의의 스피커 배치환경에 최적으로 입력 다채널 오디오 신호를 변환하는 과정을 의미하며, 채널수뿐만 아니라 각 채널들에 대한 스피커 위치도 가변될 수 있다.

2) MPEG 오디오 서브그룹은 새로운 표준을 개발하기 위해 발표된 Cfp에 응답하여 제출된 기술들을 평가한 후 가장 우수한 성능을 나타내는 기술을 참조모델로 선정한다. 이후 2~3년의 기간 동안 참조모델의 성능을 향상시킬 수 있는 요소기술들을 추가하여 표준화를 완료한다.

서정일
한국전자통신연구원(ETRI) 전문위원 / 선임연구원

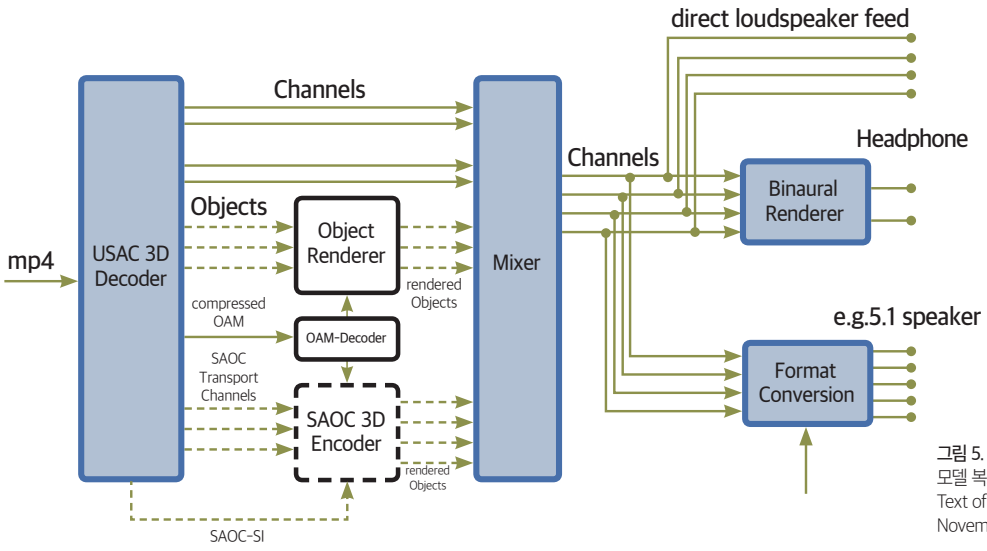

그림 5. MPEG-H 3D Audio 채널과 객체분야 참조 모델 복호화기 블록도(출처: N14060, Working Draft Text of MPEG-H 3D Audio CO RM0, Geneva, Swiss, November 2013)

[그림 5]는 채널과 객체신호 분야에서 선정된 참조모델 복호화기 블록도이다. 부호화기의 역순으로 복호화 과정이 이루어져서 입력 오디오 신호와 동일한 스피커 포맷의 신호가 믹서(Mixer)에서 출력된 후 재생환경에 따라 세 가지 모드로 동작하게 된다. 먼저 입력포맷과 동일한 스피커 출력 포맷을 구비하고 있을 경우에는 직접 스피커로 출력한다. 헤드폰을 이용한 재생환경에서는 믹서에서 출력되는 다채널 오디오 신호를 바이노럴 렌더링(binaural rendering) 과정을 거쳐 3차원 헤드폰 신호로 변환한 후 출력된다. 이때 바이노럴 프로세싱은 BRIR(Binaural Room Impulse Response)이란 특수한 형태의 충격응답(impulse response) 신호를 이용하는데, BRIR은 잘 설계된 청취룸(listening room)에서 다채널 스피커 포맷(예 22.2채널) 스피커를 직접 이용하여 녹음된 바이노럴 충격응답신호이며, BRIR을 이용하면 무향실에서 녹음된 HRTF(Head Related Transfer Function)를 이용하는 방법보다 자연스러운 3차원 바이노럴 오디오 신호를 합성할 수 있다. 마지막으로 입력 오디오 포맷과 상이한 스피커 출력 포맷을 구비하고 있을 경우에는 포맷 변환기(Format Converter)를 통하여 재생 스피커 출력 포맷에서 맞게 입력 스피커 신호를 변환하여 출력한다. 포맷 변환기는 믹서로부터 입력되는 다채널 오디오 입력신호를 개별 음원으로 가정하고 출력 스피커 포맷에 재합성하는 과정을 통하여 출력 스피커 포맷에 최적으로 적응된 오디오 신호를 출력한다.

