Voice Arrow: 声の可視化によるインタフェースツールの提案

田中聖也[†1]　外村佳伸[†2, a)]

概要：本論文では，仮想環境において声を可視化し，その声の出す向きや出し方でさまざまな操作が可能になるインタフェースツール Voice Arrow を提案する．本システムでは体の向き，声の大きさ・高さ，手のジェスチャーによる操作が可能であり，各操作と対応するインタフェース機能の有効性を調べるために，初期的なシステムを実装し，評価実験を行った結果について報告する．

1. はじめに

近年のデジタル化の進展によって，今では身の回りにデジタルな環境が当たり前にあり，仮想的な環境も活用されるようになってきた．加えて AR（拡張現実）やプロジェクションマッピングといった技術により，仮想空間を現実世界に作り出すなど，今後より一層現実と仮想の融合が進むと考えられる．

こうした環境におけるユーザーインタフェースにはさまざまなものがある．例えば特別なデバイスを装着して仮想世界の中でボタンやタッチパネルを操作したり，ジェスチャーによる操作を行うものなどがある．しかし今後デジタルな環境がより生活に身近なものとなり，現実と仮想の 3 次元融合的な広がりを持つことを考えると，我々は特別なデバイス装着の必要がなく，しかも自然な操作感で利用できるインタフェースに将来性を感じている．仮想空間で利用するインタフェースとして，従来主に手や指を用いて仮想ポインティングすることが多いことに対し，本研究では声を可視化してインタフェースの基本ツールとして使えることをめざしている．本報告では，その基礎的な検討として，操作体系および基本インタフェースツール Voice Arrow を実装例とともに提案する．

Voice Arrow では特別なデバイスの装着なしに，利用者が体の動作を伴いながら，声を使って直感的に操作を行うことができるものである．

2. 仮想空間の声と操作インタフェース

2.1 声のインタフェース

近年では音声認識を用いたインタフェースの研究や，「音声アシスタント」と呼ばれる音声言語に基づくインタフェースが広く使われている．これに対し本システムでは，声そのものを音声信号レベルで扱い，3 次元仮想空間において可視化することを基本とし，声そのものを操作インタフェースとして，声を出す方向や出し方で仮想空間内の対象に対して操作が行えることをめざした．

2.2 仮想空間における音の可視化

仮想空間における音の可視化にはさまざまな手法がある．例えば，二次元では理解が難しい音源の周囲の音の強度など空間的なサウンド情報を矢印で可視化する研究[1]や，聴覚に障害を持つ人に対し，音源の位置や特徴，種類を視覚化する研究[2]などが行われている．このように音の用途に応じてさまざまな可視化は行われているが，一般的な用途向けの手段としては適用しにくい．本システムでは声をインタフェースツールとして用いるために，声を操作オブジェクト化し，声の特徴や指向性及び伝搬を可視化することをめざした．

3. Voice Arrow

本章では Voice Arrow を具体的に提案する．

3.1 基本概念

本システムでは，利用者は自身の声を仮想環境における新しいユーザーインタフェースとして利用する．システムでは 3 次元 CG による仮想空間を作り，その中に利用者が位置づけられる．利用者は後述するような仮想空間を見せる大画面の前や，将来的には透過型の AR 用メガネを用いてもよいが，そこで利用者が仮想空間内の操作対象に向かって声を発すると，その声は指向性をもって対象に向かって進行するように可視化される．このとき，人の直観的な操作行動で使えるために，人が何かに向かって声をだす際に，図 1 のように両手を口の正面に添えるような行動をとることを念頭においた．つまり人が遠くに声を届ける際に手をメガホンのように使う姿勢である．ただし，本システムにおける声の可視化は音響的に厳密にシミュレーションするものではなく，あくまで利用者の声による操作を可能にするためのものであり，ユーザーインタフェースにおける GUI（Graphical User Interface）に該当するものである．

†1 龍谷大学大学院 理工学研究科 情報メディア学専攻
†2 龍谷大学 先端理工学部 知能情報メディア課程
a) tonomura@rins.ryukoku.ac.jp

本システムの利用者は，空間内の対象オブジェクトに対し，発声の向きや声の出し方，手のジェスチャーを調整しながら声を発することで，そのオブジェクトに対し様々な操作を行う．

