



リアルワールドメタバース実現に向けた クロスバース・プラットフォームの構築 (2) — 実時間音声伝達機構の設計と開発 —

大塚陸斗¹⁾, 中村文彦²⁾, 木村朝子²⁾, 柴田史久²⁾

1) 立命館大学大学院 情報理工学研究科 (〒 567-8570 大阪府茨木市岩倉町 2-150)

2) 立命館大学 情報理工学部 (〒 567-8570 大阪府茨木市岩倉町 2-150)

概要: 我々は、現実世界を起点としたメタバースであるリアルワールドメタバースを実現すべく、現実世界と仮想世界を相互に行き来可能なクロスバース・プラットフォームの構築を目指している。その一例として大学のキャンパスを舞台としたクロスバース・キャンパスを構想しており、そこでは、現地と遠隔地に存在するユーザ同士が、身振り手振りだけでなく、音声でもコミュニケーションを取ることが可能なシステムの構築を目標としている。本研究では、音源の位置がリアルタイムに変化するクロスバース・キャンパスにおいて、臨場感と没入感を損なわない自然な空間音響の開発と、低遅延でスケーラビリティのある音声通信システムの開発を行う。

キーワード: クロスバース・プラットフォーム, メタバース, 空間音響, 音声通信

1. はじめに

近年、ユーザがアバターを介して VR 空間で交流する「メタバース」が注目を集めている。しかし、従来のメタバースは、VR 空間のみで完結しており、現実世界とは切り離されたものであった。この課題に対して、Niantic 社は現実世界を起点とした新しいメタバースである「リアルワールドメタバース」を提唱した [1]。これは、VR 空間に現実世界のデジタルツインを構築し、現地とリモートの双方からのアクセスを可能にする新たな概念である。リアルワールドメタバースの研究事例として、Microsoft 社が開発した「Microsoft Mesh」がある [2]。これは離れた場所にいるユーザー間で 3D モデルなどのデジタルコンテンツをリアルタイムに共有し、同じ空間にいるかのような共同作業を実現するものである。

こうした背景を踏まえ、我々はリアルワールドメタバースの一例として、大学のキャンパスを舞台とした「MR キャンパス」を構想している [3]。この実現において、現地ユーザとリモートユーザが混在する環境では、音声処理やその通信に起因する遅延が会話の円滑さを著しく損なうという特有の課題が生じる。また、誰がどこで話しているかという位置的な手がかりがなければ、ユーザーの没入感も大きく低下する。そこで本研究では、MR キャンパスにおけるリアルタイム性と没入感を両立する R-V 空間での音声伝達機構の開発を目的とする。ここで、R-V 空間とは現実世界と仮想世界の両空間を指す。具体的には、音の物理的性質をより軽量な処理により再現し、高い没入感を与える「空間音響」と、多人数参加時でも遅延を抑制する「音声通信」の両面からアプローチを行い、その実装と評価を行った。

2. 実時間音声伝達機構の開発

2.1 概要

MR キャンパスの実現にあたり、R-V 空間のユーザの音声を伝達することは重要な課題である。音声は MR キャンパスにおいてユーザ同士がコミュニケーションを取る手段の一つであり、誰がどこで話しているかという情報がユーザの没入感に直結する。既存の音声通話システムでは、どのユーザが話しているかは分かるが、そのユーザがどこにいるのかという位置的な手掛かりが得られないため、複数のユーザが各地で会話を行う MR キャンパスには不向きである。

加えて、リアルタイム性も大きな課題となる。MR キャンパスでは、リモートユーザ同士の会話だけでなく、そこに現地ユーザも混在した会話が発生する。例えば、現地ユーザ同士の会話にリモートユーザが加わると、リモートユーザは必ず遅延が発生した上で会話をすることになる。R-V 空間それぞれにユーザが存在するため、音声処理や音声通信に起因した音の遅延が会話の円滑さに大きく影響する。

