VR/AR 오디오 기술 및 표준화 동향

정현주 오현오*
㈜가우디오랩 연구원
㈜가우디오랩 대표이사 *

가상현실 및 증강현실(VR/AR)에서 실감 있는 사용자 경험을 제공하기 위해서 음향은 영상 못지 않게 중요한 요소이다. 인간의 청감 특성을 기반으로 재현하는 바이노럴 오디오(Binaural Audio)는 VR/AR 환경에서 상호작용(Interaction)이 가능한 음향을 실시간 렌더링을 통해 사용자에게 제공하기 때문에 필수적인 기술이다. 본 고에서는 VR/AR에 사용되는 오디오 기술의 특징을 고찰한다. 또한, VR/AR과 관련된 국제표준화 단체의 표준화 진행 동향도 오디오 기술 관점에서 다루고자 한다.

I. 서론

'가상현실(Virtual Reality: VR)'이라는 용어는 19세기 프랑스의 시인이자 연출가인 앙토냉 아르토(Antonin Artaud)가 처음 사용한 것으로 알려져 있다. 그는 등장인물과 소품 등이 극장에서 만들어내는 마치 환영과 같은 속성을 "la realite virtuelle" 즉, "Virtual Reality"로 묘사했다[1]. 감독의 연출과 배우의 연기가 빚어낸 가상의 세계를 극장에서 수동적으로 감상하는 형태가 아닌, 능동적으로 보고 듣고 느낄 수 있는 경험으로 가능하게 하고자 하는 열망은 영상, 음향 미디어 기술 발전의 원동력이 되었다.

가상의 세계를 경험하는 방법은 사용자의 태도에 따라 수동적인 경험과 능동적인 경험, 이렇게 두 가지로 나눌 수 있다. 우리에게 가장 익숙한 TV, 극장 스크린, 하이파이 오디오 등과 같이 평면 디스플레이 기반의 미디어 기술은 수동적인 사용자 경험을 기반으로 발전했다. 사실적인 색감, 높은

* 본 내용은 정현주 연구원(☎ 02-562-1968, sc@gaudiolab.com)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.
** 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 018-0-00863, OTT/IPTV 서비스를 위한 클라우드 기반 360/VR 오디오 End-to-End 솔루션 개발 및 사업화)

해상도와 명암비의 화면, 다채널의 오디오 등 영상과 음향 기술 분야에서 특히 많은 발전을 이루었다. 현대 미디어 산업은 영상과 음향, 두 가지가 기술 경쟁의 축으로 근간을 이루게 된다. 한편, 능동적인 경험에서는 사용자가 가상 세계 속에서 얼마나 '현실적'으로 상호작용(Interaction)할 수 있는지가 가장 중요하다. 현재의 시공간이 아닌 다른 어딘가에 있는 듯한 경험(Being There)이 충족되어야 하는데, 이는 우리의 오감이 실제처럼 감각할 수 있어야 가능하다. 그러나 인간의 오감 중에서 촉각, 후각, 미각을 완전히 모사하는 기술의 완성도는 아직 사람들에게 착각을 일으킬 만큼 높지 않다. 반면, 앞서 언급한 바와 같이 상대적으로 영상과 음향 기술의 완성도는 기술적 성숙도가 높기 때문에, 적어도 시각과 청각 두 가지를 완전하게 하는 것이 가상현실을 능동적으로 경험하는 데 있어서 선행되어야 한다.

가상공간에서의 능동적인 사용자 경험을 가능하게 하는 기기로 현재까지 가장 대표적인 형태는 HMD(Head-Mounted Display)를 탑재한 헤드셋 형태의 시스템이다. 1968년 이반 서덜랜드(Ivan Sutheraland)가 그 원형을 고안한 이래 발전을 거듭하여 "VR의 원년"이라고도 불리우는 2016년에는 Oculus, HTC, Sony 등에서 HMD 기기를 연이어 출시하며 시장에 확실히 안착했다. HMD 기반의 VR 헤드셋은 스테레오스코픽 디스플레이(Stereoscopic Display), 헤드폰/이어폰 기반 입체 음향, 사용자의 머리 방향/위치 추적 센서 및 입력 컨트롤러 등을 주요 기능으로 제공하며 영상, 음향, 상호작용이 사용자 경험에서 중요한 요소임을 증명한다. 일찍이 조지 루카스(George Lucas)도 영상 경험 못지 않게 소리 경험의 중요성("I feel that sound is half the experience…")을 강조한 바 있다[2]. 그만큼 가상 세계와 현실의 경계를 없애기 위해서 소리 경험은 영상 경험과 함께 없어서는 안 되는 반드시 필요한 요소이기 때문이다.

본 고에서는 가상현실 및 증강현실(Augmented Reality: AR)에서 사용자에게 실감 있는(Immersive) 경험을 제공하기 위한 기술적 개요를 오디오 관점에서 다루고자 한다. VR/AR에서 필요한 실감오디오 기술의 개요에 대해서 설명하고, 현재 진행 중인 VR/AR 관련 국제표준화 동향에 대해서 오디오의 관점에서 살펴본다.

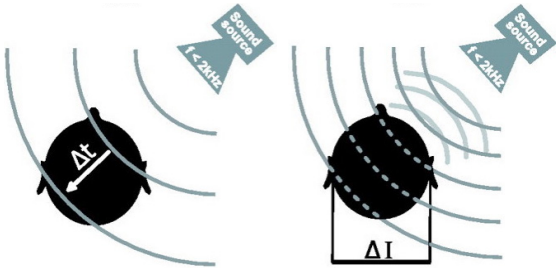
II. VR/AR 오디오 기술 개요

1. 바이노럴 오디오 - 인간의 청각 특성

사람은 두 귀를 통해서 3차원 공간의 소리를 인지한다. 따라서 사람의 두 귀에 전달되는 것과

동일한 특징을 갖는 소리를 시뮬레이션할 수 있다면, 헤드폰/이어폰과 같은 2채널 출력만으로도 입체 음향을 재현할 수 있다. 이러한 오디오 재생 방법을 바이노럴 오디오(Binaural Audio)라고 한다. 먼저 사람이 소리의 위치를 인지하는 원리를 살펴보자.