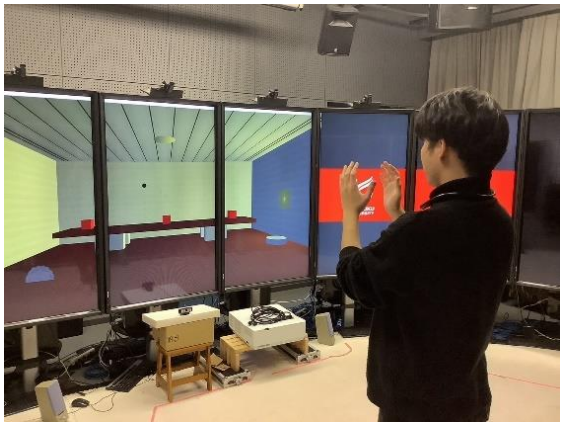

図 1　対象に向かって発声するしぐさ

3.2 基本操作とインタフェースとしてしての機能

利用者が声により行うことのできる基本操作と，対応するインタフェースとしての機能を表 1 に示す

表 1　基本操作と対応インタフェース機能

発声の基本操作	インタフェース機能
1. 声を出す	作用を起こす
2. 声の向きを変える	対象への方向・位置の指定
3. 声の大きさを変える	強弱の変更
4. 声の高さを変える	パターンの変更
5. 両手の間隔を変える	影響範囲の変更

操作 1.は，対象オブジェクトに対して声を出すことによって作用を起こすもので，本システムのあらゆる動作の出発点となる．操作 2.では体の向きを変えるなどにより声を出す方向や位置を指定することができ，この操作によって特定のオブジェクトに対して操作を行うことが可能になる．操作 3.では声の大きさ（音の振幅）によって音声オブジェクトの持つパラメータの大小を操作することができる．操作 4.では声の高さ（音の周波数）を変えることで，操作オブジェクトのもつパラメータの種類やパターンを変更することができる．扱える声の周波数は，人の話し声の平均的な周波数を参考に，男性は下限を 100Hz，上限を 600Hz，女性は下限を 200Hz，上限を 1100Hz としている．操作 5.では，利用者が口の正面に添えている両手の間隔により声の拡散範囲を制御することができ，声の影響範囲を調整することが可能になる．

3.3 声の可視化

利用者の声による操作を補助するための GUI として図 2 のような音声オブジェクトの可視化を行った．図のように声は視点方向直線上に進行する球と，その周りを円状に囲み放射状に拡散しながら進行する球集合が同時に進行するように可視化される．このとき，球集合の拡散範囲が表 1 の操作 5.に対応している．また中心の球の大きさ及び色が操作 3.と操作 4.にそれぞれ対応している．球の大きさに関しては操作 3.の声の振幅が大きいほど球も大きくなる．球の色に関しては操作 4.の声の周波数によって変わり，これは色相環に対応させている．周波数が低いほど青に，高いほど赤に近い色になる．また操作 2.に対応して画面上に黒点で照準が表示されており，この照準の正面方向に声が可視化される．

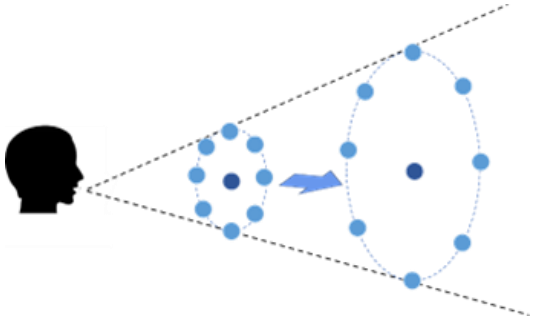

図 2　音声オブジェクトの可視化

3.4 プロトタイプシステム

仮想空間は日常的な場面での利用を想定し，部屋をイメージした空間を作成した．またそこに図 3 のように声による操作が可能な，家具をイメージしたオブジェクトを設置した．オブジェクトは「Cleaner Object（C オブジェクト）」，「Lighting Object（L オブジェクト）」，「Statue Object（S オブジェクト）」の計 3 種類ある．

