MPEG-H 3D 오디오 바이노럴 렌더링 기술 표준화

박 영 철 / 연세대학교 컴퓨터정보통신공학과
이 태 규, 윤 대 희 / 연세대학교 전기전자공학과

I. 서 론

UHD(Ultra high-definition display)와 HMD(Head mounted display)의 등장으로 사람이 보다 몰입감있는 멀티미디어를 소비할 수 있는 시대가 도래하고있다[1]. 특히 스마트폰의 대중화와 초고속 무선 통신의 발달로 고품질 멀티미디어를 모바일 환경에서 소비하는 사례가 보편화됨에 따라 많은 멀티미디어를 헤드폰 또는 이어폰을 통하여 소비하고 있다[2]. 그러나 현재 UHD나 HMD 콘텐츠에 포함되어 있는 오디오는 몰입감을 부여한다고 하기에는 아직 미숙하다고 볼 수 있다.

사람은 귀라는 두 개의 센서를 이용하여 소리를 듣는다. 단순히 두 개의 마이크로폰으로 3차원 공간을 인지하는 것은 불가능하다. 그러나 사람은 두 귀만으로도 외이, 머리 크기, 몸통 등과 같은 신체 특성과 그 특성에 의한 신호의 변화를 학습하여 3차원 공간을 인지한다. 따라서 사람은 두 귀로 발생하는 소리 자체의 음색 뿐만 아니라 공간 정보를 포함하는 많은 음향 정보를 공간을 인지하는데 사용한다. 머리 전달 함수(Head Related Transfer Function, HRTF)은 특정 3차원 공간상의 위치에서부터 사람의 두 귀까지의 전달함수로서 사람이 어떻게 3차원 공간상에서 소리의 방향을 인지하는데 사용되는 단서들은 포함하고 있다[3]. 머리 전달 함수를 사용하여 모노(mono) 오디오 신호를 필터링하면 바이노럴(binaural) 신호가 얻어지고, 이를 헤드폰 또는 이어폰으로 재생하게 될 경우 원하는 위치에 음상이 정위된다. 이 프로세스를 바이노럴 렌더링(Binaural Rendering)이라고 한다. 만일 머리 전달 함수가 소리를 듣는 사람 자신의 머리 전달 함수일 경우 모노 오디오 신호가 실제 3차원 공간에서 재생될 때와 동일한 음장감을 느낄 수 있다.

일반적으로 사람은 수평면 상의 소리의 방향을 양이단서라 불리우는 양 이간의 레벨 차이(ILD: interaural level difference)와 시간 차이(ITD: interaural time difference)를 이용하여 인지한다. 반면 높이를 지각하기 위해서는 외이에 반사되는 경로에 따라 발생하는 스펙트럼 특성을 이용한다. 사람의 귀의 모양은 모두 다르기 때문에 외이의 반사되는 경로 또한 다르고, 이에 의한 스펙트럼 특성 또한 다르기 때문에 일반화된 머리 전달 함수로는 정교한 높이 정위가 불가능하다. 한편 일반적으로 머리 전달 함수는 무향실에서 측정되는데, 사람은 무향실과 같이 소리가 없는 공간이 익숙하지 않기 때문에 머리 전달 함수를 이용하여 바이노럴 렌더링을 수행하면 소리를 듣는 사람은 매우 어색함을 느끼게 된다. 따라서 재생 공간에 대한 특성을 반영해야 한다.

재생공간의 특성을 반영하는 대표적인 방법은 원하는 재생 공간에서 머리 전달 함수를 측정하면 된다. 이를

BRIR(Binaural Room Impulse Response)이라 부른다. 그러나 BRIR은 일반적으로 HRIR(Head Related Impulse Response, HRTF의 시간영역 신호)에 비하여 상당히 긴 길이를 가지기 때문에, 많은 저장용량과 함께 높은 연산량이 요구된다. UHD나 HMD에서 재생되는 오디오 콘텐츠가 충분한 몰입감을 주지 못하는 주요한 원인이 이러한 높은 연산량과 머리 전달 함수의 개인적 차이 때문이라고 할 수 있다.

