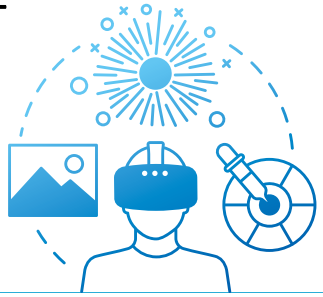
몰입형 미디어 서비스를 위한 오디오 표준

- MPEG-H 3D 오디오와 MPEG-I Immersive 오디오 -



이미숙 _ 한국전자통신연구원 책임연구원
이용주 _ 한국전자통신연구원 책임연구원
이태진 _ 한국전자통신연구원 실장

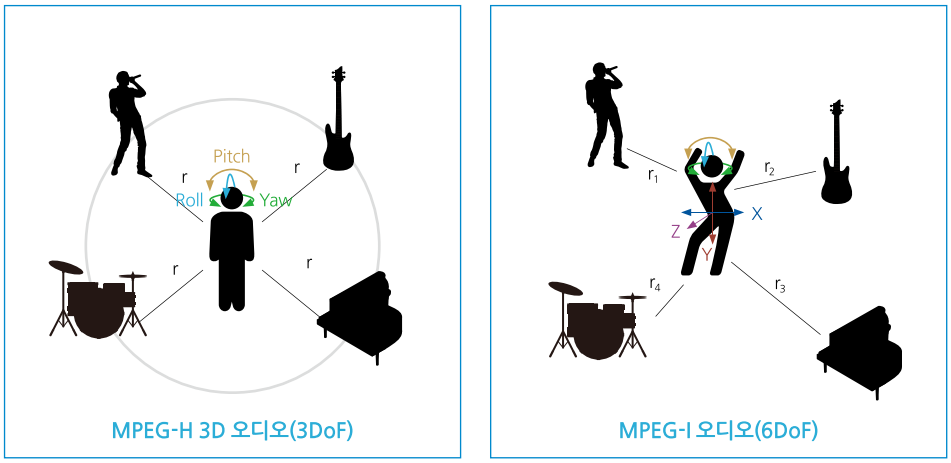
1. 머리말

MPEG에서는 현실 세계에서의 사용자 경험을 가상공간에서 재현하는 데 필요한 기술의 표준화를 위해 2016년 10월에 개최된 MPEG 116차 회의에서 몰입형 미디어(Immersive media)에 대한 MPEG-I(ISO/IEC 23090: Coded Representation of Immersive Media) 프로젝트를 공식적으로 시작하였다. 표준화 초기에는 몰입형 서비스에 필요한 기술을 8개 파트로 나누어 논의하였으나 현재는 몰입형 미디어 아키텍처, 비디오 및 오디오 등으로 구성된 12개 파트에서 표준화를 진행하고 있다. 오디오 기술은 'MPEG-I 파트 4 몰입형 오디오 코딩(MPEG-I Part 4: Immersive Audio Coding)'이라는 제목으로 표준화가 진행 중이다.

또한 MPEG에서는 빠르게 발전하는 미디어 시장에 대응하기 위해 자유도(DoF, Degree of Freedom)의 정도를 두 단계로 나누어 표준화를 진행하고 있다. 단계 1a에서는 사용자의 위치가 고정된 상태에서 x, y, z 축을 중심으로 머리를 회전(Pitch, Yaw, Roll)할 수 있는 3자유도(3DoF), 단계 1b에서는 3DoF와 같은 머리 회전과 제한된 범위 내에서 이동이 허용된 3DoF+, 그리고 단계 2에서는 머리 회전과 함께 전후, 좌우, 상하 방향으로 이동이 가능한 6자유도(6DoF)를 지원하는 기술에 대한 표준화를 진행하기로 하였다.