現実世界の空間情報をもとに、空間音響を作成する研究も存在するが、音響特性の反映に事前レンダリングが必要であることや、処理に時間を要することから、動的なシーンでの使用は困難である。そこで本研究では、空間音響の再現度を一定のレベルで確保しつつ、リアルタイム性があり、動的なシーンに適応可能な音声伝達機構を開発する。

2.2 空間音響の開発

空間音響の実装には、主に Unity Technologies 社の Unity を用いた。没入感のある空間音響を作成する上で、音の定位、距離減衰、遮音、回折、反響の 5 つの性質を実装した。

2.2.1 音の定位

まず、音の定位は Unity の標準機能を用いて実装した。「Audio Source」という音声再生機能のパラメータの一つである「Spatial Blend」の値を設定することで、ステレオパンニングにより音声が生体的に再生される。これを用いることで、音の定位を強く感じられるようにしている。

2.2.2 音の距離減衰

音の距離減衰は、線形の減衰モデルを利用した。物理法則としては指数関数的曲線の減衰モデルが使用されることが多いが、このモデルでは音が急激に減衰するため、音源との距離感の把握が難しくなる。線形的な減衰モデルは、音の減衰量が距離によって一定であり、音源との距離感が把握しやすいため、MR キャンパスのような同時参加者の多い空間において適している。

また、高周波数の音は減衰しやすく、低周波数の音は減衰しづらいという性質を再現するため、音源とリスナーの距離に応じてカットオフ周波数を調整したローパスフィルタを適用する。

2.2.3 音の遮音

音の遮音を実装するにあたっては、Nintendo が発売している「ゼルダの伝説 ティアーズオブザキングダム」で用いられている音響技術を参考に実装した [4]。具体的には、3次元空間を物体情報をもとにボクセル化し、A*アルゴリズムを用いて経路探索を行い、経路が見つからなかった場合に遮音する処理を実装した。本研究では、VSN-BA という改良された双方向探索を用いた A*アルゴリズムを使用している [5]。Unity シーン上でボクセル化を行った様子を図 1 に、経路探索を行った様子を図 2 に示す。

2.2.4 音の回折

音の回折は、音源とリスナーの直線距離と、障害物を考慮した最短経路の距離の差を求めるというアルゴリズムを実装した。この差が 0 より大きい場合、音源とリスナーの間に物体があることになり、その差分に応じてローパスフィルタのカットオフ周波数を下げることで、音の回折による高周波数の遮断を再現している。

2.2.5 音の反響

音の反響は VRChat でも多く用いられている、「Audio Reverb Zone」を用いて実装した。これは Unity の標準機能として搭載されており、一定範囲内の音源に、指定した音響特性を加えて、疑似的に反響音を再現することができる。また、音響特性はカスタマイズすることが可能で、これを調整することで、任意の材質に合わせた音の響きを再現することができる。

2.3 音声通信の実装

本研究では、WebSocket という通信プロトコルを用いて、ユーザ間の音声通信を実装した。音声通信のシステム構成を図 3 に示す。VR 空間と MR 空間それぞれのユーザはマイクを介して Unity に音声を伝達する。この時のサンプリングレートは 16000Hz である。Unity に伝達された float 型の音声配列は、byte 型に変換した後に、gzip という圧縮

形式を用いて圧縮し、サーバへ WebSocket 通信を用いて送信する。ここで、マイク音声の送信レートは 0.01s であり、この時間幅で取得した音声データを圧縮して送信している。サーバは Amazon EC2 の t2.micro インスタンスを用いている。各ユーザの音声を受け取ったサーバは、リモートユーザの音声を、VR 空間に存在するユーザと MR 空間に存在するユーザの両方に送信し、現地ユーザの音声は、VR 空間に存在するユーザにのみ送信する。

送信された音声データは、受け取ったクライアント側で、gzip の形式から byte 型のデータを介して float 型のデータに戻した後に、音声を送信したユーザのアバタ位置で再生される。音声を再生する際に、受信した音声をバッファに蓄積し、一定量の音声は蓄積されてから再生を行うことで、音声の途切れる現象を防いでいる。バッファサイズは遅延と音質のトレードオフであり、本研究では両者のバランスを考慮してそのサイズを 3000bytes に設定した。

