第 25 回日本バーチャルリアリティ学会大会論文集（2020 年 9 月）

交流型 WebVR における空間音響のオンライン評価手法

Online evaluation method of spatial acoustic design method for network collaboration WebVR

坂口塔也 [1)]，山崎勇祐 [2)]，Bredikhina Liudmila[3)]，白井暁彦 [2)]
Toya SAKAGUCHI, Yusuke YAMAZAKI, Bredikhina Liudmila , and Akihiko SHIRAI

1) 静岡大学 情報学部（〒 432-8011 静岡県浜松市中区城北 3-5-1）
2) GREE VR Studio Lab
3) Geneva University

概要: オープンソースで実装されている WebVR コラボレーションプラットフォーム「Hubs」では、空間音響に Web Audio API の PannerNode インターフェイスが実装されている。立体音響は空間や距離感を認識する上で重要である一方、ネットワークで接続された同一空間内の不特定多数のユーザによる動的に移動可能な会話環境を考慮すると、音量減衰モデルによってはお互いに干渉し合うため、適切な会話環境の設計を行うためには各モデルの特性を明らかにしたうえでの設定が必要となる。本研究では PannerNode を利用した複数人の会話を再現する評価実験を完全オンラインで実施することに成功しており、結果を共有する。

キーワード： WebVR, Hubs, PannerNode, オンライン実験

1. はじめに

不特定多数の参加者が会話ベースのコミュニケーションを行う国際会議や展覧会をオンラインで開催するにあたり、ソーシャル VR プラットフォームが注目されており、その一つである Mozilla Hubs (以下、Hubs) が実際に国際会議等に積極的に活用され始めている [1]。

Duc Anh Le et.al が UIST2019 にて、Hubs を利用した参加者にアンケート調査を行い、Hubs が国際学会に有用であることを示している [2]。また IEEE VR 2020 では専用の Hubs 会場が作られ、口頭発表の視聴に並列し、デモやポスター発表、200 件近い交流型発表が Hubs 上で行われた（`https://hubs.ieeevr.online/`）。環境問題や新型感染症といった問題が大きくなる昨今、こうした学会や交流型イベントのオンライン化の取り組みが今後より積極的に行われるのは想像に難くない。

一方で、こうしたソーシャル VR プラットフォームに共通する課題の一つに、音声混合がある。VRChat や Hubs など、いくつかのソーシャル VR プラットフォーム上において、簡易的な空間音響モデルが適用されているものの、音場再現のような特殊なデバイスや、コストの高い演算処理が必要な手法はリアルタイム処理には向いておらず、まだ普及していない。3 次元空間内の音源はすべてヘッドホンなどの音響機器によってステレオ 2 チャンネルに集約された音として知覚される。

学会のポスター発表や懇親会など、不特定多数の話者が同一空間において無秩序に会話を行う場合、電話のような 2 者間通話モデルとは異なり、音声処理における Push-to-Talk のような話者以外のミュート処理を施すことも難しい。またアバターシステムによっては、相手の表情や頷きなど高解像度を必要とする情報を考慮することも難しい。またポスターや懇親会のような環境は、Zoom のようなビデオ会議システムによる事前に話者が知られている会議や講演者が定められた Webinar と違い、「会場の賑わい」のような未知の参加者同士の会話が聞こえてくる環境の再現も重要である。

2. 減衰のモデル

ネットワーク WebVR における複数話者の会話を快適に再現する現実的な手法として、距離や話者の向きといった減衰モデルを適切に利用し、リアルタイム処理を幅広い再生環境で再現できる必要がある。Hubs では、空間音響に Web Audio API の PannerNode インターフェイスが JavaScript ベースで実装されている。この API は W3C Audio Working Group が標準化[1]し Mozilla が API として整備している[2]。

各ブラウザ環境でサポートされている実装はそれぞれ異なる（例えば、Orientation は Internet Explorer や Safari で実装がない）ため、指向性を制御する cone 値と、距離に対する減衰モデルである distanceModel（以下、減衰モデルと呼ぶ）ならびにそれに付随するパラメータが Hubs 内において制御可能である。

減衰は（linear, inverse, exponential）の 3 種類のモデル

[1] `https://www.w3.org/TR/webaudio/`
[2] `https://developer.mozilla.org/en-US/docs/Web/API/PannerNode`