사람의 양쪽 귀에 들어오는 소리 신호의 차이로부터 얻은 단서를 바이노럴 큐(Binaural Cue)라고 한다. 바이노럴 큐에는 양쪽 귀에 도달하는 신호의 크기 차이를 나타내는 양이 레벨차(Interaural Level Difference: ILD)와 양이 시간차(Interaural Time Difference: ITD)가 있는데, 이 두 가지는 음원의 주파수 대역에 따라 다른 특성을 지닌다. ILD는 사람의 머리 모양과 크기에 따라 좌·우 귀에 도달하는 소리의 크기가 다르게 나타나는 머리가림효과(Head Shadow Effect)에 기인하기 때문에 주로 고주파 영역(대략 2kHz 이상)에서 두드러진 차이를 보인다. 반면, 저주파 영역의 음원은 머리가림효과를 고려하더라도 회절이 잘 일어나 좌·우 귀에 도달하는 신호의 크기 차이가 크지 않다. 하지만 머리 크기에 따른 좌·우 귀 사이의 경로 차이는 여전히 존재하기 때문에 신호가 도달하는 시간차에서 기인한 ITD가 우세하게 작용한다. 이 두 가지 단서를 함께 활용하여 사람은 수평면 상의 소리 방향을 인지한다[3],[4].

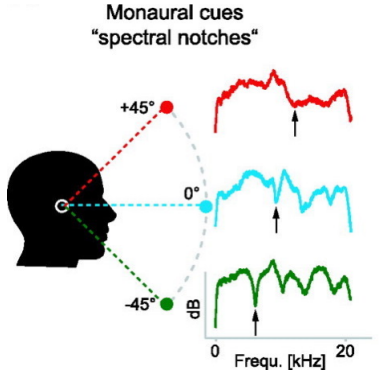


[양이 시간차(ITD)] [양이 레벨차(ILD)]

〈자료〉 Benedikt Grothe, Michael Pecka, and David McAlpine, "Mechanisms of Sound Localization in Mammals", Physiological Reviews, Vol.90, No.3, Jul. 2010, pp.983-1012.

[그림 1] 바이노럴 큐(Binaural Cues)

한편, 3차원 공간상에서 음원을 정확하게 인지하려면 수평면 상의 방향뿐만 아니라 수직면 상의 방향(높이) 또한 필요하다. 이러한 수직 방향의 음원을 인지하는 데 사용되는 단서는 머리와 귓바퀴에 의한 필터링 효과이다. 즉, 양쪽 귀에 인지되는 소리는 특정 주파수 영역에서 보강(Peak)과 감쇄(Notch) 경향을 보인다. 음원의 높이에 따라 귓바퀴의 공명 특성과 어깨에서 반사되는 반사음의 세기가 달라지기 때문이다. [그림 2]와 같이 음원의 높이에 따라 주파수 스펙트럼의 감쇄 주파수가 변하는데, 바이노럴 신호의 주파수 특성이 달라지는 현상은 두 귀에 공통으로 존재하기 때문에 모



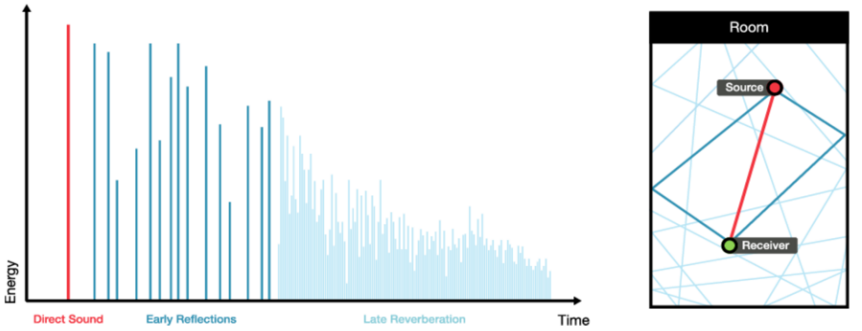
〈자료〉 Benedikt Grothe, Michael Pecka, and David McAlpine, "Mechanisms of Sound Localization in Mammals", Physiological Reviews, Vol.90, No.3, Jul. 2010, pp.983-1012.

[그림 2] 모노럴 큐(Monaural Cue): Spectral Notches

노럴 큐(Monaural Cue)라고 부른다. 이러한 주파수 특성의 변화(Spectral Notches)로부터 사람은 음원의 높이를 인지할 수 있다[4].

음원의 수평 및 수직 방향감과 더불어 음원의 위치를 정의하기 위해 필요한 또 다른 요소는 거리감/공간감이다. 사람이 음원의 거리를 인지하는 가장 중요한 단서는 소리의 크기일 것이다. 멀리 있는 소리가 더 작게 들리고 가까이 있는 소리가 크게 들리는 것은 자명하다. 그러나 동일한 음원이 멀리서 가까이, 혹은 가까이에서 멀리 움직이는 경우가 아니라면 소리의 크기만으로 거리를 판단하는 데에는 한계가 있다. 사람은 멀리에서 나는 큰 폭발음과 아주 가까이에서 들리는 작은 모기 소리를 듣고 그 거리를 쉽게 인지할 수 있다. 즉, 음원의 거리를 인지할 때 소리의 절대적인 크기 이외에 다른 단서를 함께 활용한다.

