증강현실 오디오 시스템에서의 객체 오디오 획득 및 렌더링 기술

전찬준, 김홍국
광주과학기술원

요 약

본 고에서는 증강현실 오디오에 대한 개념을 기술하고, 이를 실현하기 위하여 필요한 요소 기술들을 논한다. 먼저, 실사 오디오의 객체화를 위하여 실사 오디오의 획득 방법 및 이를 객체화하는 방법에 대하여 서술한다. 또한, 실사 오디오와 가상 오디오의 혼합을 위하여 필요한 실사 오디오의 공간 정보 추출 및 가상 오디오의 공간 정보 부여 기술을 설명한다. 그리고 오디오 증강을 위한 객체 오디오의 렌더링 기술에 대해 소개한다. 마지막으로 이러한 요소 기술들을 통합한 증강현실 오디오 시스템의 사례를 기술하는 것으로 본 고를 마무리한다.

Ⅰ. 서 론

혼합현실(Mixed Reality, MR)이라고 불리기도 하는 증강현실(Augmented Reality, AR)은 현실세계에 가상의 사물이나 정보를 합성해서 원래의 환경에 존재하는 사물처럼 보이도록 하는 기술을 의미하며, 구체적으로 다음의 세 가지 조건을 만족하는 것으로 정의된다. 즉, 첫째는 현실세계와 가상세계가 결합되어야 한다는 것, 둘째는 3차원적으로 표현되어야 한다는 것, 마지막으로 사용자와 실시간 인터랙션이 가능해야 한다는 것이다[1][2]. 최근, 증강현실 기술은 고성능 디지털 시스템 기술의 발전을 근간으로 다양한 분야로의 적용 가능성을 보여주고 있다. 대표적으로, 스마트폰에 내장된 센서를 활용하여 실제 도로에 내비게이션 정보를 덧붙여 보여주는 어플리케이션과 head mounted display (HMD)를 활용한 게임 등이 있다. 그런데 궁극적으로 증강현실을 구현하기 위해서는 이와 같이 시각적인 정보를 증강시키는 것뿐만 아니라, 총체적인 오감 정보를 증강시키는 것이 필요하다. 최근, 국외에서는 청각, 미각, 후각, 촉각에 관한 증강현실 기술에 관한 연구사례가 소개되고 있으며[3][4][5], 국내에서도 이에 관한 관심이 고조되고 있다.

따라서, 본 고에서는 청각과 관련된 증강현실 오디오 기술에 관하여 살펴본다. 이를 위해, 먼저 증강현실 오디오(Augmented Realistic Audio, ARA)에 대한 개념과 증강현실 오디오 서비스를 위한 시스템 구조에 대해 기술한다. 이어서 증강현실 오디오 시스템의 각각의 요소 기술에 대해 설명한다. 특히, 증강현실 오디오 서비스를 위하여 각각의 객체 오디오 획득 방법과 렌더링 기법, 그리고 가상 오디오와 실사 오디오의 혼합을 위하여 공간 정보 추출 및 부여 기술에 대해서 논하도록 한다.

Ⅱ. 증강현실 오디오 시스템에서의 객체 오디오 획득 및 렌더링 기술

증강현실 오디오는 서론에서 언급한 증강현실의 개념을 적용함으로써 정의할 수 있다. 즉, 증강현실 오디오는 현실세계의 소리인 실사 오디오에 가상 오디오를 혼합하는 기술을 의미하고, 구체적으로는 사용자에게 입체적으로 재생되어야 하며, 사용자와 실시간으로 인터랙션이 가능하도록 구현되어야 한다[6]. 〈그림 1〉은 증강현실 오디오 시스템의 구조를 보여준다. 그림에서 보는 바와 같이, 실사 오디오와 가상 오디오가 혼합되는 것을 볼 수 있다. 여기서, 단순하게 혼합하는 것이 아니라 실사 오디오의 공간 정보를 추출하여 가상 오디오에 공간 정보를 부여하고, 이를 혼합한다. 또한, 실사 오디오는 사용자와 실시간으로 인터랙션이 가능하도록 객체화되고 렌더링되어야 한다. 다음의 각 항에서는 증강현실 오디오를 위한 요소기술들을 살펴보도록 한다.

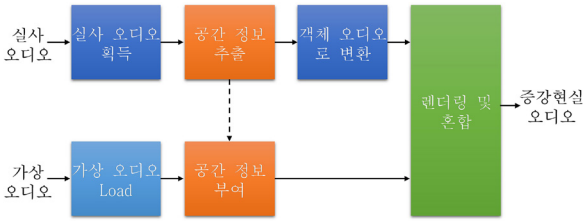

그림 1. 증강현실 오디오 시스템의 구조도

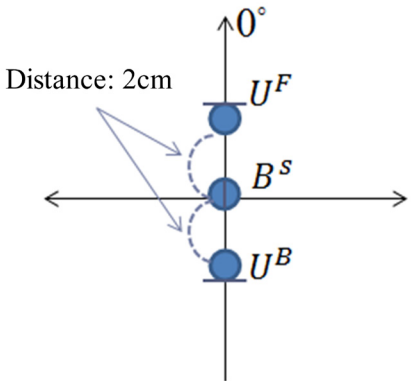

그림 2. Uni- 및 bi-directional 마이크로폰 배치도

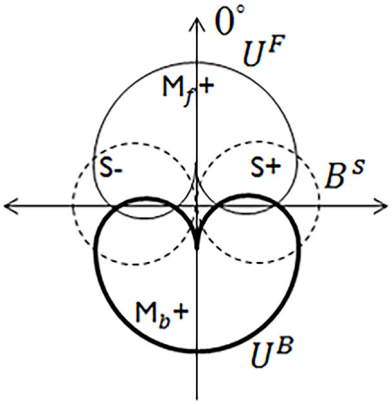

그림 3. Uni- 및 bi-directional 마이크로폰의 directivity pattern

1. 객체 오디오의 획득 기술

증강현실 오디오 시스템에서는 가상 오디오를 실사 오디오에 입체적으로 혼합한다. 또한, 실사 오디오도 사용자와 실시간 인터랙션하며 이동/회전/증강/삭제 등이 용이하여야 한다. 이를 위하여, 실사 오디오의 객체화가 필요하다.