から選択することができる。inverse モデルおよび exponential モデルは逆数もしくは指数関数的曲線で減衰を行う。音源とリスナーの距離遠くなるにつれ 0 に収束するが、どれだけ離れても完全に 0 にはならないため音が聞こえ続ける。一方、linear モデルは直線的な減衰を行う。音の減衰の終了距離は明確に設定でき、またその地点での最小ゲインを設定することが可能である。反対に inverse モデルと exponential モデルは性質が近いため、本稿では Hubs でデフォルトで利用されている inverse モデルと linear モデルの特性を比較する実験の中心とし、exponential モデルについては割愛する。それぞれの減衰モデル 式 (1), 式 (2) に示し、それらに使用されている変数を以下に記す。

$$G_{Linear} = 1 - f\frac{max\Big[min(d,\ d'_{max}),\ d'_{ref}\Big]\ -\ d'_{ref}}{d'_{max}\ -\ d'_{ref}} \tag{1}$$

$$d'_{ref} = min\ (d_{ref},\ d_{max})$$
$$d'_{max}\ =\ max(d_{ref},\ d_{max})$$
ただし、$d'_{ref} \leq d \leq d'_{max}$
（$d'_{ref}\ =\ d'_{max}$の場合、上式は $1-f$ と等価となる）

$$G_{Inverse} = \frac{d_{ref}}{d_{ref} + f[max(d, d_{ref}) - d_{ref}]} \tag{2}$$
ただし、$d_{ref} \leq d < \infty$

- d = 音源とリスナー間の距離
- d_{ref} = 音源から減衰が始まるまでの距離
- d_{max} = 音源とリスナーの間の最大距離
- f = 減衰係数、値が大きいほど減衰率が上昇する。

3. 実験

3.1 手法

本実験では、linear 減衰モデルとして式 (1) に $d_{ref} = 0, d_{max} = 3, f = 1$ をそれぞれ代入した式 (3)、inverse 減衰モデルとして式 (2) に $d_{ref} = 1, f = 1$ をそれぞれ代入した式 (4) を用いた。なお、linear モデルのパラメータは実験環境をもとに仮定したものである。また、inverse 減衰モデルおよびパラメータは、Hubs においてデフォルトで適用されている設定である。

$$G_{Linear}\ =\ 1\ -\ \frac{1}{3}max\Big[min(d,3),0\Big] \tag{3}$$
$$G_{Inverse}\ = \frac{1}{1+2[max(d,1)-1]} \tag{4}$$

図 1 は Hubs ルーム上に構築した実験環境である。被験者はキーボードの A/左矢印キー、D/右矢印キーを使って左右の移動のみ行う。被験者の面には壁があり、A〜Y の 25 のアルファベットが記載された直線の画像が壁に並行に設置した。このアルファベット直線は座標を回答するための目安であり、アルファベット間幅は 0.3 m である。いずれかのアルファベット位置に音源 A と音源 B が設置してあるが、被験者からは不可視状態となっている。音源の位置は部屋ごとに変更した。 なお、2 音源間距離を 4.2 m で固定した。これは式 (3) と式 (4) において減衰率が同じになる地点が d=1.5 m と d=2.0 m であり、アルファベット選択肢となる 0.3 の倍数の中で 2.0 m に最も近い 2.1 m 地点で理論上の同一減衰量となる地点が設定できるためである。これにより、両音源の中間地点での減衰率は G_{Linear} が 0.3、$G_{Inverse}$ が 0.3125 となる。この環境において、音源 A と音源 B の組み合わせを変更し、合計 6 種類の実験ルームを用意した。実験ルームの対応表を表 1 に示す。サイン波およびホワイトノイズは音声編集ソフト「Audacity」を利用してなお、サイン波の周波数は 440 Hz、ホワイトノイズは Amplitude = 0.1 で生成した。会話の音源は「千葉大学 3 人会話コーパス」[3] を利用した。この音源は、千葉大学で収録された大学生・院生・ポスドクを含む同性 3 人からなる友人同士 12 組の雑談を収めたものである。この会話音源を利用して、WebVR コラボレーションプラットフォームで発生しうる「複数の会話が並列に進行している状況」を再現した。女性 3 人の会話データと男性 3 人の会話データを 5 分程度に切り抜き、ホイトノイズおよびサイン波と合わせて Audacity に て Loudness Normalization で正規化を行った上で、モノラル音声に加工した。

表 1: 実験ルームの設定 (Lin.=Linear, Inv.=Inverse)

Room	1	2	3	4	5	6
減衰モデル	Lin.	Inv.	Lin.	Inv.	Lin.	Inv.
音源 A	440Hz Sin 波		会話（女性）			
音源 B	ホワイトノイズ				会話（男性）	