MPEG(Moving Picture Expert Group)에서는 이러한 환경 변화에 발맞추어 UHD급 고품질 컨텐츠 재생 환경에 걸맞는 오디오를 재생할 수 있는 새로운 오디오 코덱인 MPEG-H 3D 오디오에 주요한 구성요소로써 바이노럴 렌더링을 표준화 기술 범위에 포함하였다[4]. 현재 MPEG-H 3D 오디오는 DIS(Draft International Standard) 상태로써 가장 상용화 가능성 높은 차세대 방송용 코덱 기술이다[5].

본 기고에서는 MPEG-H 3D 오디오의 바이노럴 렌더링 기술이 표준화 되는 과정과 현황을 먼저 설명하고, MPEG-H 3D 오디오의 바이노럴 렌더링의 기술적 특징을 대략적으로 살펴보고자 한다.

II. MPEG-H 3D 오디오 바이노럴 렌더링 관련 표준화 제정 과정 및 현황

멀티미디어 국제 표준을 주도하고 있는 MPEG에서는 차세대 코딩 표준으로 MPEG-H를 추진하여 왔다[4]. MPEG-H 3D 오디오는 세가지 입력 신호, 즉 채널 기반 오디오 신호, 객체 기반 오디오 신호, 그리고 HOA(High-order Ambisionic) 기반 신호를 지원하도록 되어 있다. MPEH-H 3D 오디오는 103차 회의에서 CfP(Call for Proposal)[4]를 발행하여 105차 미팅에서 채택된 표준 기술인 RM(Reference Model)을 선정하였다. RM을 선정하기 위하여 크게 네가지의 청음 평가를 수행하였는데, 첫째는 코덱의 성능을 평가하기 위한 3개의 비트레이트에서의 음질 테스트, 둘째는 3D 오디오의 경우 다수의 사용자가 사용할 가능성이 높기 때문에 Sweet-spot이 아닌 지점에서도 충분한 퀄리티의 입체음향을 제공하는 지에 대한 Off-sweet spot에서의 퀄리티 측정, 세번째는 모바일 환경을 고려한 헤드폰에서의 바이노럴 렌더링 테스트, 마지막으로는 스피커가 정해진 위치에 배치되어 있지 않을 때에도 충분한 입체 음향을 제공하는 지에 대한 테스트 등이 이루어 졌다. 주목할 점은 CO(Channel + Object) 기반 및 HOA의 Immersive 오디오 컨텐츠를 헤드폰에서 효과적으로 재생하기 위한 기술로써 바이노럴 렌더링이 MPEG-H 표준의 범위에 포함하게 되었다는 것이다.

MPEG 오디오 표준의 경우 CE(Core Experiment)라는 과정을 거쳐 RM 코덱의 성능을 발전 시켜왔다. CE란 RM 선정 이후 RM을 기반으로 새롭게 제안된 성능을 개선하거나 새로운 기능을 제공하는 기술을 추가 또는 삽입함으로써 선정된 RM의 성능을 개선하는 과정이다. 당연히 CE 과정에서 새롭게 제안된 기술에 대하여 해당 기술의 당위성을 검증하고 성능을 측정 및 비교하는 과정을 거친다. CE는 MPEG 고유의 표준화 과정으로써 한 기관의 기술이 독점하는 것이 아니라 다양한 기관의 좋은 기술들이 추가됨으로써 보다 진보된 오디오 코덱이 만들어 질 수 있게 하는 큰 토대이다.

RM이 선정된 105차 회의에서 전자통신연구원(ETRI)에서 제출한 바이노럴 렌더링 기술이 연산량에 비하여 매우 우수한 성능을 보임을 확인하게 되었다. 이를 바탕으로 106차 회의까지 바이노럴 렌더링 CE가 진행되었다. 바이노럴 렌더링 CE의 경우 105차 회의와 106차 회의 사이에 추가적인 참여 의사를 밝힌 기관의 참여가 허가되었으며, ETRI(한국), 연세대-윌러스표준기술연구소(한국), Fraunhofer IIS(독일), Orange Lab(프랑스), Qualcomm(미국), Huawei(중국)의 6개 기관이 참가하였다. 각 기관에서 제출한 기술의 성능 검증을 위하여 위 6개 기관 외에 삼성전자(한국)가 cross-check site로서 추가되어 총 7개의 기관에서 음질 평가 실험을 수행하였다. 음질 평가 실험은 CfP의 내용과 같은 방식으로 RM0를 이용하여 512kbps로 부호화된 음원을 사용하여, 6개 기관이 각자가 제안한 시스템을 이용하여 얻은 바이노럴 오디오 신호를 제출하고, 상호간 청음평가를 수행하는 형태로 진행되었다. 실험 결과 Qualcomm을 제외한 5개 제안 시스템의 취득 점수의 신뢰구간이 서로 겹친 상황에서 평균값을 기준으로 연세대-윌러스 시스템이 가장 높은 점수를 얻었고, 그 다음 ETRI, Orange 등의 순위를 가졌다. 바이노럴 렌더링 기술은 모바일 환경에서