〈그림 2〉는 실사 오디오의 객체화를 위한 마이크로폰 배치의 일례를 보여준다[7]. 그림에서는, uni-directional 마이크로폰을 앞뒤로 배치하였고, bi-directional 마이크로폰을 uni-directional 마이크로폰 사이에 배치하였다. 마이크로폰의 간격은 2cm로 하였다. 〈그림 3〉은 〈그림 2〉에 의해 형성된 각각의 마이크로폰의 directivity pattern을 나타낸다. 그림에서 보는 바와 같이, uni-directional 마이크로폰으로 획득한 신호인 U^F와 U^B는 cardioid 형태의 모양을 갖으며, bi-directional 마이크로폰으로 획득한 신호인 B^S는 뫼비우스 띠 형태를 갖는 것을 볼 수 있다. 이렇게 3개의 마이크로폰으로 획득한 신호는 아래의 수식과 같이 표현이 가능하다.

$$U^F(n)=\left(\frac{1}{2}+\frac{1}{2}\cos\alpha\right)S(n), \quad (1)$$

$$U^B(n)=\left(\frac{1}{2}-\frac{1}{2}\cos\alpha\right)S(n), \quad (2)$$

$$B^S(n)=\left(\cos\alpha-\frac{\pi}{2}\right)S(n)=\sin(\alpha)S(n) \quad (3)$$

여기서, α는 방위각을 나타낸다. 이러한 세 개의 마이크로폰으로부터 획득된 신호를 활용하여 아래의 수식과 같이 omni-directional 및 또 다른 bi-directional 신호를 생성할 수 있다.

$$O(n)=U^F(n)+U^B(n)=S(n), \quad (4)$$

$$B^F(n)=U^F(n)-U^B(n)=\cos(\alpha)S(n). \quad (5)$$

상기의 수식으로 계산된 $O(n)$, $B^S(n)$, $B^F(n)$을 이용하여 원하는 방향의 cardioid 형태의 신호를 형성할 수 있다.

$$U^\theta(n)=\frac{1}{2}\left(O(n)+B^F(n)\cos\theta+B^S(n)\sin\theta\right) \quad (6)$$

여기서, θ는 desired angle을 가리킨다.

식 (6)을 이용하면, 〈그림 4〉와 같이 5 채널에서 사용되는 front left(−30°), front right(30°), center(0°), rear left(−110°), 그리고 rear right(110°) 채널들에 대한 방향으로 directivity pattern을 형성할 수 있다. 이렇게 형성된 5채널 신호를 활용하여 실사 오디오를 객체화가 가능한데, 이를 위하여 먼저 형성된 오디오 신호를 다음 식과 같이 푸리에 변환하여 주파수 도메인 신호로 변환한다.

$$X_j(k)=FFT\{x_j(n)\} \quad (7)$$

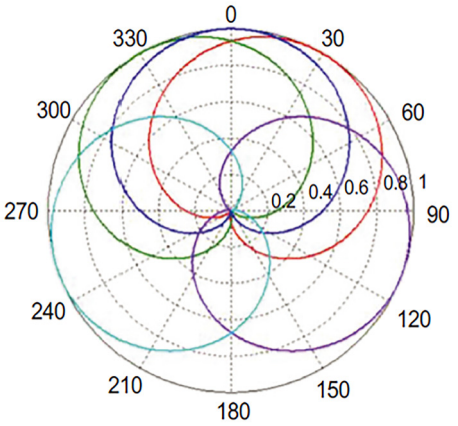

그림 4. 5채널에 대한 directivity pattern

여기서, j는 빔패턴으로 형성된 오디오 채널의 index를 나타낸다. 5.1채널 규격에 맞게 5개의 방향으로 빔패턴을 형성하였다면, j는 1에서의 5까지의 값을 가지게 된다. 편의를 위해, 본 고에서는 j는 1과 2일때의 값을 가지는 기준에서만 설명하도록 한다.

푸리에 변환된 주파수 도메인 신호를 빔형성된 신호끼리 차를 구하여 채널 간의 크기차이 정도를 아래의 수식과 같이 계산한다[8][9].

$$AF(k,i) = \left|X_1(k) - g(i)X_2(k)\right|, \tag{8}$$

$$AF(k,2\beta - i) = \left|X_2(k) - g(i)X_1(k)\right| \tag{9}$$

여기서, $g(i)$는 $0 \le i/\beta \le 1$의 값을 가지게 되고, β는 대한 분해능을 조절하는 인자로 높은 값을 가질수록, 방향에 대한 높은 분해능을 갖게 된다. 예를 들어, β는 100으로 설정될 수 있다. 이렇게 계산된 $AF(k,i)$에서 음원이 존재하는 방향에 대한 i 값에서 최소값을 가지게 된다. 아래의 수식과 같이 최소값이 형성되는 곳(null)을 peak로 변환해 주면서 채널마다 실제 소리를 방향 별로 분석할 수 있다[8][9].

$$\overline{AF}(k,i) = \begin{cases} AF(k,i), & if\ AF(k,i) = \min_i AF(k,i) \\ 0, & otherwise \end{cases} \tag{10}$$

위와 같은 방법으로 계산된 $\overline{AF}(k,i)$에서 마지막으로 d_a와 B를 설정함으로써 아래와 같이 객체 오디오를 획득할 수 있다.

$$\left|Y(k)\right| = \sum_{i=d_a - B/2}^{d_a + B/2} \overline{AF}(k,i). \tag{11}$$

2. 객체 오디오의 공간 정보 획득 기술

실사 오디오와 가상 오디오의 혼합하기 위해서는 실사 오디오의 공간 정보가 필요하다. 이는 다른 공간 정보를 가지고 있는 두 오디오 데이터를 혼합할 경우에는 이질감이 생길 우려가 있기 때문이다. 이에 따라, 본 절에서는 실사 오디오로부터 공간 정보라고 할 수 있는 잔향시간을 추정하는 방법을 살펴보도록 한다.