C オブジェクトは掃除機ロボットをイメージしたものであり，床に対して声を出すことで指定した場所に向かって C オブジェクトが動き出し，掃除を行ってくれる．

L オブジェクトは照明をイメージしたものであり，声の大きさと高さを使って，照明の明るさと色を操作することができる．照明の明るさは声の大きさに応じて 5 段階の制御が可能であり，振幅が大きいほど明るくなる．照明の色は，声の周波数によって変わり，色は声の可視化の色に対応しており計 6 種類である．

S オブジェクトは影像をイメージしたオブジェクトが机の上に計 3 つ置かれている．各オブジェクトは声の大きさと高さを使って，オブジェクトのサイズと形を操作することができる．オブジェクトのサイズは声の大きさに応じて 5 段階の制御が可能であり，振幅が大きいほどサイズも大きくなる．オブジェクトの形は声の周波数によって変わり，低い周波数の場合は球，中程度の周波数の場合は円柱，高い周波数の場合は正方形で計 3 種類である．

実際に各オブジェクトに操作を行った一例を図 4 に示す．

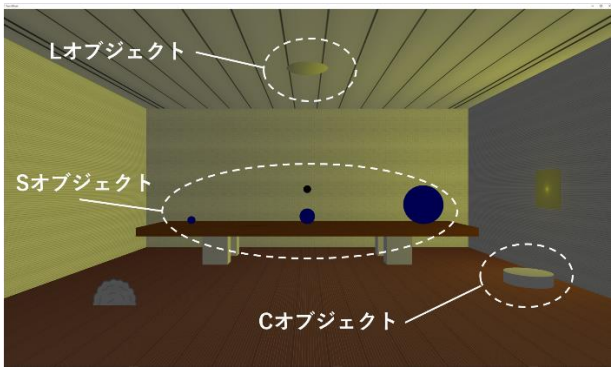

図 3　仮想空間内に配置される各オブジェクト

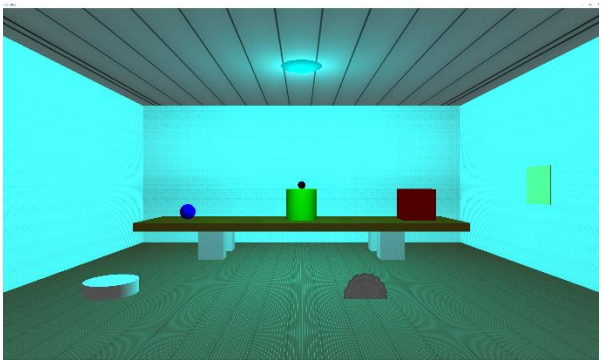

図 4　各オブジェクトに操作を行った一例

また実装したプロトタイプシステムでは,アンビエント・ウォールと呼んでいる 46 インチ液晶 7 台が連なった大画面ディスプレイの内の 3 台分の画面を用い，そこに作成した仮想空間を表示し，その前に利用者の身体動作を検知するためのカメラを設置した．利用者はその前に立ち，装着したマイクを通して声を発する（入力する）．システムの概要を図 5 に示す．