거리를 인지하는 데 활용하는 단서는 여러 가지이지만 가장 대표적인 것은 직접음(Direct Sound)과 잔향음(Reverberant Sound)의 에너지 비율(Direct-to-Reverberant Energy Ratio: DRR)이다[5]. 공간상에서 음원이 공간의 영향을 받지 않고 직접 사람의 귀로 도달하는 성분을 직접음, 공간으로부터 반사되어 직접음이 도달한 이후에 순차적으로 사람의 귀에 도달하는 부분을 잔향음으로 정의하며, 이 두 성분의 에너지 비율을 통해 음원의 거리를 판단할 수 있다. 같은 공간 상에서 음원의 거리가 가까워지면 직접음 성분의 에너지 비율이 증가하기 때문에 DRR은 상대적으로 커지게 되고 음원이 멀어지면 DRR은 작아진다.



〈자료〉 Gaudio Lab, Inc., 2018.

[그림 3] 공간 상에서 음원의 거리에 따른 직접음(빨간색)과 잔향음(파란색) 에너지 비율의 예

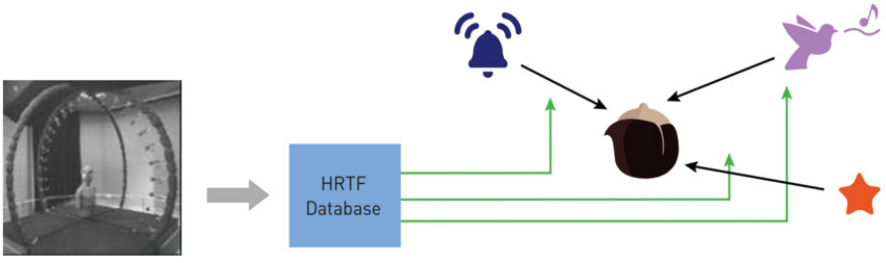
2. 바이노럴 렌더링

바이노럴 오디오를 이용하여 VR/AR 환경에서 3차원 실감 오디오를 제공하려면 이와 같은 단서를 오디오 신호에 잘 반영해야 한다. 일반적으로 이 과정은 앞서 살펴본 개별 단서를 분석하여 오디

오 신호에 반영하기보다는 머리전달함수(Head Related Transfer Function: HRTF)라는 형태로 표현하여 이를 필터링함으로써 이루어진다. HRTF는 3차원 공간의 특정 위치에 음원이 있을 때 음원으로부터 좌, 우 양쪽 귀까지의 공간에 대한 전달함수로, 사람의 머리, 몸통(어깨), 귓바퀴 등의 영향이 모두 반영된다. HRTF는 무향실(Anechoic Chamber)에서 실제 사람의 양쪽 귀에 특수한 마이크를 장착하여 측정하거나, 사람의 상반신 모양을 한 더미헤드(Dummy Head) 형태의 마이크를 통해 측정하여 데이터베이스 형태로 가공된다.

이렇게 특정 공간에서 양쪽 귀에 입력되는 신호를 바로 취득하는 것을 바이노럴 레코딩(Binaural Recording)이라고 한다. 무향실처럼 잔향이 없는 공간에서 여러 방향에서 오는 3차원 상의 신호를 실험적으로 취득하는 것은 HRTF 데이터베이스를 얻기 위한 목적이다. 이와는 다르게 실제로 취득/재현하고자 하는 음향 공간(Sound Scene)의 소리를 직접 바이노럴 레코딩하여 얻을 수도 있다. 바이노럴 레코딩의 경우 마이크가 위치한 공간 전체의 소리를 그대로 녹음하기 때문에 현실감 있는 소리를 취득할 수 있으며, 추가적인 신호처리가 필요하지 않기 때문에 재생 과정도 단순하다. 그러나 녹음하는 당시에 고정된 마이크의 위치와 각도를 재생 시에 변경할 수 없기 때문에 VR 환경에서 상호작용이 가능한 사용자 경험을 제공하는 데에는 한계가 있다.

HRTF 데이터베이스로부터 개별 오디오 신호를 필터링하여 방향감, 공간감을 반영해주는 일련의 신호처리 과정을 바이노럴 렌더링(Binaural Rendering)이라고 한다. 바이노럴 레코딩과는 다르게 개별 오디오 신호의 해당 방향과 위치 정보를 실시간으로 업데이트하여 필터링할 수 있기 때문에 상호작용이 가능한 사용자 경험이 중요한 VR 환경에서 적합하다. 그러나 미리 준비된 HRTF 데이터베이스 등은 무향실이라는 특수한 공간에서 측정한 결과이기 때문에 적당한 잔향(Reverberation)이 존재하는 실제 상황과 큰 차이가 있다. 이 차이를 극복하기 위해 무향 공간이 아닌 잔향이 있는 방에서 측정된 별도의 바이노럴 방 전달함수(Binaural Room Transfer Function, Binaural Room Impulse Response: BRIR) 데이터베이스를 사용하기도 한다. 그러나 이렇게 측정된 BRIR 데이



〈자료〉 오현오, 손주형, 곽진삼, "UHDTV 방송을 위한 차세대 오디오 기술 표준화 동향 소개", 표준특허 전문지 SEP Inside Vol.10, 2016.

[그림 4] 바이노럴 렌더링의 기본 개념도

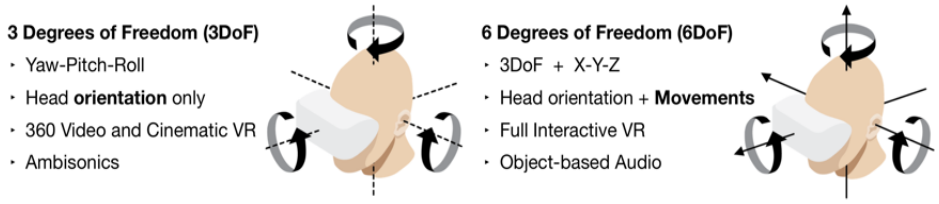
터는 해당 공간의 특정 위치에서만 정의되기 때문에 사용자가 공간 상에서 자유롭게 움직여야 하는 VR 환경에 적용하려면 제약이 따른다. 따라서 공간의 영향에 독립적인 HRTF를 이용하여 음원에 1차적인 방향감을 부여하고, 인공 잔향기(Reverberator) 또는 시뮬레이션(Room Simulation) 등의 후처리를 통해 추가적인 공간감을 제공하는 방법이 통상적으로 바이노럴 렌더링에 사용된다.