오디오 서브그룹에서는 단계 1의 3DoF와 3DoF+까지는 이미 표준이 완료된 MPEG-H 3D 오디오 저-복잡도 프로파일(Low-Complexity Profile)로 충분히 서비스 가능하다고 판단하고, MPEG-I 오디오에서는 단계 2의 6DoF를 지원하는 오디오 기술에 대한 표준화를 진행하기로 하였다. 이러한 논의 결과를 반영하여 전방향 미디어 포맷을 다루는 MPEG-I 파트 2인 OMAF(Omnidirectional MediA Format)에서는 단계 1을 지원하는 오디오 기술로 MPEG-H 3D 오디오 저-복잡도 프로파일을 정의하고 있다.

MPEG-H 3D 오디오[1]는 가상현실/증강현실(VR/AR) 서비스에 필요한 채널, 객체, 그리고 앰비소닉스(HOA, Higher Order Ambisonics) 오디오를 효율적으로 압축 및 복원할 수 있는 기술이다. 또한



[그림 1] MPEG-H 3D 오디오와 MPEG-I 오디오 기술을 이용한 서비스 환경 예

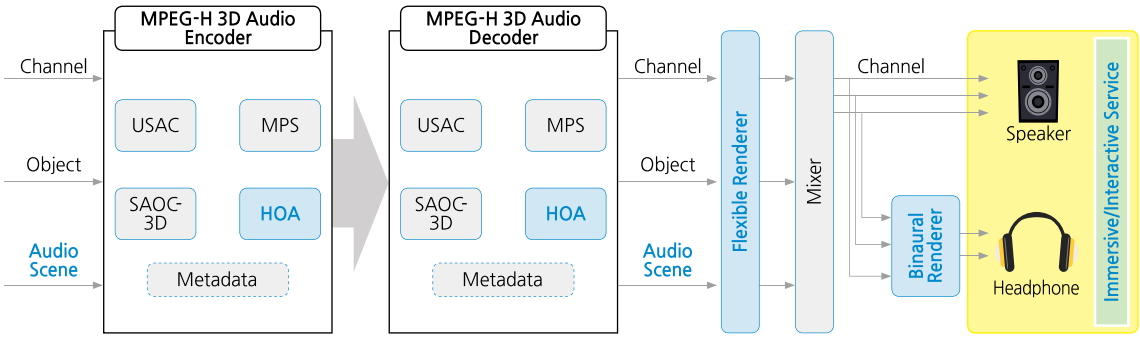
헤드폰에서 최적화된 3D 오디오를 재현하기 위한 여러 가지 렌더링(Rendering) 기술을 포함하고 있다. 따라서 오디오 서브그룹에서는 MPEG-H 3D 오디오 기술을 3DoF와 3DoF+ 서비스에 사용할 뿐만 아니라 MPEG-I 오디오에서 채널, 객체 그리고 앰비소닉스 오디오의 압축 및 복원에도 활용하기로 결정하였다. 그리고 MPEG-I 오디오 표준화에서는 6DoF를 지원하기 위해 추가적으로 필요한 메타데이터와 렌더링 기술 개발에 집중하기로 하였다.

[그림 1]은 MPEG-H 3D 오디오와 MPEG-I 오디오 기술을 통해 제공할 수 있는 오디오 서비스 환경에 대한 간단한 예를 보여주고 있다. 그림을 보면 MPEG-H 3D 오디오 기술을 사용할 경우, 사용자는 고정된 자리에서 머리만 좌우, 전후, 상하로 움직일 수 있는 상태에서 오디오 콘텐츠를 청취할 수 있다. 반면에 MPEG-I 오디오 기술을 사용하면 사용자가 머리를 움직일 수 있을 뿐만 아니라 위치도 이동하면서 오디오 콘텐츠를 청취할 수 있게 된다. 본고에서는 MPEG-H 3D 오디오 기술의 특징과 MPEG-I 오디오 기술의 표준화 진행 상황에 대해 살펴본다.

