



実写ベースの 6DoF コンテンツにおける音響制作

Audio production for live-action 6DoF content

杉本岳大¹⁾, 岩崎泰士¹⁾, 久保弘樹¹⁾

Takehiro SUGIMOTO, Taishi IWSAKI, and Hiroki KUBO

1) 日本放送協会 (〒 157-8510 東京都世田谷区砧 1-10-11, sugimoto.t-fg@nhk.or.jp)

概要: 6DoF コンテンツは映像の仮想空間と音響の仮想空間を重ね合わせて構成するため、発音点が明確な音声オブジェクトは適切に映像に合わせて定位させる必要があり、従来コンテンツにおける音響制作の要件とは異なる。そこで、実写ベースの 6DoF コンテンツにおける音響制作の課題を整理するとともに、課題に基づいてゲームエンジンを活用した 6DoF コンテンツの音響制作環境を構築し、実際に 3 種の異なるコンテンツを制作して実用性を確認した。

キーワード: 聴覚, 音響, シーン記述, 音響メタデータ

1. 緒言

NHK は、あたかもその世界に入り込んだかのような体験を提供できるイマーシブメディア [1] を未来の放送メディア技術の重点領域の一つと考え、AR (拡張現実) や VR (仮想現実) を活用したコンテンツの制作や提示技術の研究開発を進めている [2]。イマーシブメディアでは、ユーザが仮想空間内の任意の position¹ でコンテンツを楽しめることを目標としているため [3]、イマーシブメディア用の音響技術も、ユーザの自由な移動に追従して音をレンダリングできるようにユーザの six-degree of freedom (6DoF) 動作への対応が欠かせない [4,5]。AR/VR 用の音響規格である MPEG-I Immersive audio (IA) [6,7] も、6DoF に対応可能な音響技術の実現を目指している。



図 1: 6DoF コンテンツの模式図。6DoF では、ユーザの position によって発音源との幾何学的な関係が変化する。

¹本稿では、position を location + orientation と定義する。

さて近年、被写体映像をボリュメトリックキャプチャでできる映像制作手法が身近になってきており、NHK でも実写ベースの 6DoF コンテンツの制作に取り組んでいる。図 1 に、6DoF コンテンツの模式図を示す。

テレビや映画などの従来メディア用コンテンツとイマーシブメディア用の 6DoF コンテンツにおける制作要件の違いの一つに、映像オブジェクトと音声オブジェクトの定位の一致度がある。従来メディア用コンテンツでは、例えば常に出演者の口唇をトレースしながらセリフを定位させる制作は一般的でなく、多くのコンテンツでは映像の状況に関わらず、セリフは画面センターへ定位させる。しかし 6DoF コンテンツは、映像の仮想空間と音響の仮想空間を重ね合わせて構成することが前提であり、人声や楽器音などの発音点が明確な音声オブジェクトは、映像と音響で空間的な矛盾が生じないように、口唇や楽器などの発音点に定位させる必要がある。そこで本稿では、実写ベースの 6DoF コンテンツにおける音響制作の課題を整理して音響制作環境を構築するとともに、コンテンツ制作における実用性を確認する。

2. 実写ベースの 6DoF コンテンツの映像制作

今回制作したコンテンツは 3 種類で、NHK E テレの番組「おとうさんといっしょ」の主題歌 (図 2)、3 人の登場人物による会話劇「ドラマ」(図 3)、人体浮遊をテーマにした手品「マジック」(図 4) である。登場人物やキャラクターの映像オブジェクトは、個別にボリュメトリックキャプチャした映像を使用した。映像オブジェクトの形式はダイナミックメッシュ形式 [8] で、1 映像オブジェクトあたりのポリゴン数は 10 万個程度、フレームレートは 30 fps である。背景映像は 360° カメラで撮影し、正距円筒図法による矩形映像 [8] を制作して使用した。