입력 실사 오디오의 에너지 감쇄 패턴에 대한 Gaussian 확률 분포 기반 maximum likelihood 기준에 의해 최적 잔향시간을 추정할 수 있다[10][11]. 이를 위해, 우선 주어진 시간 프레임의 입력 실사 오디오의 에너지 감쇄 여부를 직전 시간 프레임과의 비교를 통해 확인한다. 이 때, 분산, 최대값, 최소값 등의 세가지 파라메터를 서로 비교하여 에너지 감쇄 여부를 결정한다. 구체적으로, 현재 프레임의 분산과 최대값이 직전 프레임의 분산과 최대값보다 작으며, 현재 프레임의 최소값이 직전 프레임의 최소값보다 크면 현재 프레임에서 에너지가 감쇄 되었다고 판정한다. 아래의 식은 이 과정을 보여준다.

$$D_E^i = \begin{cases} 1, & if\ \sigma_i^2 < \sigma_{i-1}^2\ \&\ x_i^{\max} < x_{i-1}^{\max}\ \&\ x_i^{\min} < x_{i-1}^{\min} \\ 0, & otherwise \end{cases} \tag{12}$$

여기서, i는 프레임 인덱스를 가리키며, 주어진 시간 프레임의 입력 실사 오디오 신호 중 각각 최대값과 최소값이 $x_i^{\max}$와 $x_i^{\min}$이다. 위의 식과 같이 세 가지 조건을 모두 만족할 경우 i 번째 프레임에서 에너지가 감쇄되었다고 판정하며, D_E^i에 1을 할당한다. 그렇지 않은 경우에는 D_E^i에 0을 할당한다. 에너지 감쇄가 일정 프레임 이상 연속적으로 발생할 경우, 해당 에너지 감쇄 구간에 대해 특정 잔향시간에 대한 에너지 감쇄 패턴을 적용한 Gaussian 분포 형태의 확률을 구하여 maximum likelihood 추정 방식을 통해 잔향시간을 추정한다. 또한 그 외에 구간에 대해서는 잔향시간 추정을 수행하지 않으며, 이때 해당 프레임의 잔향시간은 0으로 설정한다.

다음 식은 에너지 감쇄가 일정 프레임 이상 연속적으로 발생할 경우에 해당 구간 신호에 대한 Gaussian 확률을 구하는 과정을 보여준다.

$$L_i(T_{60}^c) = -\frac{N_D}{2}\left((N_D - 1)\ln(T_{60}^c) + M_i(T_{60}^c) + 1\right), \tag{13}$$

$$M_i(T_{60}^c) = \ln\left(\frac{2\pi}{N_D}\right)\exp\left(-3\frac{\ln 10}{T_{60}^c f}\right)\sum_{n=0}^{N_D}\left(x_i^D(n)\right)^2 \tag{14}$$

여기서, $x_i^D(n)$는 i 번째 프레임을 기준으로 이전 프레임들에서 일정 프레임 이상 연속적으로 에너지 감쇄가 이뤄진 구간의 잔향을 나타내며, 본 고에서는 에너지 감쇄가 3프레임 이상 연속적으로 일어날 경우에 대해 예상 잔향시간 후보군, T_{60}^c에 대한 에너지 감쇄 패턴을 적용한 Gaussian 확률을 구한다. 특히, T_{60}^c는 0초에서 1초사이의 구간을 0.05초 간격으로 나눈 21개의 후보 잔향시간을 나타낸다. 또한 N_D는 $x_i^D(n)$의 전체 샘플 수를 의미한다.

다음으로 각각의 후보 잔향시간에 대한 에너지 감쇄 패턴이 적용된 Gaussian 확률들을 다음과 같이 maximum likelihood 추정기에 적용하여 잔향시간을 추정한다.

$$\hat{T}_{i,60} = \max_{T_{60}^c} L_i(T_{60}^c) \tag{15}$$

위의 식을 통해 구해진 $\hat{T}_{i,60}$는 시간 프레임의 변화에 따른 값의 변동을 최소화하기 위해 다음과 같이 평활화 과정을 거치게 된다.

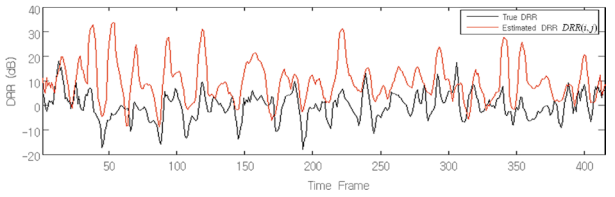

그림 5. 실제 DRR과 DRR 추정치 간 비교

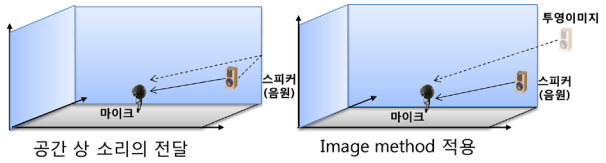

그림 6. 소리의 전달 과정 및 image method

$$T_{i,60} = \beta \hat{T}_{i,60} + (1-\beta) T_{i-1,60} \tag{16}$$

여기서, β 는 평활화 정도를 조정하는 forgetting factor로 본 고에서는 0.9의 값을 할당한다.