2. MPEG-H 3D 오디오

2015년에 표준으로 제정된 MPEG-H 3D 오디오는 대화면 고해상도로 대변되는 UHDTV 방송 서비스를 고려하여 개인 맞춤형 오디오 서비스를 제공하기 위해 채널, 객체, 그리고 앰비소닉스 오디오를 효율적으로 압축 및 복원할 수 있는 기술이다. MPEG에서는 다양한 형태의 입력신호를 처리하기 위해 [그림 2]와 같이 기존의 MPEG 오디오 표준 기술인 USAC(Unified Speech & Audio Coding)[2], MPS(MPEG-Surround)[3] 그리고 SAOC(Spatial Audio Object Coding)[4] 기술 등을 활용하고, 앰비소닉스 오디오와 다양한 재현 환경에 최적화된 오디오를 재생하기 위한 렌더링 기술을 추가하여 MPEG-H 3D 오디오 표준을 제정하였다.

MPEG-H 3D 오디오는 세 가지 프로파일(메인 프로파일, 저-복잡도 프로파일, 상위 프로파일)을 제공하고 있다. 이 중에서 저-복잡도 프로파일은 복잡도가 낮은 기술 또는 메인 프로파일에 있는 툴의 복잡도를 낮춘 기술들을 기반으로 만들어진 프로파일로 품질과 기능은 메인 프로파일과 크게 다르지 않다.



[그림 2] MPEG-H 3D 오디오 기술 개요[5]

2.1 핵심 기술

일반적으로 오디오 콘텐츠는 채널 오디오 형태로 전송되어 정해진 위치에 있는 스피커를 통해 재생된다. 국내 HDTV 방송에서는 오디오 신호를 대부분 스테레오 채널로 전송하고 일부 음악 방송에서만 5.1 채널을 사용하고 있다. 그러나 NHK에서는 22.2 채널 오디오 시스템을 개발하여 일본 UHDTV 방송 표준에 적용하였다. 이렇게 고차의 다채널에 대한 요구가 생기면서 MPEG-H 3D 오디오에서는 22.2 채널까지 지원하고 있다.

고차 다채널을 지원하기 위해서는 특히 압축 성능이 중요하므로 MPEG에서는 음성과 음악 모두에서 고른 성능을 나타낼 뿐만 아니라 압축효율이 뛰어난 USAC[2]을 MPEG-H 3D 오디오의 핵심 기술로 채택하였다. USAC에서는 스테레오 신호의 코딩을 위해 성능이 개선된 MPS(MPEG Surround) 기술[3]을 사용하고 있는데, MPEG-H 3D 오디오에서는 이 기술을 확장하여 고차의 다채널 오디오 신호를 처리하는 데 사용하고 있다.

객체 오디오는 오디오 장면을 구성하는 각각의 음원을 의미한다. 예를 들어, 스포츠 중계에서 관중석의 응원 소리와 아나운서의 해설을 별도의 객체 오디오로 볼 수 있다. 기존 채널 오디오에서는 관중석의 소리와 아나운서의 목소리가 혼합된 신호를 재생단의 스피커에 1:1로 매칭되는 채널이라는 단위로 묶어서 처리한다. 그러나 객체 단위로 신호를 압축하여 전송하면 시청자의 스피커 구성환경에 맞게 신호를 재현할 수 있다. 객체 오디오는 채널 오디오와는 달리 객체 오디오가 재생되는 스피커 위치를 제어할 수 있으며, 사용자의 선택에 따라 인터랙티브(Interactive) 서비스도 가능하다. 이러한 객체 오디오는 채널 오디오로 간주하여 USAC으로 압축하거나 SAOC-3D[6]로 압축할 수 있다. 특히, 객체 오디오의 수가 많거나 전송 비트율이 낮은 경우에는 압축효율이 좋은 SAOC-3D 기술을 적용한다.

MPEG-H 3D 오디오 표준은 또한 앰비소닉스 오디오[7]를 효율적으로 압축할 수 있는 기술을 포함하고 있다. 즉, 앰비소닉스 오디오를 메자닌 포맷(Mezzanine format)이라고 불리는 PCM 신호와 메타데이터로 표현한다. 그리고 전송 비트율을 낮추기 위해서 PCM 신호를 다른 채널 오디오와 마찬가지로 USAC으로 압축하여 전송한다.