실제 현실과 분리된 가상의 세계를 실감 있게 재현하고자 하는 것이 VR의 특징이라면, 실제 현실이 그대로 투영되는 상황에서 가상의 객체나 정보를 합성하여 원래 환경에 존재하는 사물처럼 만들고자 하는 것이 AR의 목적이다. VR 환경에서는 폐쇄형의 헤드폰이나 이어폰을 통한 바이노럴 렌더링 재생이 이상적이지만, AR 환경에서는 이러한 기기를 통해 음향을 재생하게 되면 차음 효과 때문에 현실 세계의 소리를 사용자가 듣는데 제약이 생긴다. 이러한 문제를 해결하려면 AR 환경에서는 바이노럴 렌더링 기술을 사용하는 것과 더불어 고려해야 할 사항이 있다. 폐쇄형의 헤드폰/이어폰을 장착한 경우 외부의 주변음을 들을 수 있는 추가적인 신호처리 방법을 생각해볼 수 있다. 이와 같은 기술은 외부 마이크를 통해 입력 받은 주변음을 실제로 헤드폰/이어폰 없이 듣는 것과 유사하게 왜곡 없이 사용자에게 전달하고자 하는 목적을 지닌다[7],[8]. 한편, 폐쇄형의 헤드폰/이어폰 자체를 사용하지 않는 방법은 간단하게 문제를 해결할 수 있는 접근법이다. 가장 널리 알려진 AR 헤드셋인 마이크로소프트사(Microsoft)의 홀로렌즈(HoloLens)나 매직리프원(Magic Leap One) AR 헤드셋은 헤드밴드에 부착된 형태의 스테레오 스피커를 기본 제공한다. 즉, 외부 주변음의 소리를 그대로 들으면서 가상의 렌더링된 음향이 함께 중첩되어 들리는 형태이다. 이러한 방식은 외부 소리를 잘 들을 수 있다는 장점은 있으나 렌더링된 오디오의 음상 정위(Localization) 성능이나 음질이 상대적으로 헤드폰/이어폰 재생에 비해 떨어진다는 단점이 있다.

3. VR/AR에서 사용자 자유도

VR/AR 환경에서 실감 있는 사용자 경험이 보장되려면 사용자가 가상공간에서도 현실 세계에서와 같이 자유자재로 움직일 수 있어야 한다. 사용자가 임의대로 자유롭게 움직일 수 있는 상태를 “6축의 자유도를 가진 상태”라는 의미로 6DoF(Degree of Freedom)라고 한다. 3차원의 x, y, z축으로의 공간 이동(위치)뿐만 아니라 머리 방향(회전)을 3축(Yaw, Pitch, Roll)에 대해서 움직일 수 있는 상태이다. 반면, 360도 동영상을 감상할 때와 같이 사용자가 임의의 고정된 위치에서 머리 회전만 자유로운 상태를 3DoF라고 정의한다.

진정한 의미의 가상현실을 구현하려면 6DoF가 보장되어야 하지만 아직까지 기술적인 어려움이 따른다. 게임 엔진과 같이 영상과 음향이 실시간 렌더링을 통해 제공되는 경우는 상대적으로 6DoF



[3DoF] [6DoF]

〈자료〉 Gaudio Lab, Inc., 2018.

[그림 5] VR/AR 공간에서의 사용자 자유도

의 가상 환경을 생성하는 것이 수월하지만, 카메라나 마이크 등을 통해 취득된 실사 영상, 음향을 기반으로 6DoF를 재현하기 위해서는 기술적 과제가 많이 남아 있다. 우선 영상 객체의 볼류메트릭 촬영(Volumetric Capture)이 가능해야 한다. 넓은 공간에서 여러 객체를 배경과 함께 고해상도로 6DoF가 가능하게 취득하는 연구가 활발히 이루어지고 있지만 아직 제약이 많은 것으로 알려져 있다. 반면, 오디오의 경우는 각각의 음원을 독립된 오디오 객체(Object)로 미리 녹음하여 렌더링을 통해 6DoF로 재현하는 것이 가능하다.

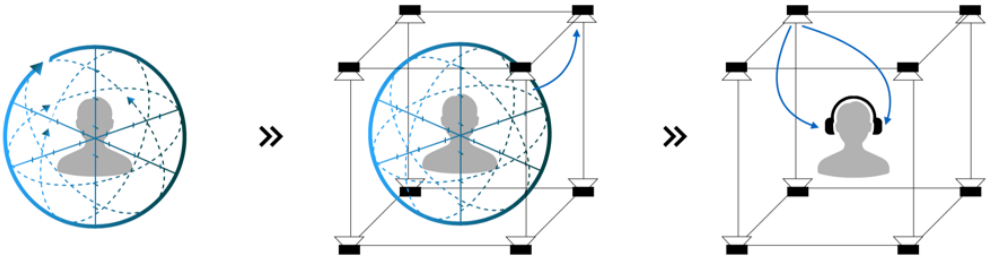
3DoF는 사용자의 위치가 고정된 상태에서 전방위(Omnidirectional)에 대한 영상, 음향 정보를 취득할 수 있으면 재현이 가능하기 때문에 6DoF에 비해 기술적 난이도가 낮다. 360도 카메라를 이용하면 전방위 영상 취득이 가능하고, 마찬가지로 오디오도 앰비소닉스(Ambisonics) 마이크를 사용하여 360도 음향 공간을 녹음할 수 있다. 현재 유튜브(Youtube), 페이스북(Facebook) 등 유수 기업에서는 자사의 플랫폼을 이용하여 360도 동영상을 제공하고 있으며, 사용자도 360도 카메라와 FOA(First Order Ambisonics) 마이크를 이용하여 간편하게 콘텐츠를 제작/업로드할 수 있다.