시스템 적용 용이성과 성능 측면을 모두 고려하여 두 가지 방식의 잔향시간 추정 방법을 적용해야 한다. 즉, 입력 실사 오디오 전체를 고려하여 상대적으로 정확한 잔향시간 추정치를 얻을 수 있는 global 잔향시간 추정방식과 global 잔향시간 추정방식에 비해 상대적으로 정확성은 다소 떨어지지만 프레임 단위 실시간 구조에 적합한 local 잔향시간 추정방식 등 두 가지 방식을 고려할 수 있다. 우선 global 잔향시간에 대한 추정치는 다음 식과 같이 각 프레임 별로 주어진 $T_{i,60}$ 의 전체 프레임에 대한 평균값을 구함으로써 얻어질 수 있다.

$$RT_G = \frac{1}{N}\sum_{i=0}^{N-1} T_{i,60} \tag{17}$$

여기서, N 은 주어진 입력 잔향음성의 전체 프레임을 나타낸다. 다음으로 local 잔향시간의 경우는 식 (16)를 통해 구해진 $T_{i,60}$ 를 그대로 활용하게 된다. 즉,

$$RT_L(i) = T_{i,60} \tag{18}$$

〈그림 5〉는 잔향성분 추정이 성능을 검증하기 위하여 실제 direct-to-reverberant ratio (DRR)과 DRR의 추정치간을 보여준다[12]. 그림에서 보는 바와 같이, 실제 DRR 패턴과 DRR 추정치의 패턴이 유사함을 알 수 있다.

3. 객체 오디오의 공간 정보 부여 기술

다음으로는 실사 오디오로부터 추정한 공간 정보에 기반하여 가상 오디오에 공간 정보를 부여하는 기술을 살펴본다. 음원으로부터 고막에 최종적으로 전달되는 소리는 음원으로부터 고막으로 직접 입사되는 직접음(direct source)과 벽이나 사물 등에 부딪혀 입사되는 반사음(reflected source)으로 나눌 수 있다[13]. 헤드폰이나 이어폰을 통해 음원을 재생할 경우 직접음은 청취자의 귀로 전달되지만 음의 전달 경로에서 발생하는 공간 정보 중 특히 벽이나 신체 등에 반사되어 오는 반사음은 들을 수 없게 된다. 일반적으로 스튜디오 등에서 녹음된 음성이나 음악 등은 반사음이 충분히 반영되어 있지 않으므로, 공간 정보를 부여하기 위해 인공적으로 반사음 및 잔향을 추가하는 방법을 사용한다. 이러한 반사음 및 잔향을 획득하는 방법은 여러 가지가 있을 수 있는데, 직접 룸 임펄스 응답을 획득하여 이용하는 방법, 인공적으로 룸 환경을 구축하고 그 룸에서의 반사음 및 잔향을 모의실험하여 획득하는 방법 등이 대표적이다[14][15]. 본 고에서는 공간정보 기반의 잔향을 생성하기 위해 사용된 image method에 대해서 설명한다[16][17].

Image method란 공간상에서 전달되는 소리에 대해 벽에 부딪혀 오는 반사음을 벽에 부딪친 거리만큼 떨어진 곳에 위치한 투영 이미지로 고려함으로써 가상의 반사음을 설계하는 기법으로 그 개념도는 〈그림 6〉과 같다. 그림에서 보는 바와 같이, 실제 음원이 위치해 있는 곳으로부터 청취자에 해당하는 마이크까지의 경로에 대해 직접파가 존재하게 되고, 벽에 부딪혀서 전달되어 오는 반사파는 벽과 음원과의 거리만큼 더 떨어진 곳에 가상의 음원, 즉 투영된 이미지가 존재하는 것으로 가정하고 투영된 이미지로부터 마이크로 직접 소리가 전달되어 오는 것으로 룸 임펄스 응답을 설계한다. 〈그림 6〉에서 나타낸 소리의 전달 경로를 1차원 환경을 가정하면 가상 음원의 x 좌표는 아래 식과 같이 구할 수 있다.

$$x_v = (-1)^i x_s + \left[i + \frac{1-(-1)^i}{2} \right] x_r \tag{19}$$

여기서, x_v 는 가상 음원의 좌표, x_s 는 음원의 x 좌표, x_r 은 x 축 상에서의 공간의 길이, 즉 방의 길이를 의미한다.

i 번째 가상 음원의 위치는 식 (19)에 의해 결정되고 만약에 i 가 음수인 경우는 가상 음원은 x 축 상의 반대쪽에 위치하는 것을 의미한다. 또한 i 가 0인 경우에는 가상 음원은 실제 음원에 해당한다. i 번째 가상 오디오와 마이크(청취자) 간의 거리, x_i 는 아래의 수식과 같이 식 (19)에서 구한 가상 음원의 x 좌표, x_v 에서 마이크의 x 좌표, x_m 을 빼줌으로써 얻을 수 있다.

$$x_i = (-1)^i x_s + \left[i + \frac{1-(-1)^i}{2} \right] x_r - x_m . \quad (20)$$

이와 동일하게 3차원 공간상의 y 축 상에서의 가상 음원과 마이크(청취자) 간의 거리 및 z 축 상에서의 가상 음원과 마이크(청취자) 간의 거리를 아래의 수식을 통해 각각 구할 수 있다.