2.2 렌더링 기술

MPEG-H 3D 오디오 기술을 통해 22.2 채널 오디오 콘텐츠를 제공하더라도 모든 시청자가 22.2 채널

의 스피커를 정해진 표준 위치에 설치하는 것이 아니기 때문에 스테레오 스피커나 다른 위치에 배치된 스피커 또는 헤드폰 사용자에게도 콘텐츠 제작자의 의도에 충실한 오디오 장면을 제공할 수 있어야 한다. MPEG-H 3D 오디오의 주요 특징 중 하나는 바로 다양한 스피커 배치환경과 헤드폰에 최적화된 3D 오디오를 재현할 수 있는 렌더링 기술이라고 할 수 있다.

스피커가 아닌 헤드폰을 통해 오디오 신호를 재현할 경우에는 바이노럴 렌더러(Binaural Renderer)를 통해 원 콘텐츠의 효과를 최대한 반영하는 스테레오 신호를 생성한다. 바이노럴 렌더러는 복원된 고차의 다채널 신호를 스테레오 신호로 변환하는 과정에서 공간상의 스피커 위치에서 발생하는 전달함수를 적용하여 헤드폰을 통해 3D 오디오를 경험할 수 있도록 하는 기술이다.

3. MPEG-I Immersive 오디오

오디오 서브그룹에서는 2016년 10월에 개최된 116차 MPEG 회의를 기점으로 가상현실/증강현실 서비스에 필요한 오디오 기술에 대한 논의를 시작하였다. 오디오 서브그룹에서는 MPEG-H 3D 오디오 기술로 3DoF와 3DoF+ 서비스가 가능하다고 판단하고 사용자가 가상공간에서 현실에서처럼 자유롭게 움직일 수 있는 6DoF를 가능하게 하는 오디오 기술에 대한 논의에 집중하기로 하였다. 현재 MPEG-I 파트 4 몰입형 오디오 코딩(Immersive Audio Coding)이라는 이름으로 6DoF를 제공하기 위한 오디오 기술에 대한 표준화가 진행 중이다.

현재 오디오 서브그룹에서는 2019년 1월에 개최된 125차 회의에서 MPEG-I 오디오의 아키텍처와 요구사항을 확정하고 MPEG-I 오디오 표준 기술 선정에 필요한 절차를 논의 중이다. 현재까지 논의된 표준화 일정은 128차 회의(2019. 10월)에서 인코더 입력 포맷과 평가 플랫폼 확정, 129차 회의(2020. 1월)에서 제안요청서(CfP, Call for Proposal) 진행, 132차 회의(2020. 10월)에서 작업 초안(WD, Working Draft) 발간, 134차 회의(2021. 4월)에서 위원회 초안(CD, Committee Draft) 발간을 목표로 하고 있다.

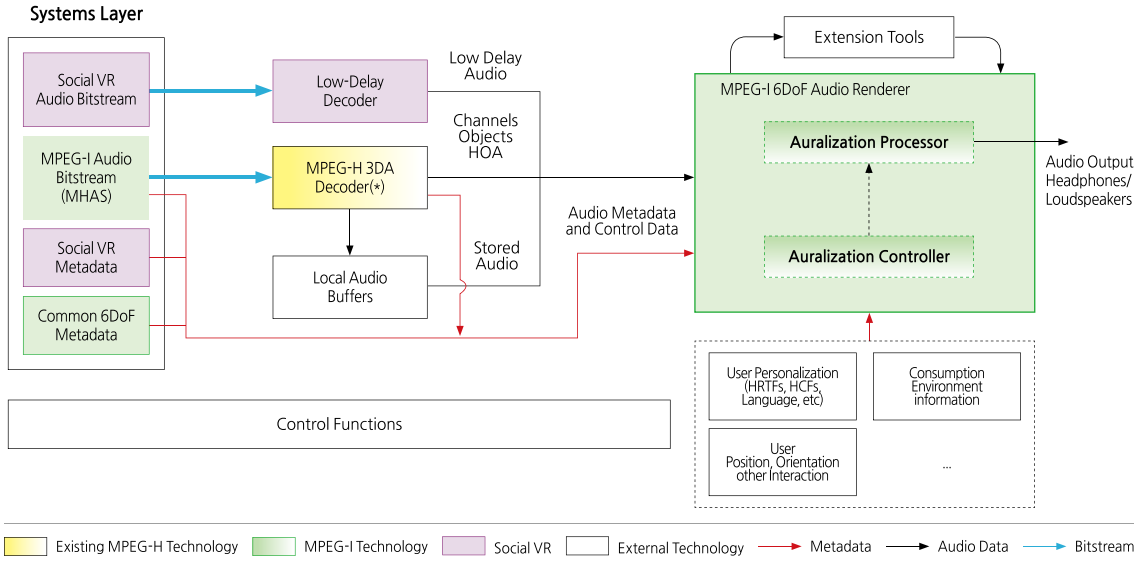
본 절에서는 MPEG-I 오디오의 아키텍처와 요구사항 그리고 MPEG-I 오디오 기술 평가에 활용하게 될 인코더 입력 포맷과 평가 플랫폼에 대해 간단히 살펴본다.