3DoF+는 3DoF에서 6DoF로 발전하기 위해서 풀어야 할 기술적인 과제를 단계적으로 풀어나가기 위해 도입된 과도기적 개념이다. 기본적으로 6DoF와 같이 머리 회전과 위치 이동이 가능한 형태의 자유도이지만, x, y, z 방향의 이동이 매우 제한적인 범위 내에서(수십 cm 이내)만 허용된다. 예를 들면, 사용자가 의자에 앉아서 일어나지 않은 상태로 머리를 돌아보고 움직일 수 있는 정도의 자유도와 같다. 3DoF와 마찬가지로 3DoF+는 고정된 좌석 위치에서 TV나 극장 스크린을 통해 콘텐츠를 소비하는 데 익숙한 사람들에게 가장 친숙하게 다가갈 수 있는 미디어 형태이다. 영상의 경우는 시점의 변환에 따라 달라지는 시차(Parallax), 객체의 가림(Occlusion) 현상을 해결하기 위해 3DoF+를 별도로 정의하여 접근하고 있으나 오디오 기술의 관점에서는 6DoF와 차이가 없다고 보기 때문에 3DoF+의 문제를 따로 구분해서 다루지는 않고 있다.

4. 오디오 포맷: 채널, 앰비소닉스, 객체

채널은 전통적인 모노, 스테레오, 멀티채널(5.1채널, 7.1채널, …) 등의 오디오 전송/재생 방식이다. 믹싱 또는 마스터링을 통해 제작 단에서 만들어진 형태 그대로 압축, 전송되고 별도의 오디오 렌더링 없이 동일하게 재생되기 때문에 재생하고자 하는 음향 공간(Sound Scene) 그대로를 사전 정의된 스피커 위치에서 일대일 매핑(Mapping)하여 출력한다. 그렇기 때문에 사전 정의된 스피커의 위치에서 배치가 벗어나면 제작자가 의도하지 않은 음상(Localization) 왜곡이 나타날 수 있고, VR/AR에서 필수 요소인 상호작용을 재현하는 데 한계를 보인다.

앰비소닉스 포맷은 음향 공간 상의 특정 위치에서 정의된 모든 방향의 오디오를 획득한 신호이며, 공간의 음장을 구면 조화 함수(Spherical Harmonics) 형태로 표현한 방식이다. 음향 공간의 특성을 그대로 반영하기 때문에 Scene-based Audio라고 표현하기도 한다. 채널 포맷과 달리 신호 자체가 스피커 신호를 포함하지 않기 때문에 재생하려면 가상의 채널 신호로 변환하는 과정이 필수적이다. [그림 6]에서는 앰비소닉 신호를 가상의 스피커 레이아웃 신호로 디코딩하고 해당 스피커 신호를 다시 바이노럴 렌더링하여 재생하는 단계를 나타낸다. 또한, 앰비소닉스 포맷은 차수가 올라갈수록 높은 공간 해상도를 가지는 특성이 있다. 가장 일반적으로 널리 쓰이는 1차 앰비소닉스는 FOA(First Order Ambisonics)라고 하며 B-Format으로도 불린다. 3~4차 이상의 앰비소닉스 형태는 HOA(Higher Order Ambisonics)로 통칭하기도 한다. 구면 조화 함수의 특성 상 신호의 좌표계를 회전시키는 연산이 간단하게 처리되며, 이러한 특성은 Yaw, Pitch, Roll 등, 3DoF의 자유도가 필수적인 VR/AR 또는 360도 동영상에 맞는 오디오 신호로 각광받는 계기가 되었다. 그러나 특정 위치에서 정의되는 음장을 표현할 수밖에 없기 때문에 6DoF의 오디오를 재현하기에는 제약이 있다. 공간 상에서 복수의 위치에 대해서 획득된 표본 신호가 있을 때 임의의 위치를 보간



[Rotatio(3DoF)] [Ambisonics to Channels] [Channels to Binaural]

〈자료〉 Gaudio Lab, Inc., 2018.

[그림 6] 앰비소닉스 포맷의 가상 채널 레이아웃을 이용한 바이노럴 렌더링 방법

(Interpolation)하여 음장을 합성하는 방식이 실험적으로 소개되었으나 아직 상용화되기에는 만족할 만한 음질과 음상 정위 성능을 제공하지 못하는 것으로 알려져 있다[9].

마지막으로 객체 포맷은 개별 음원을 독립된 오디오 신호로 각각 전송하는 방식이다. 개별 음원이 공간 상의 해당 위치 정보(x, y, z 좌표, 방향) 등을 포함하는 메타데이터와 함께 전송되고 재생단에서는 사용자의 현재 위치, 방향에 따라 실시간으로 렌더링하여 재생하기 때문에 6DoF를 포함한 VR/AR 환경에서 가장 이상적인 포맷이라고 할 수 있다. 그러나 객체의 수가 증가함에 따라 전송해야 할 객체 신호의 수가 함께 증가하고, 오디오 렌더링을 위해 필요한 연산량도 선형적으로 증가할 수 있는 단점이 있다. 또한, 거리감, 공간감을 재현하려면 적절한 시뮬레이션(Room Simulation)이 수반되어야 하는데 실제 공간에서 직접 획득한 공간감 대비 아직 완전하지 않다는 한계가 있다.