$$y_j = (-1)^i y_s + \left[i + \frac{1-(-1)^i}{2} \right] y_r - y_m , \quad (21)$$

$$z_k = (-1)^i z_s + \left[i + \frac{1-(-1)^i}{2} \right] z_r - z_m \quad (22)$$

여기서, y_s와 z_s는 가상 음원의 y 좌표 및 z 좌표를 각각 의미하고, y_r과 z_r은 공간의 y 축 상의 길이 및 z 축 상의 길이를 각각 의미한다. 또한 y_m 및 z_m은 마이크의 y 좌표 및 z 좌표를 각각 의미한다. 각 가상 음원의 거리는 식 (20), (21), 그리고 (22)에서 구한 x_i, y_j 그리고 z_k를 피타고라스 방정식에 대입함으로써 다음 식과 같이 구할 수 있다.

$$d_{ijk} = \sqrt{x_i^2 + y_j^2 + z_k^2} \quad (23)$$

여기서, d_{ijk}는 3차원 배열로 표현된 각 가상 음원들의 거리를 나타낸다.

상기의 식으로부터 각 가상 음원의 단위 임펄스 응답 함수를 구하기 위해서 아래 식을 먼저 구한다.

$$u_{ijk}(t) = t - \frac{d_{ijk}}{c} \quad (24)$$

여기서, t는 시간, d_{ijk}는 식 (23)에서 구한 거리, 그리고 c는 소리의 속도를 의미한다. 다음으로 단위 임펄스 응답 함수는 아래의 수식과 같이 구하게 된다.

$$a_{ijk}(u_{ijk}) = \begin{cases} 1, & if\ u_{ijk} = 0 \\ 0, & otherwise \end{cases} \quad (25)$$

여기서, a_{ijk}는 단위 임펄스 응답의 크기를 나타내며 구하는 과정은 아래와 같다. 즉, a_{ijk}를 구하기 위해서 공간상에서 음원으로부터 마이크까지 전달되는 반사음의 크기에 영향을 끼치는 두가지 성분을 고려해야 한다. 첫째, 음원에서 마이크로 직접 전달되는 성분인 b_{ijk} 이며 이는 아래 식과 같은 관계를 가진다.

$$b_{ijk} \propto \frac{1}{d_{ijk}} . \quad (26)$$

둘째, 음이 전달되는 동안에 발생하는 반사음들의 수이다. 만약 벽면이 가지는 반사계수가 모두 동일한 경우, 벽면의 반사계수, r_w는 벽면에 의해 생기는 반사음의 총 개수인 $|i|+|j|+|k|$만큼 지수승으로 증가하게 된다. 이와 같이, i, j 그리고 k에 대한 가상 음원의 총 반사계수는 아래의 식으로 표현된다.

$$r_{ijk} = r_w^{|i|+|j|+|k|} . \quad (27)$$

만약 공간의 각 벽이 가지는 반사계수가 각기 다르다면, 각 벽의 반사계수에 대한 반사 성분을 모두 고려해 주어야 한다. 공간의 x 축에 대한 i 번째 가상 음원 성분의 반사성분은 다음 식과 같이 주어진다.

$$r_{x_i} = r_{x=0}^{\left|\frac{1}{2}i - \frac{1}{4} + \frac{1}{4}(-1)^i\right|} \; r_{x=1}^{\left|\frac{1}{2}i + \frac{1}{4} + \frac{1}{4}(-1)^i\right|} \quad (28)$$

여기서, $r_{x=0}$은 음원으로부터 가까이에 있는 x 축 상의 벽을 의미하고, $r_{x=1}$은 먼거리에 떨어져 있는 x 축 상의 벽을 각각 의미한다. 이와 유사하게 y 축 및 z 축에 대한 가상 오디오의 반사성분은 다음의 식 (29) 및 (30)과 같이 각각 구할 수 있다.

$$r_{y_j} = r_{y=0}^{\left|\frac{1}{2}j - \frac{1}{4} + \frac{1}{4}(-1)^j\right|} \; r_{y=1}^{\left|\frac{1}{2}j + \frac{1}{4} + \frac{1}{4}(-1)^j\right|} \quad (29)$$

$$r_{y_j} = r_{y=0}^{\left|\frac{1}{2}j - \frac{1}{4} + \frac{1}{4}(-1)^j\right|} \; r_{y=1}^{\left|\frac{1}{2}j + \frac{1}{4} + \frac{1}{4}(-1)^j\right|} \quad (30)$$

따라서, i, j 그리고 k에 대한 가상 음원의 총 반사계수는 식 (28), (29) 및 (30)으로부터 구한 r_{x_i}, r_{y_j}, 그리고 r_{z_k}를 모두 곱함으로써 구할 수 있다.

$$r_{ijk} = r_{x_i} r_{y_j} r_{z_k} . \quad (31)$$

마지막으로 공간 임펄스 응답열은 식 (25)에서 구한 단위 임펄스 응답과 식 (31)로부터 구한 임펄스 응답의 크기를 곱하고 이와 동시에 모든 축에 대해 더해줌으로써 다음 식과 같이 얻을 수 있다.

$$h(t) = \sum_{i=-n}^{n} \sum_{j=-n}^{n} \sum_{k=-n}^{n} a_{ijk} e_{ijk} . \quad (32)$$