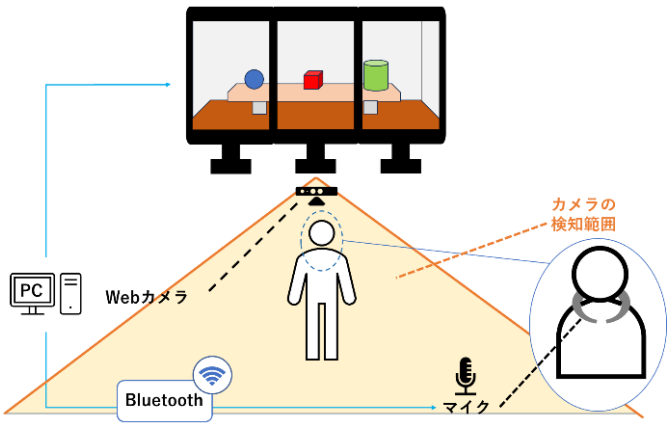

図 5　システム概要

3.5　ソフトウェア処理

本システムでは大きく身体動作の処理，仮想空間の描画処理，音声処理の 3 つのモジュールからなる．ソフトウェア処理の概要を図 6 に示す．

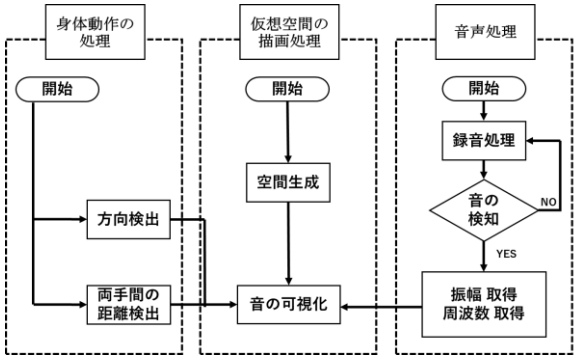

図 6　ソフトウェア処理概要

3.5.1 身体動作の検出

身体動作の検出には MediaPipe を用いた．MediaPipe は Google 社が提供するハンドトラッキングやポーズ検出，顔検出などの画像認識を行うことが可能なオープンソースのフレームワークである．本システムでは計 32 点からなる骨格データ（図 7）を基にポーズ検出を行い，利用者の正面方向の検出と両手間の距離の検出を行った．また MediaPipe から得られる骨格データの画像の一例を図 8 に示す．

両手間の距離に関しては，両手首の X 座標の差を参照した．利用者の正面方向の検出に関しては，上下方向の検出と左右方向の検出を別々に行い，各方向で検出した角度を組み合わせたものを利用者の正面方向とした．上下方向の角度に関しては利用者の基本姿勢を 3.1 の図 1 のような姿勢とすることから，口と手の Y，Z 座標を繋いだときにできる直線の角度を参照した．また，左右方向の角度に関しては上下方向同様に，口と手の X，Z 座標を繋いだときにできる直線の角度を参照する方法（手法 1）に加え，両肩の X，Z 座標を繋いだときにできる直線の角度を参照する方法（手法 2）の計 2 手法を用意し手法を用意し，左右方向の検出にどちらの手法が適しているか予備実験で調べた．

予備実験では上下左右中央の各方向に設置したオブジェクトを見ているときの x，y 座標系における手と口の相対位置を調べた．照準を表示せず行ったときを自然な姿勢とし，各手法において照準を表示した状態で再度計測し，どちらが自然な姿勢に近いかを調べた．実験には 20 代の男性 4 人に協力してもらい，実験終了後にアンケートを行った．計測した結果の平均を取りまとめたものを図 9 に示す．各点の位置が原点を口の位置としたときの相対位置であり，黒点が自然な姿勢の場合，赤点が手法 1 の場合，青点が手法 2 の場合である．