VR/AR 오디오에서 실감 음향을 재현하기 위해서는 위의 세 가지 오디오 포맷의 상호보완적인 특성을 잘 활용해야 한다. 예를 들면, 믹싱 혹은 앰비소닉스 마이크를 통해 취득한 공간 음향은 채널 또는 앰비소닉스 포맷을 이용하여 3DoF가 가능한 장면에 대해서 표현하고, 심도있는 6DoF 상호작용이 요구되는 음원에 대해서는 객체 포맷으로 전송/렌더링하여 실시간 렌더링을 통해 재생하는 방법을 고려할 수 있다. 따라서 재생 단에서 제작자의 의도대로 동일한 음향 품질을 제공하기 위해서는 바이노럴 렌더링을 수행하는 오디오 렌더러(Audio Renderer)의 성능이 보장되어야 한다. 사용자의 움직임에 반응한 소리가 재생되기까지의 지연시간(Motion-to-Sound Latency)이 인지 불가능할 정도로 매우 작어야 상호작용에 문제가 없으며, 렌더링되어 재생되는 음향 역시 각각의 음원이 공간상의 해당 위치에 잘 정위되어 들려야 영상과 음향 사이의 불일치가 발생하지 않는다. 또한, 가상공간에서 느껴지는 공간감(반사음, 잔향 등의 음향학적 효과)이 잘 표현되어야 한다. 이러한 오디오 렌더러의 중요성 때문에, VR/AR 오디오의 표준화 과정에서 레퍼런스 렌더러(Reference Audio Renderer)의 선정은 매우 중요하게 다루어진다.

5. VR/AR 오디오 제작 S/W

앞서 설명한 세 종류의 오디오 포맷을 입력으로 VR/AR 미디어의 오디오를 편집/제작하는 방법은 재생 플랫폼의 형태에 따라 크게 두 가지로 분류된다. 먼저 360도 동영상과 같이 실사 영상 기반의 3DoF 미디어를 제작하는 경우에는 360도 동영상과 합치되는 오디오 편집이 중요하기 때문에 기존의 평면 디스플레이 기반의 동영상 미디어 제작 방식과 유사하게 DAW(Digital Audio Workstation) 상에서 VR/AR 오디오 제작을 위한 플러그인 형태의 S/W를 이용한다. 반면에, 현재까지 6DoF가 가능한 미디어 형태(파일 포맷)의 콘텐츠는 아직 정의되지 않았으며 여러 국제 표준

[Facebook Spatial Workstation] [Gaudio Works] [Audio Ease 360pan Suite]

[그림 7] DAW 기반의 360도 동영상용 VR/AR 오디오 제작 툴의 예

기구에서 표준화를 진행 중이다. Unity, Unreal 등의 게임 엔진을 통해서 6DoF의 콘텐츠 제작이 가능하며, 관련된 오디오 역시 게임 엔진을 기반으로 제작된다. 게임 엔진에 따라서 사용 가능한 Audio Rendering SDK를 제작자가 선택하여 프로그래밍에 적용할 수 있다.

III. VR/AR 오디오 표준화 동향

VR/AR 헤드셋과 같은 관련 기기들이 연이어 출시되면서 실감 미디어 애플리케이션과 서비스도 점점 더 다양한 형태로 출시되고 있다. 그러나 애플리케이션과 기기 간에 사용되는 플랫폼의 차이로 인해 상호운용성(interoperability) 및 호환성(compatibility)에 아직 제약이 많으며, 이를 가능하게 하는 국제 표준 또한 미흡하다. 관련 표준을 제정함으로써, 재생 플랫폼에 따라 다른 방식으로 제작을 해야 하는 불편을 해소함과 동시에 시장에서 다양한 애플리케이션의 확산을 기대해볼 수 있을 것이다.

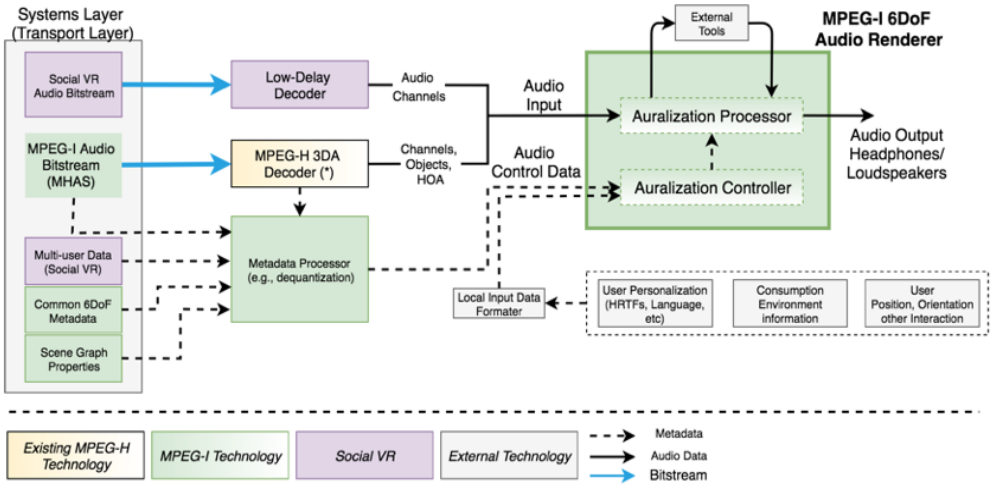
1. MPEG

비디오, 오디오 등을 포함한 멀티미디어 데이터의 국제 표준 개발을 담당하는 MPEG(Moving Picture Experts Group)은 2015년에 MPEG-H 3D 오디오 기술을 실감 오디오(Immersive Audio) 기술의 국제 표준으로 제정했다[13]. MPEG-H 3D 오디오 표준은 본래 UHD TV 방송에 도입할 기술로 시작되었으나, 마침 VR 시장의 급부상에 맞추어 VR 오디오 기술에 대응하기 위한 확장을 시도한 것이다. MPEG-H 3D 오디오 기술은 VR/AR에서 필요한 기본 포맷인 채널, 앰비소닉스(HOA), 객체 오디오에 모두 대응 가능하고 효율적인 연산량으로 처리 가능한 바이노럴 렌더러

[14]를 포함하고 있어 VR/AR 오디오로의 확장에 용이하다. 그러나 3DoF의 자유도까지만 재생 가능하다는 점 때문에 완전한 VR/AR 오디오 기술로 사용되기에는 근본적인 한계를 안고 있었다.