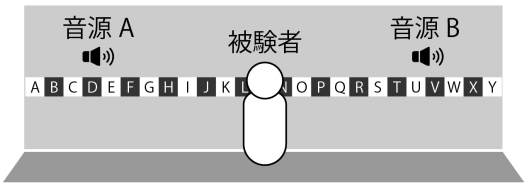

図 1: 実験環境概念図

3.2 手順

本実験は全てオンライン上で行った。被験者は、実験のガイドならびに実験ルームのリンクが記載された Google フォームのリンク (`https://bit.ly/VR20200805`) が与えられた。また事前に、該当する年齢層と性別を回答した。被験者は初めに実験手順が説明された画像が並ぶ Hubs ルームに入室した。このルームでは Hubs の操作方法を確認できる。Hubs 標準のルーム内での移動モデルには加速度があり、現実空間に比べて高速であり、調整法での実験に向いていな

い。そのため Hubs 内のユーザ設定項目である Movement speed modifier (移動速度変化量) の値を標準の 1.0 より 0.1 に変更するように指示した。これにより、実験環境内での移動速度は約 0.32 m/s となった。バーチャルな予備室にて動作確認を完了後、被験者は実験用ルーム **Room1〜6** に入室する。実験用ルームは全て定員が 1 名に設定されており、ユーザー同士の干渉を防止した。一連の実験操作はすべてオンラインで各自の協力者がヘッドフォンを装着し、静音環境において時間制限なく実施することを要請している。音源は入室すると自動で再生される。設問はそれぞれ、**Q1**:音源 A の位置、**Q2**:音源 A が聞こえなくなる境界の位置、**Q3**:音源 B の位置、**Q4**:音源 B が聞こえなくなる境界の位置、**Q5**:音源 A と音源 B が同程度に混ざって聞こえた位置とした。以上 5 点について、アルファベットで回答する。なお、**Q2** と **Q4** については、境界が無いと判断した場合は「無し」と回答できる。実験順によるバイアスを避けるため、**Room 2** での終了後被験者に {クマさん, ウサギさん} のいずれかの実験グループを選択 (もしくは指示により同数となるよう) 実施した。**クマさん** を選択した場合は Linear → Inverse の順、すなわち **Room 3, 4, 5, 6** まで順に入室し、**ウサギさん** を選択した場合は Inverse → Linear の順、すなわち **Room 4, 3, 6, 5** の順に入室して実験を行うルートが設定される。

3.3 結果

24 時間の実施期間において参加した実験協力者は男性 12 名 (10 代: 3 名, 20 代: 7 名, 40 代: 1 名, 50 代: 1 名) 、女性 2 名 (30 代: 2 名) の計 14 名であった。ウサギ (以下 **U**) とクマ (以下 **K**) の比率は 1:1 (各 7 名) であった。実験結果のうち、減衰モデルが linear であるものを図 2、 inverse であるものを図 3 に示す。各設問において、回答が次の条件を満たしたときに正答とし、その正答率を表 2 に示した。

・**Q1** および **Q3**: 回答が音源 A, B の位置と一致した場合（例: Room1 の場合は Q1=D, Q2=R)。

・**Q2** および **Q4**: 音源 B からの音量が 0 になった位置と合致した場合（例: Room1 の場合は Q2=N, Q4=H)。

・**Q5**: 両音源からの音量が理論上等しくなる位置（例: Room1 の場合は Q5=K)。

また実験全体の正答率と実験の順番を入れ替えた場合 (K: linear → inverse, U: inverse → linear) の正答率の差異を表 2 に示した。

表 2 から、linear 減衰モデルの場合、サイン波とホワイトノイズに関しては音源位置特定の正解率が高いが（**Room1** の **Q1,Q3**、**Room3** の **Q3**)、会話音源（**Room3** の **Q1**, **Room5** の **Q1, Q3**）では正解率が低かった。inverse 減衰モデルの場合、ホワイトノイズの音源位置特定に関する正答率はいずれも 80%近くあるが、その他音源ではいずれも正答率が 50%を下回っていた。このモデルでは音量が 0.1 より小さくなることは無いが、20 〜 40%程度の参加者が「ある点で音量が 0 になった」と誤答している。いずれの減衰モデルの場合についても、**Q5** の正答率は全て 50%を下回っており、両音源からの音量が等しくなる位置を検出できていなかったことがわかる。正答以外の回答に着目すると、図 2, 3 の Q5 の回答結果が、両音源の音量が理論上同等となる点 (紫) よりも、**Room 1** ではホワイトノイズ、**Room 3, 4, 5, 6** では会話 (女性) に近い位置で回答される傾向がある。また実験の順番に関して、表 2 下から、**Room5** の **Q2** を除き、K の方が U よりも正答率が同等か大きい。