가장 빈번하게 사용 되기 때문에 연산량에 매우 민감하다. 따라서 바이노럴 렌더링 기술은 성능 뿐만 아니라 연산량 또한 매우 중요한 요소이며, 이 또한 바이노럴 렌더링 기술을 채택하는데 큰 요소로 논의되었다. 그러나 입력 신호의 도메인, 연산량 추정법 등에 대하여 매우 치열한 토론이 계속 되었으며, 마지막으로 다양한 BRIR에 대하여 적용 가능한 기술인지를 검증해야된다는 주장에 의하여 보다 길거나 짧은 BRIR에 대하여 107차 회의까지 검증하기로 하였다.

107차 회의에서는 연세대/에트리/윌러스가 제안한 기술이 세가지 BRIR에 대하여 음질 및 연산량에서 우수하며 기존의 MPEG 표준 기술를 재사용 하는 CfP의 철학과 선정된 RM 디코더와의 조화, 마지막으로 기술의 복잡도가 상대적으로 낮다는 등을 근거로 주파수 영역 바이노럴 렌더링 기술로 채택되었다. 연세대/에트리/윌러스가 제안한 구조는 프라운호퍼 연구소의 이공간향 기술을 채용하였으며 계산량 면에서 다른 기술에 비해 매우 효율적인 구조를 가진다.

108차 회의 이후로 채택된 바이노럴 렌더러의 세부적인 알고리즘을 기술하는 CD(Committee Draft) 텍스트와 소프트웨어가 레퍼런스 소프트웨어에 통합되어 공개되었다. 해당 기술은 렌더링 기술 뿐만 아니라 해당 렌더링에 필요한 파라미터화 기술까지 Normative로 포함되었으며 고차 Ambisonics에 사용되었던 시간 영역 바이노럴 렌더링, 파라미터화 기술 또한 Normative로 포함되었다. 110차 회의에서는 DIS (Draft International Standard) 문서가 발행되었으며, FDIS(Final DIS)를 생략하고 바로 IS(International Standard)를 발행하는 빠른 표준화 방식을 진행 할 예정이다.

III. MPEG-H 3D 오디오 바이노럴 렌더링 기술 개요

MPEG-H 3D 오디오 바이노럴 렌더링의 경우 주어진 멀티채널 BRIR에 대하여 고음질을 유지하면서 최대한 효율적으로 멀티채널 신호를 바이노럴 신호로 변환하는데 있다. 이는 스마트폰과 클라우드 서비스 등 모바일에서의 한정된 연산 능력과 저장 능력을 반영하기 위함이다. MPEG-H의 경우 22채널 입력 신호에 대하여 44개의 서로