3.1 아키텍처 및 요구사항

오디오 서브그룹에서는 오디오 신호는 MPEG-H 3D 오디오 저-복잡도 프로파일을 사용하여 압축 및 복원하고 6DoF를 제공하기 위해 추가적으로 필요한 메타데이터와 렌더링 기술을 MPEG-I 오디오의 표준화 대상으로 하고 있다. [그림 3]은 125차 회의에서 확정된 MPEG-I 오디오 참조 아키텍처를 나타낸다[8].

[그림 3]에서 연두색으로 표시된 부분이 MPEG-I 오디오를 통해 표준화가 진행되고 있는 기술이다. Social VR은 가상의 공간에서 두 사용자가 만나는 유스케이스를 지원하기 위한 기술로, 이 기능을 지원하기 위한 인터페이스와 메타데이터는 표준화 범위에 포함되지만 오디오 신호를 압축하고 복원하는 기술은 표준화 대상이 아니다.

오디오 렌더러는 실제로 신호처리를 수행하는 가청화 처리기(Auralization processor)와 가청화 처리기를 업데이트하고 오디오 장면 전환에 필요한 처리를 수행하는 가청화 제어기(Auralization controller)로 구성된다. 또한 오디오 렌더러는 인코더로부터 오디오와 메타데이터를 입력받기 위한 API 외에도 외



[그림 3] MPEG-I 오디오 참조 아키텍처[8]

부 메타데이터를 활용하기 위한 API와 외부자원 또는 향후 MPEG 기술로 실현될 수 있는 기능을 활용하기 위한 API를 가지고 있다.

또한 사용자가 가상공간에서 움직일 때 현실 세계에서 움직이는 것과 동일한 느낌을 갖도록 하는 데 필요한 기술들에 대한 요구사항을 정의하였다. 요구사항에는 오디오 품질과 가상공간의 사실적 표현 등에 대한 일반적인 내용, 오디오 렌더러, 인터페이스, Social VR, 재현 환경, 그리고 3DoF와 6DoF 플랫폼 간의 호환성으로 구분된 총 27개 항목이 정의되어 있다.

3.2 인코더 입력 포맷

오디오 서브그룹에서는 CfP에 제안된 기술을 평가하기 위해 인코더 입력 포맷(EIF, Encoder Input Format)을 정의하고 있다[9]. 인코더 입력 포맷은 오디오 장면을 표현하는데 필요한 메타데이터를 MPEG-I 오디오 인코더에 입력하기 위해 사용되는 입력 형식으로 XML 기반의 파일이다. 따라서 제안된 기술의 평가에 사용되는 모든 콘텐츠는 인코더 입력 포맷에 따라 오디오 장면을 서술하는 XML 파일을 갖게 된다.

인코더 입력 포맷에는 사용자의 움직임에 따라 객체, 채널 및 앰비소닉스 오디오의 변화를 표현하는 데 필요한 음원의 위치, 방향 및 지향성과 같은 여러 가지 속성이 정의되어 있다. 또한 공간 정보를 반영하는 실내 음향효과, 특히 회절 및 폐색을 표현하는 데 필수적인 기하학적 구조를 표현하기 위한 프리미티브와 메시에 대한 정의를 포함하고 있다. 공간의 음향 특성을 표현하기 위해 공간 안에 위치한 사물의 반사 및 투과율과 같은 특성에 대해서도 정의하고 있다. 또한 움직이는 오디오 장면을 표현할 수 있도록 시간, 사용자의 인터랙션 또는 특정 조건에 따라 오디오 장면을 수정 및 업데이트하는 방법에 대해서도 정의하고 있다.