実装した音声通信システムの動作確認として、クライアントからサーバに音声を送信し、それを他のユーザが受信するまでの遅延を計測した。参加ユーザの合計が 2 人の場合と 4 人の場合で計測した結果、平均遅延はそれぞれ 23.2ms と 20.0ms となった。ITU-T G.114 によると、高い音声品

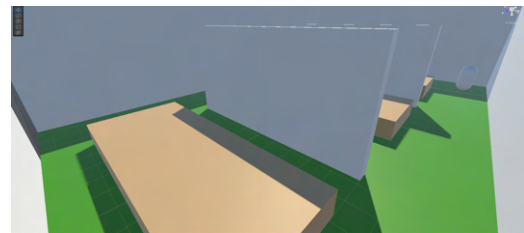


図 1: ボクセル化を行った様子

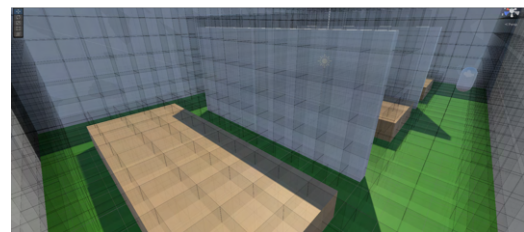


図 2: 経路探索を行った様子

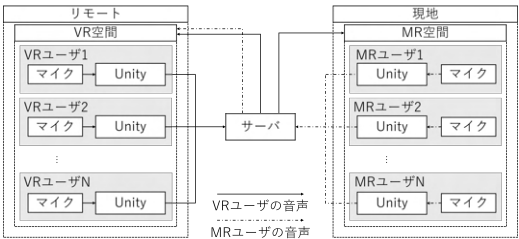


図 3: 音声通信のシステム構成図

質を実現するために必要な最大遅延として、150ms を推奨している [6]。今回の計測結果では、ネットワーク通信の部分のみを計測しているため、音声処理などを含めると遅延が増加することが予想されるが、通信速度としては十分であるといえる。

3. 評価実験

3.1 実験概要

本研究で開発した音声伝達機構の中で、特に空間音響の再現度が、MR キャンパスで使用するにあたって十分であるかの評価をすべく、評価実験を行った。本実験では、被験者に HMD を装着して VR 空間に参加してもらい、イヤホンを用いて周囲の音を聞くタスクを行わせた。具体的には、聞こえた音のボリュームや定位から、指定した音声の方向と位置を推定するタスクを行わせた。実験用のデバイスとして HMD は Meta 社の Meta Quest 3 を用い、イヤホンはオープンイヤー型の有線イヤホンである、nwm 社の MWE001 を使用した。本実験の参加者は男性 12 名で、平均年齢は 21.9 歳であった。

3.2 実験手順

本実験は以下の 6 つのステップで行った。

- 1. 準備シーンで音量調整
- 2. 練習シーンで実験の進め方を説明
- 3. 遮蔽物なし条件の実験を実施
- 4. ステップ 3 の主観的評価に回答
- 5. 遮蔽物あり条件の実験を実施
- 6. ステップ 5 の主観的評価に回答

実験で用いたシーンは図 4 に示す通りである。また、実験時の被験者と音源の位置関係の一例を図 5 に示す。音源の配置は全部で 3 種類あり、遮蔽物がある場合とない場合で、いずれも同じ音源配置を使用した。

始めに、被験者ごとに適切な音量設定が異なるため、準備シーンに入ってもらい、音量調整を行った。次に、練習シーンで実験の練習を行った。

実験は部屋の中央に立った状態から始まり、数秒後に 3 種類の音声と同時に 10 秒間再生される。この時、3 種類の音声のうち一つが指定される。再生後、指定された音声の方向を向き、コントローラーのトリガーを引くことで方向を決定する。その後、指定された音声の位置にコントロー