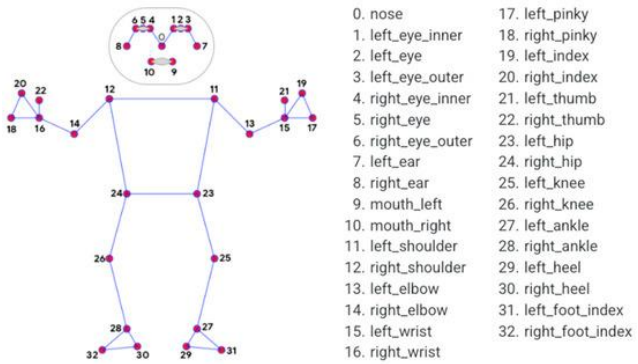

図 7　MediaPipe Pose による骨格ノード割り当て

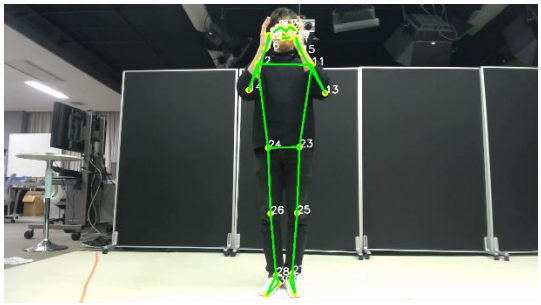

図 8　MediaPipe から得られる骨格データの画像

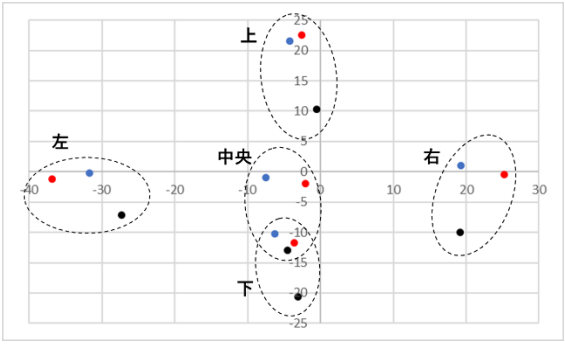

図 9　各方向における口と手の相対座標

各方向での座標の違いを比べると，手法 1，2 ともに y 軸方向には上に同程度ずれていることがわかる．しかし，x 軸方向を見ると，手法 2 である青点の方が左右の方向を見た時のずれが少ないことがわかる．また行ったアンケートでは，どちらの手法が照準を合わせやすいか聞いたところ 4 人中 3 人が手法 2 のほうが良いと答えた．よって本システムでは左右方向の角度の検出に肩の座標を用いた手法 2 を用いるものとする．

3.5.2 仮想空間の描画処理

3 次元 CG の仮想空間は JavaFX を用いて構築した．音声の入力を検知すると，音声処理で得た振幅，周波数，身体動作の処理で取得した方向に応じて音オブジェクトの集合が飛んでいくアニメーションを行う．

3.5.3 音声処理

音声処理では声の大きさの取得と声の高さの取得を行っている．声の大きさは 0.3 秒ごとに利用者の声を録音し，その音声データから振幅の平均を取得している．また声の高さは精度の観点から 1 秒ごとに利用者の声を録音し，その音声データから主要な周波数成分を抽出した．大まかな手順としては，音声データに FFT（高速フーリエ変換）を行い，振幅スペクトルを取得する．その後，振幅スペクトルからも音も強い成分の周波数を取得した．

4.　評価実験

実装したプロトタイプシステムを用いて，今回用意した体や声を使ったユーザーインタフェースの各操作機能を自然に行うことができるかを検証するため評価実験を行った．

本実験は 20 代の男性 7 名，女性 2 名の計 9 名に対して以下の通りに行った．

① 被験者に本システムでの基本姿勢と仮想空間上にあるオブジェクトに向かって声を出すと何かが起こるとだけ伝え，システムを 1 分間体験してもらった．その後アンケートを行い「声の向き」,「声の大きさ」,「声の高さ」,「両手の間隔」がそれぞれ，操作やオブジェクトにどのような影響を与えているかを予想してもらった．アンケート完了後，各操作の使い方と機能について説明した．

② 各オブジェクトを用いた実験を行った．それぞれオブジェクトの詳細と用いる機能を伝え，何度か体験してもらった後，タスクを行ってもらった．

I. C オブジェクトを用いた実験
仮想空間の床に落ちているごみのオブジェクトに向かって声を発し C オブジェクトに掃除をさせるといった操作を 1 分間行ってもらった．
実験終了後，アンケートを行った．

II. L オブジェクトを用いた実験
色と明るさを指定し，1 分間以内に指定した状態の照明に調整できるかを試してもらった．ただし明るさに関しては被験者が細かな違いを直感的に理解しにくいため「明るい」「暗い」といった程度の指示を行った．
実験終了後，アンケートを行った．

III. S オブジェクトを用いた実験．
すべてのオブジェクトを同じ「大きさ」「形」に調整するタスクと，すべて違う「大きさ」「形」に調整するタスクの 2 つを行ってもらった．それぞれ制限時間は 1 分間とした．
実験終了後，アンケートを行った．

③ 全てのオブジェクトや各操作機能を理解してもらったうえで，3 分間自由にシステムを体験してもらった．

実験①において各操作の機能に気づいた人の人数を表 2 に，実験②，③のアンケートの結果をそれぞれ以下図 10～14 に示す．評価方法は 5 段階評価で行い，質問内容通りの操作ができたと思っている場合は 5 に近い数字，できていないと思った場合は 1 に近い数字を選んでもらった．