4. 객체 오디오의 렌더링 기술

본 절에서는 객체 오디오의 렌더링 기술에 대하여 살펴본다. 실사 오디오와 가상 오디오 모두다 사용자에게 입체적으로 재

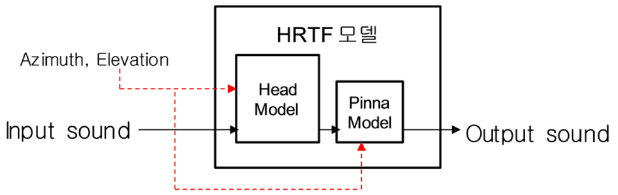

그림 7. 구조적 머리전달함수 모델의 전체 구조

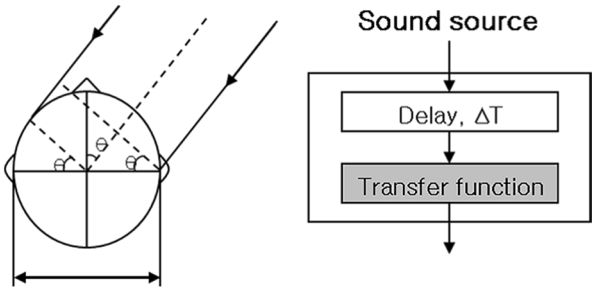

그림 8. 구조적 HRTF에서의 head model

현되었을 때 진정한 증강현실 오디오라고 할 수 있다. 따라서 사용자 개개인에 적합한 머리전달함수가 존재할 경우 이를 활용하여 손쉽게 입체적으로 증강이 가능하다[18]. 하지만, 실제 머리전달함수를 측정하기 위해서는 높은 기술적 난이도와 고비용이 수반된다. 이러한 단점을 극복하기 위한 방법 중의 하나로 머리전달함수를 수학적으로 모델링하는 방법이 연구되어 왔다[19][20]. 이러한 수학적 머리전달함수를 이용하면 방위각, 고도 및 거리를 변수로 가지기 때문에 각각의 경우에 대하여 머리전달함수를 측정하지 않아도 될 뿐만 아니라 알고리즘 적용에 있어서도 메모리 사용면에서 효율적이라 할 수 있다. 또한 청취자의 신체조건을 변수로 표현하여, 청취자에 적합한 머리전달함수를 모델링할 수 있다.

본 고에서 [20]에서 제안된 구조적 머리전달함수 모델(structural HRTF model)에 대해 설명하도록 한다. 구조적 머리전달함수 모델의 전체 구조는 〈그림 7〉과 같다. 그림에서 보는 바와 같이, 구조적 머리전달함수 모델은 청취자의 머리 사이즈 및 음원의 방위각, 고도를 입력으로 받게 된다. 구조적 머리전달함수 모델을 통한 음원의 처리과정은 다음과 같다. 첫째, 입력된 음원에 대해 head model을 통해 머리에 의한 음향 특성을 반영하게 된다. 또한 〈그림 8〉에서와 같이, 머리에 입사되는 음원에 대하여 인간이 음원의 위치를 인지하는데 사용되는 양 귀간 상호 시간 차이(Interaural Time Difference, ITD)를 적용하는 부분과 양 귀간 상호 크기 차이(Interaural Level Difference, ILD)를 반영하는 두 부분으로 나뉜다. ITD는 아래의 식과 같이 표현된다.

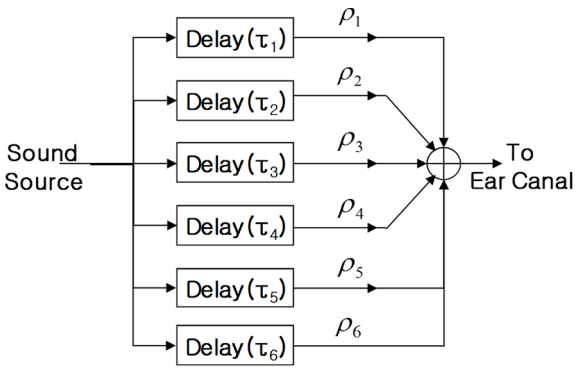

그림 9. 구조적 HRTF에서의 pinna model

$$\Delta T(ITD) = \frac{r(\theta + \sin\theta)}{c} \tag{33}$$

여기서, r은 머리의 반지름, c는 소리의 입사속도, θ는 소리의 입사각을 각각 의미한다. Head shadow 효과에 따른 양 귀간 상호 크기 차이는 아래의 식과 같은 one-pole/one-zero 전달함수로 표현된다.

$$H(z,\theta) = \frac{(2\alpha(\theta) + \beta T) + (\beta T - 2\alpha(\theta))\, z^{-1}}{(2 + \beta T) + (\beta T - 2) z^{-1}} \tag{34}$$

여기서, $\beta = 2c/a$ 이고 a는 머리의 반지름을, T는 샘플링 주기를 의미한다. 그리고 왼쪽 및 오른쪽 head shadowing을 표현하는 계수인 $\alpha_L(\theta)$ 및 $\alpha_R(\theta)$는 음원의 입사 방향각 θ에 따라 아래의 식들과 같이 각각 주어진다.

$$\alpha_L(\theta) = 1 - \sin(\theta), \tag{35}$$

$$\alpha_R(\theta) = 1 + \sin(\theta) \tag{36}$$

여기서, $\alpha_L(\theta)$과 $\alpha_R(\theta)$는 0~2의 값을 가지며, 특히 0인 경우는 최대 head shadowing을 의미하고, 이 때 소리는 해당 귀의 반대편으로부터 전달됨을 의미한다. 즉, $\alpha_L(\theta) = 0$ 인 경우는 오른쪽으로 소리가 전달되는 것을 나타낸다. 또한 $\alpha_L(\theta) = 2$ 인 경우는 소리가 해당 귀와 동일한 방향에서 직접 전달됨을 의미하고, 고주파 영역에서 6 dB만큼의 boosting 된다.