映像のシーン記述には、glTF 2.0 [9] に基づいて MPEG

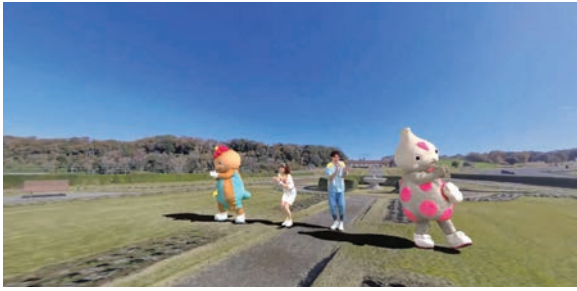


図 2: 制作した「おとうさんといっしょ」.



図 3: 制作した「ドラマ」.



図 4: 制作した「マジック」.

が拡張した形式 [10] をさらに独自拡張した形式 [8] を用いた。映像オブジェクトは、1 フレーム目のバウンディングボックスの中心座標を規準として、映像の仮想空間内の position を指定することで配置した。映像オブジェクトの動作はバウンディングボックス内に限定し、コンテンツの進行に伴うバウンディングボックスの移動はおこなわなかった。

3. 実写ベースの 6DoF コンテンツの音響制作における課題

3.1 6DoF コンテンツの音響制作の要点

6DoF コンテンツの音響制作は、音声信号つまり波形の制作と、音響メタデータによる音響シーンの記述の 2 ステップに分かれる。今回のコンテンツにおける音声信号の制作は、従来のチャンネルベース音響方式 [11] での制作と大きく変えず、マイクロホンを使った收音や収録音の整音などを主たる工程とした。

一方、音響メタデータの制作は、従来の音響制作には含まれない工程である。音響メタデータには、音声オブジェクトの position や放射特性、残響や反射などの音響シーンを特徴づける音響特性などを記述する [12, 13]。コンテンツ全体の構成を規定するという観点からは、音響メタデータ

の制作こそが、6DoF コンテンツの音響制作の要と言える。今回のコンテンツの音響メタデータには、MPEG-I IA 規格 [6] で規定されている Encoder Input Format (EIF) を独自に拡張した形式を用いた。

3.2 音響制作における課題

ボリュメトリックキャプチャした映像を用いた 6DoF コンテンツの音響制作を進めるにあたり、以下のように課題を整理した。

1. 音響メタデータと映像のシーン記述とで、空間を記述する単位系が異なること

EIF の location の単位は実空間に対応した m 系であるが、glTF では実空間と必ずしも一致しない任意の単位系であるため、音響と映像で独立にコンテンツを制作すると、組み合わせて提示した際に仮想空間でのオブジェクトの大きさや position が一致しない。

2. 音声オブジェクトに対応する発音点が、映像オブジェクト上に明示されないこと

6DoF コンテンツの音響制作においては、例えば人声は映像オブジェクトの口唇に、クラップ音は映像オブジェクトの手に定位させる必要がある。しかし、実写のボリュメトリック映像のデータ内に、発音点に当たる人体の部位を明示する情報は存在しない。

3. 音響メタデータをリアルタイムに編集できないこと

EIF は xml 形式のテキストファイルであるため、リアルタイムにレンダリングしながら編集することはできず、編集の都度、全音響メタデータの読み込み直しが発生する。

4. 6DoF コンテンツの音響制作環境の構築および制作手法の考案

前節の課題を解決する目的で、ゲームエンジン (Unity) を活用した 6DoF コンテンツの音響制作環境を構築した。図 5 に、構築した音響制作環境を示す。ユーザ情報は HMD やゲームコントローラを介してゲームエンジンに送信され、その後ゲームエンジンを介して音響レンダラにも送信される。なお音響レンダラは、MPEG-I Immersive audio 規格 [6] を独自に機能拡張したソフトウェアを用いた。