[그림 4]는 지향성을 갖는 하나의 객체 오디오(트럼

```
<Audioscene>
  <Audiostream id="signal:trumpet"
               file="armstrong.wav"
               vstchannels="0, 1" />

  <SourceDirectivity id="dir:trumpet"
                     file="trumpet.sofa" />

  <Objectsource id="src:trumpet"
               position="2 1.7 -1.25"
               orientation="30 -12 0"
               signal="signal:trumpet"
               directivity="dir:trumpet:
               gainDb="-2"
               active="true" />
</Audioscene>
```

[그림 4] MPEG-I 오디오 기술 평가를 위한 인코더 입력 포맷 예제[9]

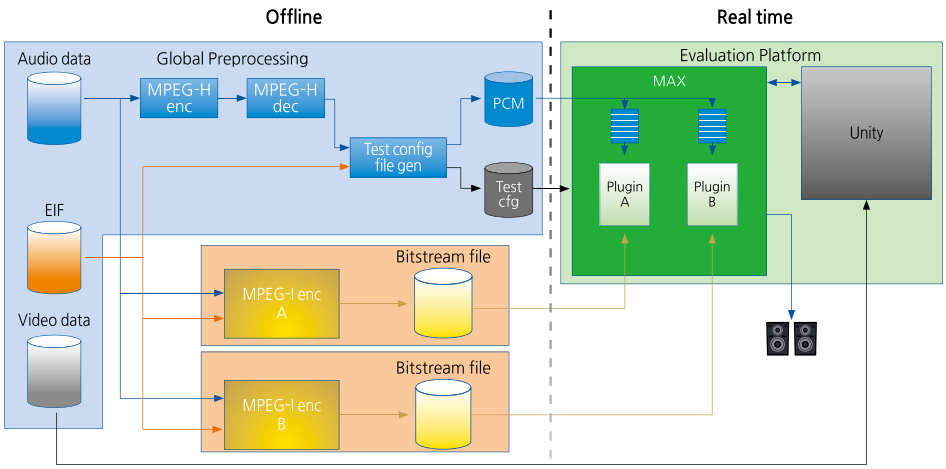
펫 연주)가 있는 오디오 장면을 인코더 입력 포맷에 따라 서술한 XML 파일의 예이다.

3.3 평가 플랫폼

오디오 서브그룹에서는 평가자가 가상 환경을 자유롭게 탐색하면서 제안된 기술을 실시간으로 동시에 평가할 수 있도록 오디오 평가 플랫폼(AEP, Audio Evaluation Platform)을 개발하고 있다[10]. [그림 5]는 오디오 평가 플랫폼의 개요도로 오프라인으로 처리된 데이터가 실시간 평가 플랫폼에 어떻게 연결되는지를 보여주고 있다.

제안된 기술을 평가하기 위해서는 PCM 데이터, 평가에 사용되는 오디오 장면과 평가방법 등에 대한 정보를 갖고 있는 환경 설정 데이터, 그리고 MPEG-I 인코더의 출력 비트스트림을 오프라인으로 미리 구해놓는다. 그리고 실제 평가에서는 오프라인으로 처리된 데이터와 사용자의 움직임 정보를 입력받아 실시간으로 렌더링한 후에 영상은 HMD로, 오디오는 헤드폰 또는 스피커를 통해 출력한다.

오디오 평가 플랫폼은 Unity와 Max를 활용하여 구현하였다. Unity3d는 모든 시각적 요소 및 사용자 인터페이스 제어 등을 수행하는 그래픽 렌더링 엔진으로 사용되고, MaxMSP 8은 모든 CfP에 제안된 렌더러, 테스트 구성 설정, 오디오 콘텐츠, 데이터 처리 및 모니터링을 수행하는 오디오 렌더링 엔진으로 사용된다. [그림 5]에서 MPEG-I Enc A와 B가 6DoF를 제공하기 위한 메타데이터를 압축하는 기술이고 Plugin A와 B는 오디오 신호 렌더링 기술로 각 기관의 CfP 제안 기술을 나타낸다.