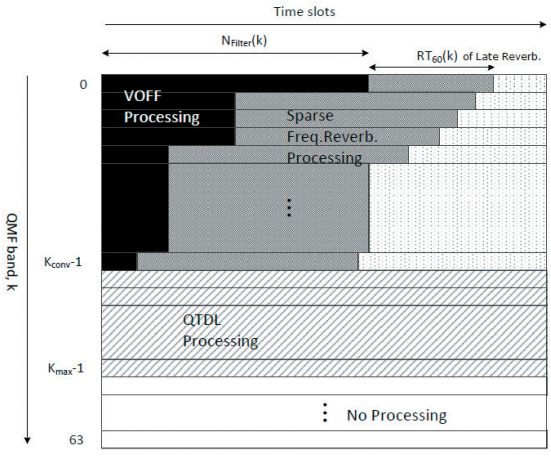

그림 1 MPEG-H 3D 오디오 바이노럴 렌더링의 개념도

다른 필터(BRIR)를 효율적으로 처리하는 것이 가장 중요한 문제가 된다. 전문적인 청음실의 경우 BRIR 필터 길이가 48000샘플, 즉 1초가량 되기 때문에, 48kHz 샘플링 주파수인 경우 22채널 입력신호를 바이노럴 신호로 변환하는데는 MAC(Multiply-accumulate)를 단일 사이클에 수행하는 프로세서를 사용하더라도 22(채널)×2(left/ right)×48000(길이)×48000(샘플율)=101,378MIPS(약 102GHz)라는 엄청난 연산량을 요구하게 된다. MPEG-H 3D 오디오 주파수 영역 바이노럴 렌더링 기술은 위의 연산량을 최소화하면서도 음질을 유지하기 위하여 시간, 주파수 별로 중요도에 따라 다른 처리 방법을 사용하는 것이 특징이다. 그림 1은 MPEG-H 3D 오디오에 대한 이해를 돕기 위한 개념도이다. BRIR을 시간-주파수 그리드로 표현했을 때 각 그리드마다 서로 다른 세가지 기술, VOFF(Variable Order Filtering in Frequency-domain), SFR(Sparse Frequency-domain Reverberator), 그리고 QTDL(QMF Domain Tapped-Delay line) 중 하나로 처리한다.

VOFF는 각 채널에 대하여 각 BRIR의 직접음과 초기 잔향음을 처리하기 위한 알고리즘이다. 직접음과 초기 반사음은 공간을 인지하는데 있어 매우 중요한 요소이기 때문에 직접적인 필터링 과정이 바람직하다. 그러나 직접음과 초기 반사음 구간은 주파수 대역별로 그 길이가 다르기 때문에 대역별로 다른 길이의 필터로 고속 컨볼루션을 수행하게 된다[6]. SFR은 후기 잔향음을 처리하기 위한 알고리즘으로 주어진 BRIR의 후기 잔향음에 상응하는 주파수 종속적인 잔향시간과 바이노럴 형태의

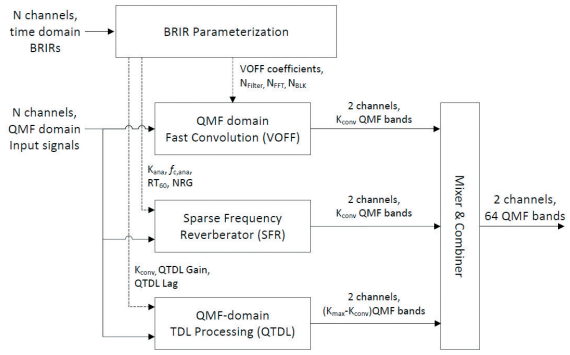

그림 2 MPEG-H 3D 오디오 바이노럴 렌더링의 블록도

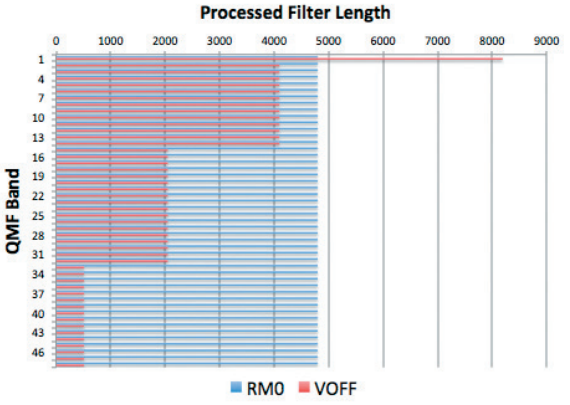

그림 3 종래 기술과 비교하였을 때 VOFF의 효율성

후기 잔향음을 만들어 준다[7]. QTDL은 고주파수 영역에서 보다 효율적으로 바이노럴 렌더링을 하기 위한 방법으로 각 QMF 대역와 각 채널마다 주요한 바이노럴 큐인 레벨 차이와 위상 차이를 반영하기 위하여 각 부대역 BRIR를 1개의 피크로 모델링하여 바이노럴 방 충격 응답을 근사한다[6]. QTDL이 작동하는 대역에서는 VOFF와 SFR이 작동하지 않는다.