둘째, head model을 통과한 출력 음원은 pinna model을 통해 귓바퀴에 의해 변하게 되는 음향 특성을 반영하게 된다. Pinna model을 통해 귓바퀴에 입사되는 음원은 〈그림 9〉와 같이 모델링되며, 귓바퀴에 의한 반사로 인해 크기 및 도달시간이 바뀌게 된 오디오 신호는 다음과 같이 근사화될 수 있다.

$$s_{out}(n) = s_{in}(n) + \sum_{k=1}^{5} \rho_k s_{in}(n - \tau_k(\theta, \phi)) \tag{37}$$

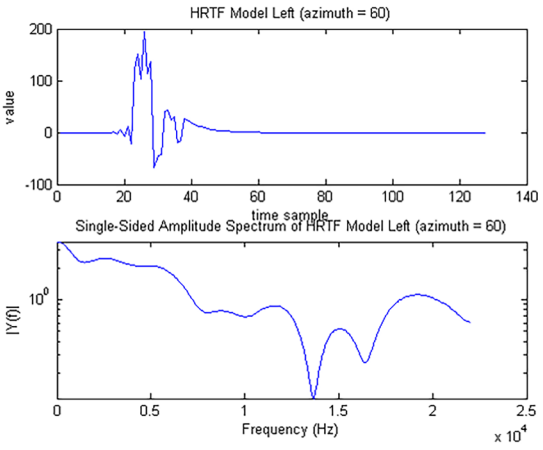

그림 10. 전방 60도에 대한 왼쪽 귀의 HRTF

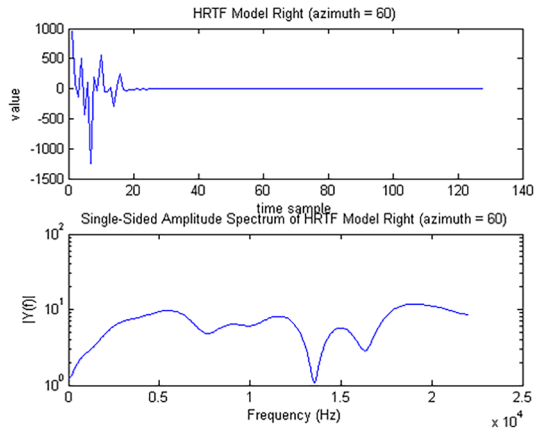

그림 11. 전방 60도에 대한 오른쪽 귀의 HRTF

여기서, $s_{out}(n)$은 pinna model을 통과한 출력신호를, $s_{in}(n)$은 pinna model에 입력되는 입력신호를 나타낸다. 또한, k는 반사음 차수, θ와 ϕ는 방향각과 고도를 각각 의미한다. 식 (37)에서 $\tau_k(\theta,\phi)$는 귓바퀴에 의한 도달 지연시간을 의미하며 다음 식을 통해 근사화된다.

$$\tau_k(\theta,\phi) = A_k \cos(\theta/2)\sin(D_k(\pi/2-\phi)) + B_k. \quad (38)$$

〈그림 10〉 및 〈그림 11〉은 위 과정을 통해 생성된 60도에 대한 왼쪽 귀 및 오른쪽 귀의 HRTF를 각각 보여 준다.

Ⅲ. 증강현실 오디오 시스템

〈그림 12〉는 증강현실 오디오 요소 기술을 통합한 시뮬레이터 예를 보여준다. 그림에서 보는 바와 같이, 증강현실 오디오

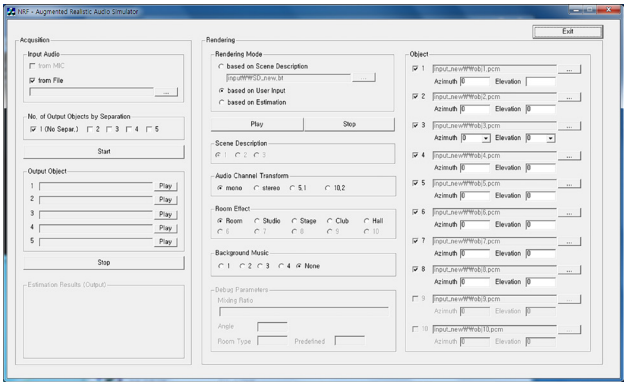

그림 12. 증강현실 오디오 요소기술 통합 Simulator

시뮬레이터는 acquisition 부분과 rendering 부분으로 나눌 수 있다. 여기서, acquisition 부분은 객체 오디오 획득에 해당하며, rendering 부분은 실사 및 가상 오디오의 렌더링에 해당하는 부분이다.

Acquisition 부분에서는 먼저 마이크로폰으로 음원을 직접 입력 받거나, 혹은 사전에 저장된 오디오 파일을 load한다. 그리고 나서, 실사 오디오로부터 지금까지 소개된 기술들을 활용하여 객체 오디오를 분리한다. Rendering 부분에서는 기존의 어떤 방식으로 렌더링할 것인가를 저장해 둔 batch 파일을 활용하는 방법과 사용자가 직접 설정하는 방식으로 구동된다. 사용자가 직접 모든 오디오 객체를 조정한다고 가정한다면, 각각의 오디오 객체를 load하여 azimuth 및 elevation 입력이 가능하다. 입력된 azimuth 및 elevation에 따라 사용자가 느끼는 입체감이 조정되며, "Audio Channel Transform" 부분에서 어떤 환경에서 재생할 것인가를 조정할 수 있다. 또한, room effect 부분에서는 room size를 결정할 수 있다. 사전에 저장된 room size를 몇 가지 설정해두어 여기에 맞는 room이 결정될 수도 있고, 공간 정보기반의 잔향 추정 및 잔향 생성 방법을 활용하여 그에 걸맞은 room effect 적용도 가능하다.