以下に、今回考案した制作手法を示す。

1. EIF と glTF の間の location 変換ルールを規定。
2. ゲームエンジンを用いて、コンテンツとは直接関連しない仮想空間を構築。
3. 個々のボリュメトリック映像 (動画) の取り込み。
4. 発音点を決定するために、音声オブジェクトに対応する球形のダミーオブジェクトを準備。
5. HMD とヘッドホンを装着し、仮想空間内でボリュメトリック映像とレンダリング音を試聴しながら、音声オブジェクトの position を決定。

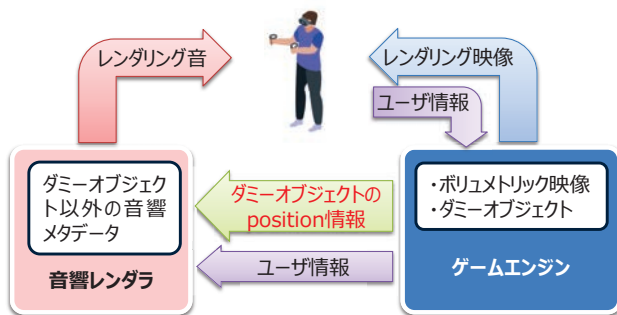


図 5: 考案した音響制作環境。オーディオオブジェクトの position 情報やユーザ情報は OSC を用いて送信した。

音声オブジェクトの position は映像のフレームごとに逐一指定する必要があるため、制作コストが膨大になる。そこで、画像解析を用いて口唇の location を自動的に抽出することで、作業の効率化を図った。画像解析で抽出できなかった orientation は、手作業で追加した。また、着ぐるみや手足などの画像解析に適さない部位は、全て手作業で音声オブジェクトの position を決定した。完成した音声オブジェクトの時系列 position データは、オーディオオブジェクトのアニメーションデータとしてゲームエンジン内に記録し、Open Sound Control (OSC) を用いて音響レンダラにリアルタイム送信した。

5. 6DoF コンテンツの音デザイン

5.1 「おとうさんといっしょ」

コンテンツを構成する音素材は、4 名分の歌声、4 名分のクラブ音、歌の伴奏、アンビエンスである。図 6 に、歌声とクラブ音のオーディオオブジェクトを示す。

歌声は登場人物の口唇に配置し、放射特性を付与した。クラブ音は、メタデータ制作を効率化するため、歌声の position に対して前方 20cm+下方 20cm のオフセットを与えた position を基準として、手を頭上に伸ばすシーンなど大幅に基準からずれたフレームのみ修正した。対応する映像オブジェクトがない歌の伴奏は、聴感上の違和感を避けるために、モノで制作してユーザの頭上 2m に配置した。アンビエンスは中層 4 ch+上層 1 ch の 5 ch で構成した。

5.2 「ドラマ」

コンテンツを構成する音素材は、3 名分の人声、3 名分の足音、カメラのシャッター音、電話の呼び出し音、アンビエンスである。図 7 に、人声・シャッター音・足音のオーディオオブジェクトを示す。

人声は登場人物の口唇に配置し、放射特性を付与した。足音はモノ信号とし、左右の靴の中間に配置した。効果音はそれぞれ発音点に合わせて配置した。アンビエンスは中層 4 ch+上層 1 ch の 5 ch で構成した。

5.3 「マジック」

コンテンツを構成する音素材は、1 名分の人声、スナップ音、背景音楽である。図 8 に、人声とスナップ音のオーディオオブジェクトを示す。スナップ音用のオーディオオブジェクトが 3



図 6: 「おとうさんといっしょ」の音響メタデータ制作の様子。黄と白のオーディオオブジェクトの position が、音声オブジェクトの position に対応する。