최종적으로 VOFF출력과 SFR의 출력이 더해지고 QTDL의 출력과 결합되어 QMF 영역의 바이노럴 출력신호로 만들게 된다. 전체 바이노럴 렌더링 과정을 나타내면 그림 2와 같다. 다음은 MPEG-H 3D 오디오 바이노럴 렌더링에서 사용하는 세가지 핵심 기술에 대해 살펴본다.

1. VOFF 블록

VOFF 기술은 MPEG-H 표준에서뿐 아니라, QMF 영역으로신호가 출력되는 기존의 HE AAC 기반 코덱의 후처리로써 바이노럴 렌더링에 적용 가능한 기술이다. 그림 3은 종래 기술(RM0)와 비교하였을 때 VOFF의 효율성을 보여주는 그림이다. 일반적으로 사람은 공간을 인지하는데 있어 저주파수 대역에 더 민감하며, 고주파 대역에 비해 더 많은 에너지가 분포한다. 종래 기술의 경우 모든 대역에 대하여 같은 길이의 필터로 처리하지만 VOFF는 저주파 대역에서는 필터 길이를 길게하고 고주파 대역에서는 필터 길이를 짧게 설정함으로써 음질의 저하없이 연산량을 줄일 수 있다.

VOFF를 처리하기 위해서는 BRIR 신호를 ISO/IEC 23003-1:2006, Annex B에 따라 복소수 QMF 영역으로 변환하여야 한다. VOFF는 QMF 영역 BRIR을 대역별로 다른 필터 길이로 자른 후(Parameterization 블록) 이를 고속 컨볼루션에 사용한다.

2. SFR 블록

SFR은 QMF 영역의 인공 잔향기로 후기 잔향을 만들기 위하여 사용된다. 입력신호는 주파수 대역별 잔향시간과 후기 잔향음의 에너지값 그리고 QMF영역으로 스테레오 다운믹스된 신호가 된다. 스테레오로 다운믹스된 신호는 입력 신호의 가중합으로 소리를 듣는 사람의 머리를 중심으로 왼쪽 반구에 해당하는 스피커 위치의 출력신호는 왼쪽 채널로, 오른쪽 반구에 해당하는 스피커 위치의 출력신호는 오른쪽 채널로 결정되며 모두 가중치 1을 가진다. 중앙면에 해당하는 스피커는 양쪽 채널모두에 0.7071 가중치를 패닝된다. 인공잔향기의 신호 처리 과정은 각 QMF 대역에 대하여 이루어지며, 입력 믹서는 스테레오신호중 하나를 90도 위상차를 주고 더하여 모노신호로 만든 후, 이 신호를 잔향기에 통과시켜 렌더링을 하게된다. 주어진 주파수 종속 잔향시간과 에너지가 일치하도록 두 개의 상관성 제거기(Decorrelator)의 밀도를 결정한다. 상관성 제거을 위한 위상은 곱하기 연산을 제외하는 등 적은 연산량을 총 4가지 옵션만 제공되며, 상관성이 제거된 두개의 신호는 ICC(Interaural Cross Correlation) 보정 모듈을 통과하면서 원하는 양이간의 코히어런스를 가지는 후기 잔향신호로 만들어지게 된다.