MPEG은 당시 시장에서 널리 퍼지기 시작한 360도 동영상 미디어에 대응하기 위해서 우선적으로 전방위 미디어의 포맷을 다루는 OMAF(Omnidirectional MediA Format) 표준 활동을 2015년 말부터 최초로 시작했다[15]. 이후 이와 관련한 비디오와 오디오, PCC(Point Cloud Compression) 기술 등을 포함하는 새로운 MPEG-I 프로젝트로 확장하며, MPEG-I 사용자 유즈케이스(Use Case)에 따라 빠른 사업화가 가능하도록 파트별로 서로 다른 타임라인으로 단계적인 표준화를 진행 중이다. 기존의 360도 동영상 서비스를 지원하기 위한 3DoF를 고려하는 유스케이스는 가장 먼저 Phase 1a에서 OMAF(MPEG-I Part 2)를 중심으로 진행하고, 제한된 범위의 이동이 추가된 3DoF+는 Phase1b에서 진행한다. 이와 관련된 오디오 기술은 유스케이스가 3DoF 혹은 3DoF+까지로 제한되며, 이는 이미 표준화가 완료된 MPEG-H 3D 오디오에서 충분히 다루어졌다. 따라서 MPEG-I Phase1에서의 오디오 표준화는 별도로 진행되지 않고 OMAF의 오디오 기술로는 MPEG-H 3D 오디오의 저연산 프로파일(Low Complexity Profile)이 포함되었다.

MPEG-I의 오디오 기술은 Part 4에서 Immersive Audio라는 제목으로 진행하고 있으며, MPEG-I Phase2인 6DoF의 유스케이스에 대해 곧바로 표준화가 진행되고 있다. MPEG-I 오디오 분야에서 오디오 압축 코덱으로는 이미 채널, HOA, 객체에 대해 충분한 수준의 압축률을 보이는



*(*) **MPEG-H 3DADecoder** is defined as the core decoder of the MPEG-H 3D Audio Low Complexity (LC) Profile receiving as input in the form of an MHAS stream and providing as output decoded PCM audio (channels, objects HOA) together with all metadata available in the MHAS packets.*

〈자료〉 N18158, ISO/IEC JTC1/SC29/WG11, "MPEG-I Audio Architecture and Requirements", MPEG2019, 2019.

[그림 8] MPEG-I 6DoF Audio Renderer의 아키텍처(Architecture) 구조도

MPEG-H 3D 오디오의 코어 코덱(Core Codec)을 재사용하는 것으로 합의했으며, MPEG-I 오디오에서는 6DoF를 가능하게 하는 오디오 렌더러의 선정과 메타데이터 정의에 집중하고 있다. CfP(Call for Proposal) 발행을 위한 요구사항(Requirements)이 2019년 초 확정되었고[16] 2021년경까지 표준화를 완료하는 것이 목표이다.

2. 3GPP

3GPP(3rd Generation Partnership Project) 또한 3GPP 이동통신 규격에 따라 VR/AR 미디어 서비스용 표준화 작업을 진행 중이다. VR Media Services Over 3GPP라는 주제로 VR 미디어 서비스와 관련된 실감 영상, 음향 기술에 대한 기술 보고서를 발간했다[17]. 해당 보고서에서는 오디오 시스템 및 획득, 렌더링 방법, 음질 평가 방법론에 관해 폭넓게 검토되었다. 3GPP에서 바라보는 VR을 위한 오디오 시스템은 Production, Distribution, Rendering의 세 가지 범주 안에 ① Audio Capture System, ② Content Production Workflow, ③ Audio Production Format, ④ Audio Storage Format, ⑤ Audio Distribution Format, ⑥ Audio Rendering System의 여섯 가지로 구성된다. 또한, 오디오의 품질을 평가하는 지표(Quality of Experience: QoE)는 음상정위 정확도, 음색 변화, 공간감 등 여섯 가지 인자로 정리했으며 각 요소별로 얼마나 상관관계가 높은지 음질 평가 방법에 반영했다.

3GPP에서 진행되는 VR 관련 오디오 표준화 연구 과제(Work Item) 중 단방향의 VR 미디어 전송을 위한 VRStream(Virtual Reality Profiles for Streaming Media) 규격은 2018년 10월에 표준이 완료되었다. Release 15로 발간된 VRStream의 오디오 기술은 4개 기관이 참여하여 평가를 진행했으며, 최종적으로 MPEG-H 3D Audio Low Complexity Profile을 사용하는 OMAF

Quality Features	1,5	1,7	2,0	2,2	2,4	2,6	average
Localisation accuracy	1,4	3,2	1,8	3,2	3,0	3,5	2,7
Timbre	1,8	0,4	2,3	4,0	2,8	2,8	2,3
Auditory spaciousness	1,6	2,0	1,8	2,3	3,0	3,3	2,3
Artefacts	1,2	2,4	3,5	1,8	0,8	2,7	2,0
Reverberance	2,6	0,8	1,5	0,8	2,5	2,0	1,7
Dynamic accuracy	0,5	1,4	1,5	1,3	2,3	1,7	1,4
Quality Elements	Content	Head tracking	Audio processing, coding	HRTFs	Recording technique, mic setup	Binaural Rendering	

〈자료〉 3GPP TR 26.918: "Virtual Reality(VR) media services over 3GPP", Release 15, 2018.