[그림 5] 오디오 평가 플랫폼 개요도[10]

4. 맺음말

본고에서는 MPEG에서 진행 중인 MPEG-I 오디오 표준화 동향과 MPEG-H 3D 오디오 기술에 대해 살펴보았다. MPEG-H 3D 오디오 기술은 가상현실/증강현실 서비스에 필요한 채널, 객체, 그리고 앰비소닉스 오디오를 효율적으로 압축할 뿐만 아니라 헤드폰에서 3D 오디오를 재현할 수 있는 기술로 3DoF와 3DoF+ 서비스 제공이 가능하다. 따라서 오디오 서브그룹에서는 6DoF를 제공하기 위해 추가적으로 필요한 메타데이터와 렌더링 기술을 MPEG-I 오디오의 표준화 대상으로 결정하였다. 그리고 MPEG-I 오디오에서 채널, 객체, 앰비소닉스로 표현되는 오디오의 압축 및 복원에는 MPEG-H 3D 오디오 기술을 사용하기로 하였다.

MPEG-I 오디오의 아키텍처와 요구사항은 2019년 1월 회의에서 확정되었으며 인코더 입력 포맷과 오디오 평가 플랫폼은 마무리 단계에 있다고 볼 수 있다. 그러나 CfP 제안 기술을 평가하기 위한 구체적인 방법이나 평가에 사용될 콘텐츠에 대해서는 추가적인 논의가 필요한 상황이다. 따라서 10월에 개최되는 128차 회의 결과에 따라 표준화 일정에 변동이 있을 수도 있다.

최근 5세대 이동통신(5G) 상용화로 초고속, 초저지연의 데이터 전송이 가능해지면서 가상현실/증강현실과 같은 몰입형 미디어 시장의 확대는 더욱 가속화될 것으로 예상된다. 따라서 서로 다른 플랫폼 간의 호환성을 고려한 MPEG-I 오디오 기술 표준화가 몰입형 미디어 시장 활성화에 기여할 수 있을 것으로 기대한다.

※ 본 원고는 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No.2017-0-00072, 초실감 테라미디어를 위한 AV부호화 및 LF미디어 원천기술 개발].

[참고문헌]

[1] ISO/IEC 23008-3:2015, 'Information Technology-High efficiency coding and media delivery in heterogeneous environments-Part 3: 3D Audio, Amendment 3'
[2] ISO/IEC 23003-3, 'Information technology-MPEG audio technologies-Part 3: Unified speech and audio coding'.
[3] ISO/IEC 23003-1:2007, 'Information technology-MPEG audio technologies- Part 1: MPEG Surround'.
[4] ISO/IEC 23003-2, 'Information technology-MPEG audio technologies-Part 2: Spatial Audio Object Coding (SAOC)'
[5] 이미숙, 백승권, 이태진, 'UHDTV 방송 서비스를 위한 MPEG-H 3D 오디오 기술', TTA 저널 제167호, 2016.
[6] A. Murtaza, J. Herre, J. Paulus, L. Terentiv, H. Fuchs, S. Disch: 'ISO/MPEG-H 3D Audio: SAOC 3D Decoding and Rendering', 139th AES Convention, New York, USA, 2015.
[7] S. Spors and J. Ahrens, 'A comparison of wave field synthesis and higher-order Ambisonics with respect to physical properties and spatial sampling', in Paper 7556, 125th AES Conv., San Francisco, CA, USA, Oct. 2008.
[8] N18158, 'MPEG-I Audio Architecture and Requirements', Jan., 2019.
[9] N18618, 'Draft MPEG-I 6DoF Audio Encoder Input Format', July, 2019.
[10] N18627, 'Draft Documentation for the MPEG-I Audio Evaluation Platform', July, 2019.