表 2　操作と機能の関連性に気づいた人数

発声の基本操作	気づいた人数（最大 9 人）
声の向きを変える	9 人
声の大きさを変える	4 人
声の高さを変える	5 人
両手の間隔を変える	8 人

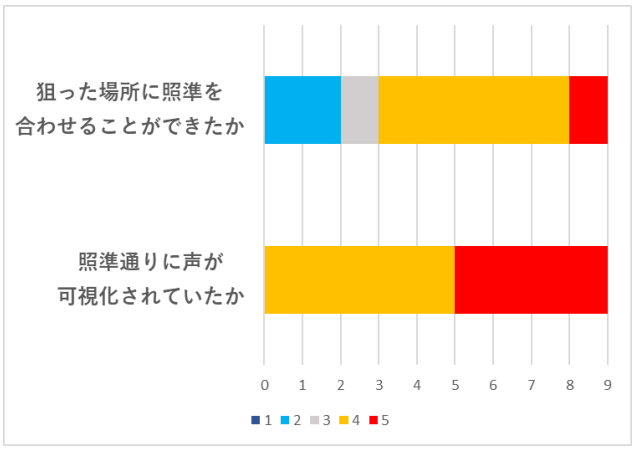

図 10　C オブジェクトを用いた
実験のアンケート結果

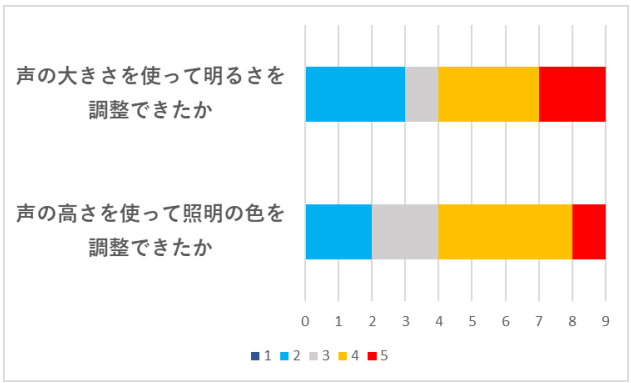

図 11　L オブジェクトを用いた
実験のアンケート結果

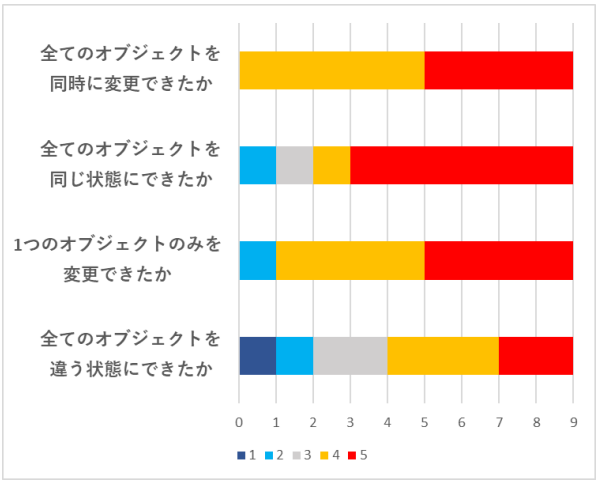

図 12　S オブジェクトを用いた
実験のアンケート結果 1

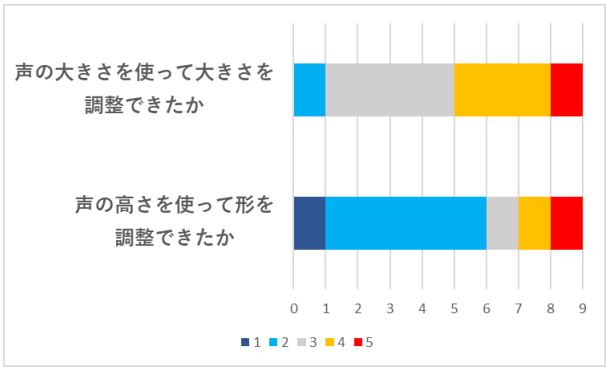

図 13　S オブジェクトを用いた
実験のアンケート結果 2

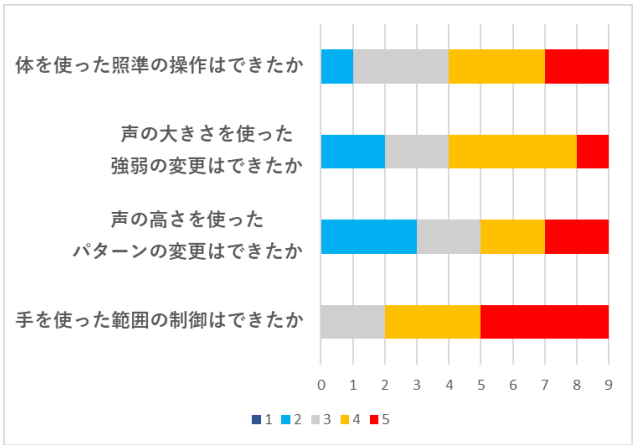

図 14　全体評価のアンケート結果

5. 考察

本章では 4 章の実験結果をふまえて，各発声の操作とそのインタフェース機能の有効性について考察を行う．

5.1 声の向きによる操作と機能の評価

声の向きによる操作は全体的に高評価であったといえる．この操作は表 2 から利用者がこの操作を直感的に行えていることがわかり，また図 10 からも照準をもとにねらった方向・位置に声を出せていることから，方向・位置決定の機能として有効であることがわかる．

5.2 声の大きさと高さによる操作と機能の評価

まず利用者が直感的に操作を行えたかについては，表 2 から半数の人は声の大きさと高さによる操作とその機能に関して気づけていないことがわかる．主な原因として，大きさや色が変わっていることには気が付いているが，それらが声の大きさと高さのどちらに対応しているのかわかりづらいことが挙げられる．

また両操作の精度としては図 11，図 13，図 14 の各該当項目を見比べると声の大きさのほうが高評価であることがわかる．やはり声の大きさは直感的に変更できるため，操作しやすく実験のタスクにおいても狙い通り行えたと考えられる．対して声の高さは図 14 や各実験のタスクの結果から意識的にパターンを変化させることはある程度可能であることがわかるが，特定の狙ったパターンに調整することが難しいということがわかった．これはシステム上 100Hz 以下などの低い周波数が認識しづらかったことや，対応する周波数の範囲を男女それぞれで固定していたため，人によってはその範囲があっていなかったものと思われる．そのため男女だけでなく，利用者ごとに周波数のキャリブレーションを行う機能を実装すべきであると考える．