ラーを用いて移動し、位置を決定する。これを 1 試行として、各条件で 3 回ずつ行った。被験者は実験中、HMD とイヤホンを装着した状態で立位の姿勢を取らせた。

実験条件として、壁の有無のみが異なる「遮蔽物なし条件」と「遮蔽物あり条件」の 2 種類を設定した。これは壁の有無が音源の方向と位置の特定に影響を与えるかを調べる目的がある。実験中に使用する音声は、女性の声、男性の声、電話の着信音の 3 つであり、実際に大学のキャンパスにおいて頻繁に耳にするとと思われるものを選出した。

3.3 評価項目

本実験では、客観的評価として被験者が推定した音源の方向と位置が、実際の音源とどれだけ離れていたかという差分を用い、主観的評価として、以下の 3 つの評価項目を用いた。各評価項目に対して、5 段階のリッカート尺度で評価を行わせた。

項目 1 音のみで音源の位置を特定できたか

項目 2 複数の音源に対して識別ができたか

項目 3 音がどの程度不自然に聞こえたか

客観的評価の項目では、異なる 3 つの音声と同時に鳴っていた時、自分が聞き分けたい音の方向と位置をどれだけ正確に特定できるかを判定した。主観的評価の項目では、本システムの空間音響がユーザーの感覚と一致しているか、空間音響に不自然な点がないかを調査した。

3.4 実験結果

条件ごとの結果は表 1、表 2 に示す通りである。方向差分（絶対値）の平均は、遮蔽物なし条件で 13.70 度、遮蔽物あり条件で 17.36 度であった。位置差分については、条件による差はわずかであり、どちらも 5m 前後に収まっていた。ただし、標準偏差に着目すると遮蔽物あり条件の方が大きく、被験者間の差が大きいことが分かる。

主観的評価の結果は図 6 に示す通りである。主観的評価の結果、項目 1（位置特定）ではスコアの標準偏差が 1.0 を上回っており、音の位置に対する評価は被験者ごとの差が大きかった。項目 2（音源識別）ではどちらの条件もスコアの平均値が 4.0 を上回っており、複数音源の識別ができたと考える被験者が多かった。項目 3（不自然さ）ではどちらの条件もスコアの平均値が 2.0 を下回っており、音の聞こえ方に対して不自然さを感じる被験者は少なかった。