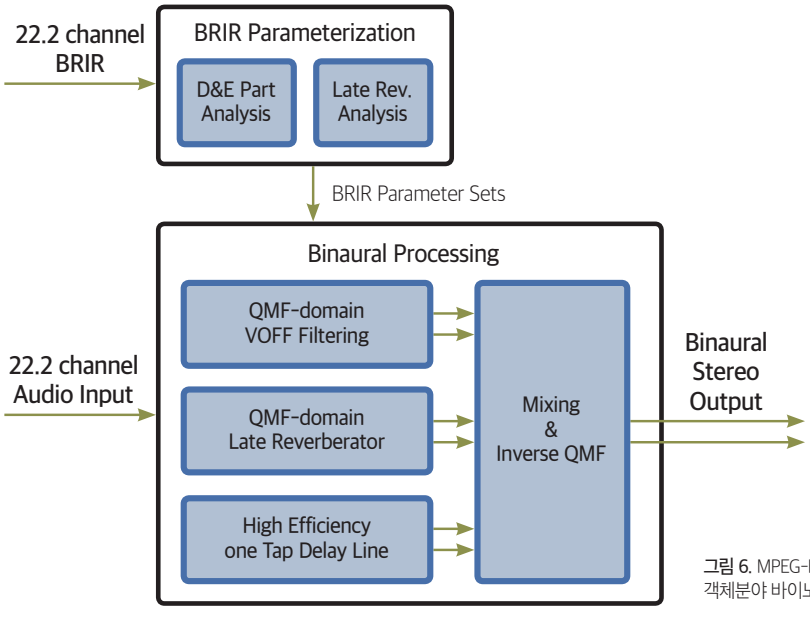

그림 6. MPEG-H 3D Audio 채널과 객체분야 바이노럴 렌더링 블록도

[그림 4와 5]에 표현된 USAC 3D는 기존의 USAC을 다채널 지원이 가능하도록 스테레오 부호화 블록을 확장한 것 외에는 큰 변화가 없으며, SAOC 3D는 다객체 오디오 부호화 기술인 SAOC 표준에서 고음질 오디오 객체를 지원하는 EAO(Enhanced Audio Object) 기능을 포함한 일부 기능만으로 간략하게 구성하여 활용하였다. 즉 3D Audio 채널과 객체분야 참조모델을 부호화 기술 관점에서 살펴보면 1990년 후반부터 개발된 MPEG 오디오 부호화 기술들인 AAC, SBR(Spectral Band Replication), MPS(MPEG Surround), SAOC를 결합한 형태라고 할 수 있다.