4. 考察

音源の位置の特定 (Q1,Q3) において inverse モデルで正答率が低い理由は、減衰開始位置まで 1.0 m 離れていることによって、$G_{Inverse}$が最大空間が広いためと考える。両音源の中央の特定 ($Q5$) において *inverse* モデルで正答率が低い理由は、音源から離れるほど減衰がなだらかになり、距離による音量の差が少ないためと考える。よって *inverse* 音源では、遠距離における定位の難易度が上がると言える。会話の定位精度が低い点について、実験協力者からのコメントとして「実験で利用した女性の会話は抑揚や声量の変化が大きく定位しづらい」という意見がみられた。会話においては抑揚や話者毎の声量の差などが要因となり、完全なノーマライズが難しいことが背景にある。改善方法として話者毎の音声が単一のチャンネルに含まれている状態でノーマライズ処理を行うことが考えられるがリアルタイム処理は難しいだろう。サイン波およびホワイトノイズの音位置の定位精度は高く、会話では低い傾向にあった。特にホワイトノイズについては、位置関係が変わる際に聞こえる音の周波数が変わり、音量以外の情報を得ていると考える。順序バイアス差は、K において *linear* と *inverse* モデルを交互に実験するため、その法則性に気づいて正答率が上がったことが考えられる。本実験では *linear* モデルの減衰開始距

表 2: 上表:実験全体の正答率 (%)(n = 14) ／下表:順序バイアス考慮。K の正答率 (%) - 正答率 (%) (各 n=7)。

	Q1	Q2	Q3	Q4	Q5
Room1	78.6	64.3	78.6	64.3	35.7
Room2	42.9	78.6	78.6	64.3	28.6
Room3	50.0	57.1	100.0	64.3	28.6
Room4	14.3	92.9	78.6	57.1	7.1
Room5	57.1	50.0	28.6	57.1	14.3
Room6	35.7	71.4	35.7	57.1	14.3

	Q1	Q2	Q3	Q4	Q5
Room3	21.4	0.0	0.0	21.4	0.0
Room4	0.0	7.1	7.1	0.0	7.1
Room5	28.6	-7.1	0.0	14.3	0.0
Room6	21.4	14.3	21.4	28.6	0.0

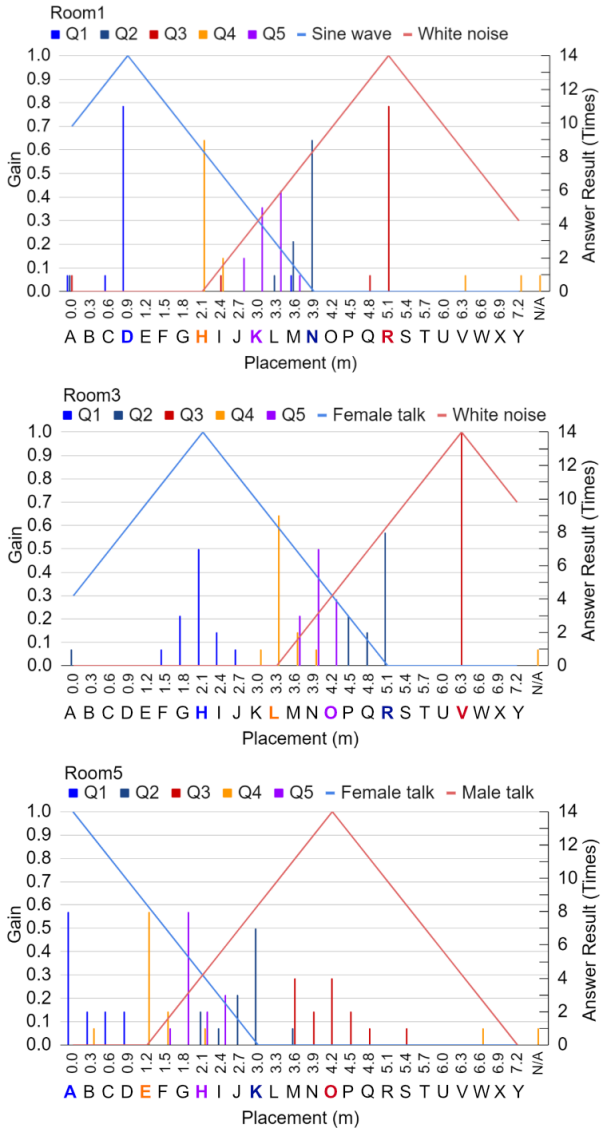

図 2: linear での結果: 横軸は図 1 における A を 0 としたときの距離 (m) と回答位置、配色は各問での正答を表す。404 は「無し」の回答数。縦軸は各モデルの Gain 理論値 (折線) および回答数 (バー)。