3. QTDL 블록

일반적으로 BRIR의 고주파수 대역은 대부분 소수의

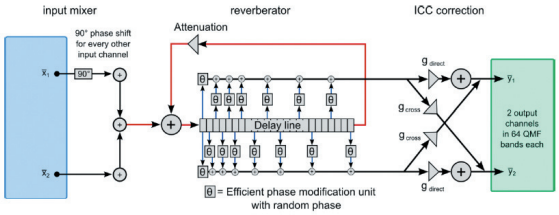

그림 4 인공잔향기의 블록 다이어그램

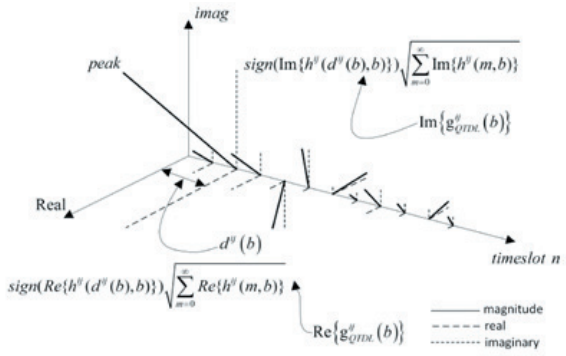

그림 5 QTDL의 개념도

반사음 피크로 이루어져 있으며 사람은 공간을 인지하는데 있어 저주파 대역에 비하여 고주파 대역에 둔감하다. 또한 SBR(Spectral Band Replication)을 사용하는 기존 오디오 코덱들은 고주파수 영역 신호가 다소 왜곡되게 나타날 수 밖에 없다. 이런 특징을 이용하여 QTDL은 VOFF와 SFR이 작동하지 않는 고주파수 영역에 적용된다. QTDL은 고주파 대역 BRIR을 단순 지연된 충격 함수로 근사하여 처리함으로써 가장 중요한 바이노럴 단서인 ITD와 ILD를 각 채널과 QMF 대역별로 일치하도록 근사화 한다.

그림 5는 QTDL의 개념도로써 실선은 고주파 대역의 QMF영역 BRIR을 의미한다. 따라서 BRIR의 반사음 피크 위치에 해당 BRIR의 필터 에너지를 갖는 충격 함수로 근사함으로써 각 QMF 대역의 반향을 1개의 이득과 지연선으로 구현할 수 있다. 이런 단순화 과정을 통해 굉장히 낮은 연산량으로 BRIR을 처리할 수 있다.

IV. 결 론

UHD, HMD 등 많은 고 몰입도 멀티미디어 시스템의 등장으로 사람이 가상현실을 일상적으로 사용하는 날이 멀지 않았다. 본 기고에서는 이러한 고 몰입도 멀티미디어 시스템에서 사용될 수 있는 MPEG-H 바이노럴 렌더링 기술을 소개하였다. 이 기술은 높은 복잡도를 줄이기 위하여 많은 새로운 접근법들을 채용함으로써 고효율의 바이노럴 렌더링 시스템을 구성할 수 있도록 한다. 그러나 아직 머리 전달 함수의 개인화나 고음질 컨텐츠의 부족등의 문제는 해결되지 못한 채로 남아있다. 그럼에도 불구하고 스마트폰의 발전, 클라우드 서비스의 발전 및 보편화 등을 통하여 모바일 환경에서의 바이노럴 렌더링 기술은 다양한 방식으로 진화 될 것이라 전망된다.

참고문헌

[1] K. Hamasaki, T. Nishiguchi, R. Okumura, Y. Nakayama, and A. Ando, "A 22.2 multichannel sound system for ultrahigh-definition TV (UHDTV)." SMPTE Motion Imaging Journal, vol. 117, No. 3, pp. 40-49, 2008

[2] G. Shapiro, "Consumer Electronics Association's Five Technology Trends to Watch: Exploring New Tech That Will Impact Our Lives," IEEE Consumer Electronics Magazine, vol.2, no.1, pp.32-35, Jan. 2013.

[3] Gardner, William G. 3-D audio using loudspeakers. Springer, 1998.

[4] ISO/IEC JCT1 SC29 WG11 Output Doc. N14747, Text of ISO/IEC 23008-3/DIS, 3D audio, The 110th MPEG Meeting, Valencia, Aug. 2014.

[5] ISO/IEC JCT1 SC29 WG11 Output Doc. N13411, "Call for Proposals for 3D Audio," The 103rd MPEG Meeting, Geneva, Jan. 2013.

[6] ISO/IEC JCT1 SC29 WG11 Input Doc. M31297, "Description of Yonsei proposal for MPEG-H 3D Audio Binaural CE," The 106th MPEG Meeting, Geneva, Oct. 2013.

[7] ISO/IEC JCT1 SC29 WG11 Input Doc. M32223, "Technical Description of ETRI/Yonsei/WILUS Binaural CE proposal in MPEG-H 3D Audio," The 107th MPEG Meeting, San Jose, Jan. 2014.