바이노럴 렌더링에 대해서는 2013년 10월부터 2014년 1월까지 ETRI, 연세대, 윌러스, Fraunhofer, Orange Labs, Qualcomm, Huawei 등이 참여하여 추가적인 기술선정 과정을 수행하였으며, 2014년 1월 ETRI, 연세대, 윌러스, Fraunhofer가 공동으로 제안한 QMF(Quadrature Mirror Filter) 기반 바이노럴 렌더링 기술이 참조모델로 채택되었다. [그림 6]은 참조모델로 채택된 바이노럴 렌더링에 대한 블록도이다. 바이노럴 렌더러는 [그림 6]과 같이 BRIR 신호를 분석하여 BRIR parameter sets를 추출하는 BRIR parameterization 블록과 입력 다채널 오디오 신호에 상기 파라미터들을 이용하여 바이노럴 스테레오 신호를 출력하는 BRIR Processing 블록으로 구성된다. BRIR parameterization 블록은 QMF 영역에서 BRIR의 직접음과 초기 반사음(Direct & Early Reflection)을 모델링하는(D&E Part Analysis) 과정과 BRIR의 잔향성분을 분석(Late Reverberation Analysis)하는 과정을 통하여 BRIR parameter sets를 추출한다. BRIR Processing 블록은 BRIR 신호의 주파수 대역별 특성을 반영하기 위하여 QMF 대역별로 상이한 길이를 가지는 BRIR 필터 구조로 구성되는 VOFF(Variable Order Frequency Filtering) 기술을 이용하여 입력 다채널 오디오와 BRIR을 필터링(QMF-domain VOFF Filtering)하며, 저주파수 대역에 비하여 상대적으로 둔감한 고주파수 대역은 1개의 지연과 이득만을 가지는 필터로 처리하는 High Efficiency One Tap Delay Line 기술을 이용하여 직접음과 초기 반사음만을 재생한다. 또한 QMF 영역에서 BRIR의 잔향성분을 적용(QMF-domain Late Reverberator)하는 과정을 추가로 수행한다.

차세대 실감방송에 대한 국가단위 표준화가 한국, 일본, 미국, 유럽 등지에서 2015년을 전후하여 이루어질 예정이므로 MPEG은 3D Audio에 대한 표준화를 신속하게 진행하여 2015년 2월까지 1차 버전을 완성할 예정이다. 제한된 표준화 일정 때문에 2013년 7월 결정된 참조모델의 큰 변화는 없을 것으로 예상되며, 수신기에서 음량(volume)을 능동적으로 제어하는 기술인 DRC(Dynamic Range Control) 기술만 추가될 것으로 예상된다. 또한 22.2채널을 128kbps 이하로 저비트율로 압축할 수 있는 MPEG-H 3D Audio 2차 버전은 2014년 하반기부터 시작하여 2015년 말까지 표준화될 예정이다.

본 글을 통하여 UHDTV와 같은 차세대 실감방송을 위하여 MPEG에서 표준화가 진행 중인 MPEG-H 3D Audio의 배경 및 진행과정과 채택기술들에 대해 살펴보았다. 아직도 대부분의 방송 프로그램은 스테레오 포맷으로 제작되고 있으며, 5.1채널 방송 프로그램을 위한 제작환경이나 전문 음향엔지니어가 부족한 실정이다. 이러한 상황에서 5.1채널보다 훨씬 많은 채널들로 구성되는 22.2채널과 같은 고차 다채널 오디오 프로그램을 제작할 수 있는 환경이 보급되기까지는 많은 시간이 소요될 것으로 예상된다. 또한 일반 가정에 이렇게 많은 수의 스피커들을 설치하도록 유도하는 것도 쉬운 일이 아닐 것이다. 이러한 현실에서 과연 이렇게 많은 채널들로 구성되는 오디오 서비스가 보급되기 어렵다고 생각할 수 있지만, 전통적으로 새로운 오디오 서비스는 영화로부터 시작된 역사를 되짚어 보면 비관적이지만은 않다. Dolby ATMOS, AURO 3D, 소닉티어 14.2채널/30.2채널과 같은 다채널 영화음향 기술들이 주목받고 있으며, 시장에서도 긍정적인 반응을 얻고 있다. 영화에서부터 이러한 고차 다채널 오디오로 구성된 콘텐츠들이 제작되기 시작하면 방송 등과 같은 다양한 분야에서도 활용될 것이다. 따라서 MPEG-H 3D Audio 표준이 실제 방송에 활용되는 시점이 우리가 생각하는 것보다 훨씬 빨라질 수도 있다.