[그림 9] VR 오디오의 음질 평가를 위한 평가 요소별 상관관계(3GPP)

3D Audio Baseline Media Profile이 선정되었다[18]-[20]. 양방향의 VR 미디어 전송 및 음성 통신을 위한 IVAS(EVS Codec Extension for Immersive Voice and Audio Services) 표준은 Release 16에서 완료를 목표로 진행 중이다.

3. VRIF

VR 관련 산업체를 중심으로 발족된 VRIF(Virtual Reality Industry Forum)는 VR/AR 서비스의 상호 호환성을 위한 가이드라인을 제공하는 역할을 담당한다. VRIF에서는 CES 2018에 맞추어 첫 번째 가이드라인 문서를 발행했다[21]. 이 문서는 주로 3DoF를 지원하는 360도 동영상 미디어의 제작-전송-재생과 관련된 생태계를 다루고 있다. 기본 미디어 포맷을 MPEG의 OMAF로 제안하기 때문에 관련 오디오 코덱으로는 MPEG-H 3D 오디오를 지원할 수 있다. VRIF에서 발행하는 가이드라인 문서는 의무 사항이라기보다는 관련 산업체가 참고하여 상호운용성을 최대한 만족시킬 수 있도록 제안하는 성격이 강하다. 현재는 실시간 방송 스트리밍 서비스(Live VR Service)와 HDR(High Dynamic Range)을 지원하는 내용을 포함하는 2.0 버전의 가이드라인 문서 작업을 진행 중이다.

IV. 결론 및 시사점

2014년 Facebook의 Oculus 인수로 시작된 VR 관련 산업에 대한 투자 확대와 2016년 다양한 VR HMD 출시를 계기로 선도적 수요층에서부터 시장 확대가 이루어졌지만 현재 그 상승세가 다소 소강된 분위기다. 그러나 VR HMD 제작사들은 저렴한 가격으로, 편리하게 사용 가능한 경쟁력 있는 HMD 기기를 출시하며 새로운 수요를 끌어들이려 노력하고 있다. 특히, AR 시장은 VR 시장 규모와 비교하여 더욱 큰 잠재력이 예측되므로 관련 기술 개발의 투자가 지속적으로 확대될 것으로 예상된다. 또한, 5세대 이동통신(5G) 시장 개막을 앞둔 현 시점에서 대용량 데이터 전송을 필요로 하는 VR/AR 시장의 확대는 더욱 가속화될 전망이다. 실감 있는 사용자 경험을 제공하기 위해 필수적인 오디오 기술 또한 그 중요성이 더욱 부각될 것이다. VR/AR 오디오에서 필요한 통일된 형태의 포맷과 코덱은 아직 정의되지 않았지만, 이와 관련된 국제 표준 기술을 제정하여 서로 다른 플랫폼, 단말기 간의 호환성을 높이고 콘텐츠 제작/소비가 수월해지면 VR/AR 시장 확대에 견인 역할을 할 것으로 기대한다.

[참고문헌]

[1] Antonin Artaud, The Theatre and its Double Trans. Mary Caroline Richards, New York: Grove Weidenfeld, 1958.

[2] George Lucas, "Technology and the Art of Filmmaking", by Larry Blake, Mix Magazine, Nov.1, 2004.

[3] Rayleigh L. XII. "On our perception of sound direction", The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 1907.

[4] Benedikt Grothe, Michael Pecka, and David McAlpine, "Mechanisms of Sound Localization in Mammals", Physiological Reviews, Vol.90, Mo.3, Jul. 2010, pp.983-1012.

[5] Kolarik, Andrew J. et al. "Auditory Distance Perception in Humans: A Review of Cues, Development, Neuronal Bases, and Effects of Sensory Loss." Attention, Perception & Psychophysics, Vol.78, No.2, Feb. 2016, pp.393-395.

[6] 오현오, 손주형, 곽진삼, "UHDTV 방송을 위한 차세대 오디오 기술 표준화 동향 소개", 표준특허 전문지 SEP Inside Vol.10, 2016.

[7] Guido Baldovino and Michele Geronazzo, "Audio augmented reality headset: a product requirements research in today's available technologies", AES Conference on Headphone Technology, Audio Engineering Society, 2016.

[8] Georgios Marentakis and Rudolfs Liepins. 2014. "Evaluation of hear-through sound localization", In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2014.

[9] J. G. Tylka and E. Choueiri, "Soundfield navigation using an array of higher-order ambisonics microphones", 2016 AES International Conference on Audio for Virtual and Augmented Reality, Audio Engineering Society, 2016.

[10] https://facebook360.fb.com/spatial-workstation/

[11] https://www.gaudiolab.com/sol-vr360-sdk/

[12] https://www.audioease.com/360/

[13] International Standard, ISO/IEC 23008-3:2015, Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio, ISO/IEC, 2015.

[14] Taegyu Lee, et al. "Scalable Multiband Binaural Renderer for MPEG-H 3D Audio", IEEE Journal of Selected Topics in Signal Processing, Vol.9, No.5, Aug. 2015, pp.907-920.

[15] Commette Draft, ISO/IEC 23090-2, Information technology - Coded representation of immersive media - Part 2: Omnidirectional media format, ISO/IEC, 2017.

[16] N18158, ISO/IEC JTC1/SC29/WG11, "MPEG-I Audio Architecture and Requirements", MPEG2019, 2019.

[17] 3GPP TR 26.918: "Virtual Reality(VR) media services over 3GPP", Release 15, 2018.

[18] 3GPP TS 26.259: "Subjective test methodologies for the evaluation of immersive audio systems", Release 15, 2018.

[19] 3GPP TR 26.818: "Virtual Reality(VR) streaming audio: Characterization test results", Release 15, 2018.

[20] 3GPP TS 26.118: "3GPP Virtual reality profiles for streaming applications", Release 15, 2018.

[21] VR Industry Forum Guidelines(Version 1.1), http://www.vr-if.org/guidelines/