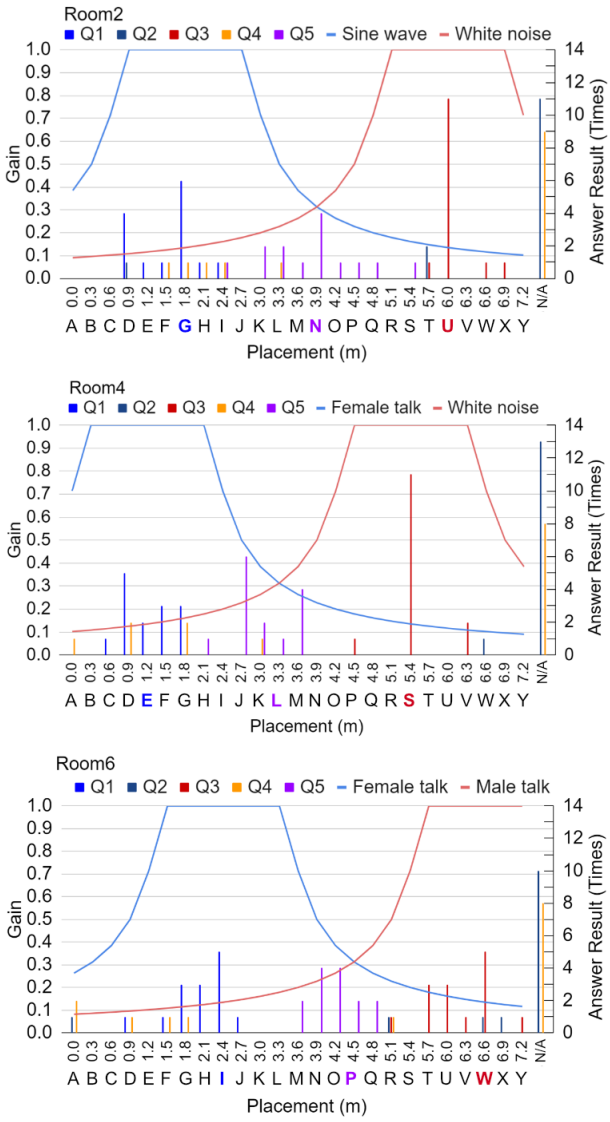

図 3: inverse での結果 (描画は図 2 同様)

離を 0m としているが、両モデルの定位や特性をより対照的に比較するのであれば、減衰開始距離を合わせて同実験を行うことが妥当であろう。なお今回の実験では *linear* 減衰の傾きを 1.0 としているが、これを 1.0 以下の数値にすることで、どれだけ離れても $Gain_{Linear}$ が 0 にならない音源を生成可能である。これらを考慮して追加実験を行うことで、より適切な設定を求めることができると考える。

5. 結論

減衰開始距離を短くすることで視覚情報や音声の指向性がなくても音源の定位が容易になる。ただし、ある音源の定位が容易になることは、他の音源が聞こえにくくなることを意味する。よって、賑わいを表現したい場合や、多くの人物の音声を混在させる必要がある場合にも有益な知見を得た。inverse モデルを利用した場合、音の減衰の強さが距離によって異なるため、音源との距離感を音声から掴むのが難しくなる。よって音源が多い環境、つまり同時参加者が多いイベント等においては linear モデルが適している。

参考文献

[1] 加藤史洋他. 小特集：学会オンライン化・VR 開催の幕開け. 日本バーチャルリアリティー学会誌, pp. 35–43, 6 2020.

[2] Duc Anh Le, Blair MacIntyre, and Jessica Outlaw. Enhancing the experience of virtual conferences in social virtual environments. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 485–494. IEEE, 2020.

[3] Yasuharu Den and Mika Enomoto. *A Scientific Approach to Conversational Informatics: Description, Analysis, and Modeling of Human Conversation*, chapter 17, pp. 305–330. John Wiley & Sons, Ltd, 2007.