また全体を通して声の大きさと高さを同時に調整することが困難であることや，現在の声の大きさと高さがどの程度かわからないといった被験者の意見から，インジケーターのようなもの画面上に表示する必要があると考える．

5.3 両手の間隔による操作と機能の評価

両手の間隔による操作に関しては表 2 からも直感的な操作であることがわかる．また図 12 や図 14 の該当項目からも範囲の調整を活かした操作が行えていることがわかる．しかし，被験者の意見としてどの範囲に飛ぶのか声を出すまでわからないという意見もあったため，照準の周りに現在の影響範囲を示すような表示を行う機能について検討したいと考える．

S オブジェクトを用いた実験では，図 12 より，全部を一度に変えたり，一つだけをねらって変えることがかなりできていることがわかる．これは，本インタフェース手法により範囲の調整が柔軟かつ連続的にできることを示しており，他の手法にはない大きな特徴と言える．

6. おわりに

本論文では，仮想環境における新たなインタフェースとして，声を用いた操作が可能な Voice Arrow を提案し，声の操作とそれに応じたインタフェース機能を評価するため，プロトタイプシステムを実装し被験者評価実験を行った．その結果，声の向きや両手の間隔を用いた操作は直感的かつ精度も高く，インタフェース機能として有効であることがわかった．しかし，声の大きさと高さによる操作は現状では利用者によって評価が大きく変わるためインタフェース機能としては課題が多くあることがわかった．そのため今後は各パラメータのキャリブレーション機能や声の大きさや高さのインジケーターを実装することで，利用者の差に関係なく直感的で容易に操作が可能なシステムの構築をめざしていきたい．

謝辞 被験者実験への参加者全員に謹んで感謝の意を表する．

参考文献

[1] Atsuto Inoue，Kohei Yatabe，Yasuhiro Oikawa，Yusuke Ikeda，Visualization of 3D sound field using see-through head mounted display，
ACM SIGGRAPH 2017，No．34， pp 1–2， 2017．

[2] Ziming Li，Shannon Connell，Wendy Dannels，Roshan Peiris，SoundVizVR: Sound Indicators for Accessible Sounds in Virtual Reality for Deaf or Hard-of-Hearing Users，
ASSETS '22: Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility，No．5， pp 1–13， 2022．