表 1: 実験結果（遮蔽物なし）

	方向差分 [°]	位置差分 [m]
平均値	13.70	4.56
標準偏差	10.66	3.11

表 2: 実験結果（遮蔽物あり）

	方向差分 [°]	位置差分 [m]
平均値	17.36	4.65
標準偏差	19.97	3.39

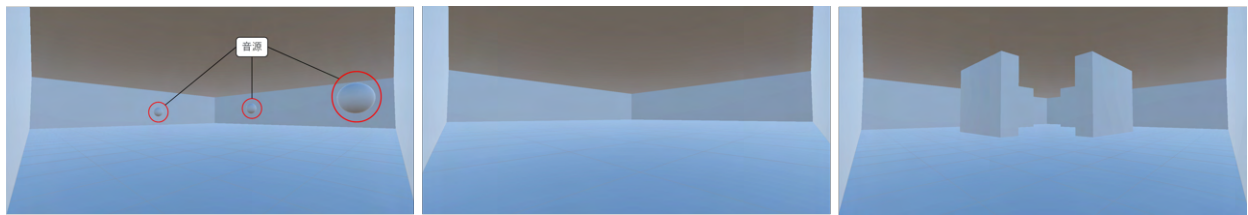


図 4: 実験シーン（左：準備シーン，中央：遮蔽物なし条件，右：遮蔽物あり条件）

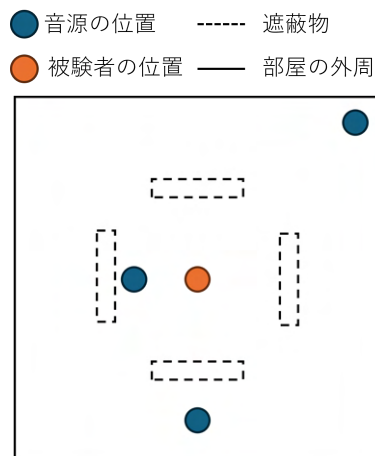


図 5: 音源の配置

3.5 考察

まず，方向差分がおよそ 20 度以内に収まっていることを確認した．中林らによると，提示された音像と映像の間隔が 20 度以上の場合，音像と映像が分離して知覚されると報告されている [7]．また，西田らによると，スピーカのような視覚情報が音源定位の特定の精度を向上させると報告されている [8]．本研究では実験のために音源の視覚情報をなくしているが，実用の際には，音声を発する人間などの視覚情報が付与されるため，方向差分がより小さくなると予想される．よって平均値に着目すると，本研究の空間音響は音源の位置や方向を特定する上で十分な性能を持っているといえる．

ここで，どちらの差分でも標準偏差が大きく，ユーザごとの個人差が激しいという問題が生じた．ユーザごとの差は，音の聞こえ方に個人差があるために生じたと考えられるが，これがユーザ体験や没入感に及ぼす影響については調査が必要である．

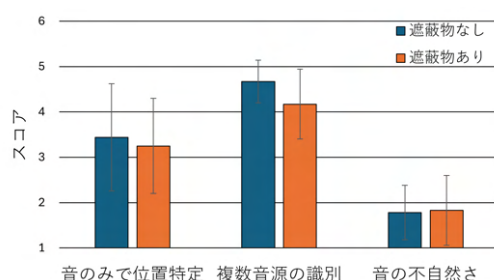


図 6: 主観的評価の結果

4. むすび

本研究では，リアルワールドメタバースの実現に向け，空間音響と音声通信の 2 つの面から音声伝達機構を開発した．開発した音声伝達機構では，音の物理的特性を簡易的に実装することで，動的なシーンに適した軽量の空間音響を再現し，WebSocket を用いることで，低遅延な音声通信を実現した．一方で，評価実験により，空間音響の聞こえ方に個人差があることが判明した．MR キャンパスでの実用に向けて，空間音響の改善や音声通信のスケーラビリティ向上を目指し，検討していく必要がある．

参考文献

- [1] John Hanke. The Metaverse is a Dystopian Nightmare. Let's Build a Better Reality. <https://nianticlabs.com/news/real-world-metaverse>, (2021-08-10).
- [2] Microsoft. Microsoft Mesh の概要 | これまでにならぬつながりを実現. <https://www.microsoft.com/ja-jp/microsoft-teams/microsoft-mesh>, (2025-01-02).
- [3] 柳澤祥太 他. リアルワールドメタバース実現に向けたクロスバース・プラットフォームの構築 (1) —3 次元マップの生成と選択的配信機構—. 第 30 回バーチャルリアリティ学会大会論文集, (2025).
- [4] GDCVault. The Number One Educational Resource for the Game Industry. <https://gdcvault.com/play/1034667/Tunes-of-the-Kingdom-Evolving>, (2025-01-07).
- [5] Peng Li, et al. VNS-BA*: An Improved Bidirectional A* Path Planning Algorithm Based on Variable Neighborhood Search. *Sensors*, Vol. 24, No. 21, (2024).
- [6] International Telecommunication Union. One-way transmission time. Recommendation G.114, Telecommunication Standardization Sector (ITU-T), 2003.
- [7] 中林克己 他. テレビ映像とステレオ音像の相互作用に関する基礎実験：テレビ・ステレオにおける音像定位. *テレビジョン学会技術報告*, Vol. 6, No. 28, pp. 55–59, 1982.
- [8] 西田鶴代 他. 音源定位における視覚情報の影響：Flmp による視覚情報の影響の定量化. *日本音響学会誌*, Vol. 55, No. 11, pp. 735–741, 1999.