Ⅳ. 결 론

본 고에서는 증강현실 기술에서 증강현실 오디오에 대한 개념이 무엇인지 살펴보고, 이를 실현한기 위한 요소 기술이라고 할 수 있는 객체 오디오 획득 방법, 공간 정보 추출 및 부여 기술, 객체 오디오의 렌더링 기술을 살펴보았다. 이는 증강현실 애플리케이션에서 임의의 객체를 삽입, 삭제, 이동 등의 효과를 구현하기 위해 적용될 수 있다.

Acknowledgement

본 연구는 미래창조과학부 및 정보통신기술진흥센터의 정보통신 · 방송 연구개발 사업[R01261510340002003, 채널/객체 융합형 하 이브리드 오디오 콘텐츠 제작 및 재생기술 개발]과 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2015R1A2A1A05001687).

참 고 문 헌

[1] 이건, "증강현실 기술의 현재와 미래," TTA 저널, 제133호, pp. 88-93, 2011년 11월.

[2] R. Azuma, "A survey of augmented reality," Presence: Teleoperators and Virtual Environments, vol. 6, no. 6, pp. 355- 385, Aug. 1977.

[3] K. Lyons, M. Gandy, and T. Starner, "Guided by voices: an audio augmented reality system," in Proc. of International Conference on Auditory Display, Atlanta, GA, pp. 57-62, 2000.

[4] D.W.F. van Krevelen and R. Poelman, "A survey of augmented reality technologies, applications and limitations," International Journal of Virtual Reality, vol. 9, no. 2, pp. 1-20, June 2010.

[5] O. Bau and I. Poupyrev, "REVEL: tactile feedback technology for augmented reality," ACM Transactions on Graphics, vol. 31, no. 4, article 89, July 2012.

[6] 강진아, 전찬준, 정석희, 김홍국, "증강현실 오디오를 위한 마이크로폰 어레이 설계 및 오디오 객체 획득 기술," 방송공학회지, vol. 19, no. 1, pp. 56-64, 2014년 1월.

[7] J. K. Kim, C. J. Chun, and H. K. Kim, "Design of a coincident microphone array for 5.1-channel audio recording using the mid-side recording technique," Advanced Science and Technology Letters, vol. 14, pp. 61-64, Aug. 2012.

[8] D. Barry, B. Lawlor, and E. Coyle, "Sound source separation: azimuth discrimination and resyntesis," in Proc. of 17th International Conference on Digital Audio Effects (DAFX-04), Naples, Italy, pp. 240-244, Oct. 2004.

[9] 전찬준, 김홍국, "채널 기반에서 객체 기반의 오디오 콘텐츠로의 변환을 위한 비균등 선형 마이크로폰 어레이 기반의 음원분리 방법," 방송공학회논문지, vol. 21, no. 2, pp. 169-179, 2016년 3월.

[10] H. W. Lollmann, E. Yilmaz, M. Jeub, and P. Vary, "An improved algorithm for blind reverberation time estimation," in Proc. of International Workshop on Acoustic Echo and Noise Control, Tel Aviv, Israel, vol. 1, no. 2, pp. 1-4, Sept. 2010.

[11] J. H. Park and H. K. Kim, "Soft-masking based denoising and dereverberation for binaural speech separation in reverberant environments," ICIC Express Letters, vol. 3, no. 3(A), pp. 681-686, Mar. 2013.

[12] F. Rumsey, Spatial Audio, Focal Press, Oxford and Boston, 2001.

[13] J. Blauert, Spatial Hearing: The Psychophysics of Human Sound Localization, MIT Press, Cambridge, MA, 1997.

[14] 서정일, 이용주, 장인선, 유재현, 강경옥, "청취환경 차이에 따른 3차원 오디오 기술 개발 동향," 한국방송공학회지, vol. 13, no. 1, pp. 82-96, 2008년 3월.

[15] P. Rubak, "Headphone signal processing system for out-of-the-head localization," in Proc. of 90th AES Convention, Paris, France, Preprint 3063, Feb. 1991.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," Journal of the Acoustical Society of America, vol. 65, no. 4, pp. 943-951, Apr. 1979.

[17] S. G. McGovern, "Fast image method for impulse response calculations of box-shaped rooms," Journal of Applied Acoustics, vol. 70, no. 1, pp. 182-189, Apr. 2008.

[18] D. R. Begault, 3-D Sound for Virtual Reality and Multimedia, Academic Press, Cambridge, MA, 1994.

[19] D. J. Kistler and F. L. Wightman, "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction," Journal of the Acoustical Society of America, vol. 91, no. 3, pp. 1637-1647, Mar. 1992.

[20] C. P. Brown and R. O. Duda, "An efficient HRTF model for 3-D sound," in Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, pp. 298-301, Oct. 1997.

약 력

전 찬 준

2009년 한국기술교육대학교 전자공학과 공학사
2011년 광주과학기술원 정보통신공학부 공학석사
2011년~현재 광주과학기술원
전기전자컴퓨터공학부 박사과정
관심분야: 오디오 신호처리, 음원 분리, 3D 오디오

김 홍 국

1988년 서울대학교 제어계측공학과 공학사.
1990년 한국과학기술원 전기 및 전자공학과
공학석사
1994년 한국과학기술원 전기 및 전자공학과
공학박사
1990년~1998년 삼성종합기술원 전문연구원.
1998년~1998년 MMC Technology 선임연구원.
1998년~2003년 AT&T Labs-Research Senior
Member Technical Staff.
2014년~2015년 City University of New York,
Visiting Professor
2003년~현재 광주과학기술원
전기전자컴퓨터공학부 교수
관심분야: 음성인식, 음성 및 오디오 신호처리, 3D
오디오, 딥런닝