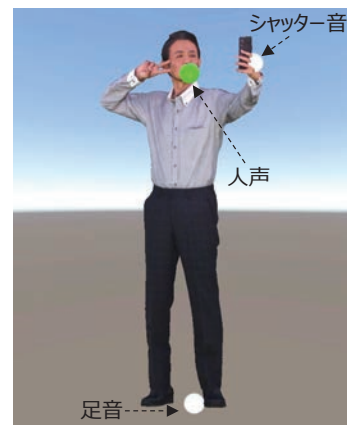


図 7: 「ドラマ」の音響メタデータ制作の様子。緑と白のオーディオオブジェクトの position が、音声オブジェクトの position に対応する。



図 8: 「マジック」の音響メタデータ制作の様子。黄と白のオーディオオブジェクトの position が、音声オブジェクトの position に対応する。

個あるのは、コンテンツ中の異なる 3 か所で、異なるタイミングで発音するためである。

人声は登場人物の口唇に配置し、放射特性を付与した。スナップ音はモノ信号とし、指に配置した。背景音楽は手品の進行に合わせた劇伴として制作し、中層 4 ch+上層 1 ch

の 5 ch で構成した。

6. 結言

本稿では、実写ベースの 6DoF コンテンツにおける音響制作の課題を整理し、課題を解決するための音響制作環境の構築について報告した。また、ジャンルの異なる 3 種のコンテンツを実際に制作し、実用性を確認した。結果として、映像の仮想空間と音響の仮想空間との間で演出的な矛盾を生じることなく、制作意図通りにコンテンツを完成させることができた。制作したコンテンツは、NHK 技研公開 2025 にて一般公開し、多くの来場者に体験してもらった [14]。

今後は、映像の仮想空間と音響の仮想空間の連携をさらに強化することで、より没入感の高いコンテンツ制作の実現を目指していく。

参考文献

- [1] Rep. ITU-R BT.2420-7, Collection of usage scenarios of advanced immersive sensory media systems, 2024.
- [2] NHK 放送技術研究所, Future Vision 2030-2040, https://www.nhk.or.jp/str/future_vision/index.html
- [3] 青木秀一: より自由な映像体験に向けたイマーシブ映像の標準化動向 —360 度映像からイマーシブメディアへの進化—, 日本音響学会誌, 78 巻 3 号, pp. 150–156, 2022.
- [4] 杉本岳大, 堀内俊治: イマーシブメディア実現に向けた音響技術, 映像情報メディア学誌, Vol. 77, pp. 17–21, 2023.
- [5] T. Sugimoto, T. Hasegawa, T. Iwasaki, H. Kubo, S. Oode, and H. Okubo: Requirements for 6DoF audio service and 6DoF audio metadata, 2023 AES Intl. Conf. Spatial and Immersive Audio, Conference paper 33, 2023.
- [6] ISO/IEC 23090-4, Immersive audio, (規格化中) .
- [7] J. Herre and S. Disch: MPEG-I Immersive audio – Reference model for the virtual/augmented reality audio standard, J. Audio Eng. Soc., Vol. 71, pp. 229–240, 2023.
- [8] 青木秀一, 大川裕司, 林田哲哉, 瀧口吉郎: クラウドを利用したイマーシブコンテンツのレンダリングシステム, 情報処理学会論文誌, Vol. 66, No. 6, pp. 988–998, 2025.
- [9] ISO/IEC 12113, Khronos glTF™ 2.0, 2022.
- [10] ISO/IEC 23090-14, Scene description 2023.
- [11] 杉本岳大: 音響規格における音響方式の位置づけ —チャンネルベース・オブジェクトベース・シーンベースの 3 音響方式—, 日本音響学会誌 80 巻 12 号, pp. 675–683, 2023.
- [12] 杉本岳大, 長谷川知美, 大出訓史, 6DoF 音響サービスに用いる音響メタデータの要求条件, 日本バーチャルリアリティ学会大会論文集, 1D1-0, 2023
- [13] 長谷川知美, 杉本岳大, 久保弘樹, 大出訓史, 岩崎泰士, 大久保洋幸: 6DoF 音響サービスのための発音源メタデータの記述法, 日本バーチャルリアリティ学会大会論文集, 1D1-02, 2023 .
- [14] NHK 技研公開, <https://www.nhk.or.jp/str/open2025/index.html>