人工知能学会研究会資料
JSAI Technical Report
SIG-Challenge-B402-05

臨場感の伝わる遠隔操作システムのデザイン：
マイクロフォンアレイ処理を用いた音環境の再構築
Design of tele-operation system for tele-presence: Recreating auditory scene using microphone arrays

劉超然[1]，カルロス石井[1]，石黒浩[2]，萩田紀博[1]
Chaoran LIU, Carlos ISHI, Hiroshi ISHIGURO, Norihiro HAGITA
国際電気通信基礎技術研究所
[1]知能ロボティックス研究所
[2]石黒特別研究所
[1]ATR/IRC
[2]ATR/HIL
chaoran.liu@irl.sys.es.osaka-u.ac.jp, carlos@atr.jp, ishiguro@sys.es.osaka-u.ac.jp, hagita@atr.jp

Abstract

コミュニケーションロボット遠隔操作システムにおいて，ロボット側の空間的音環境を操作者に再現することは，臨場感の伝達に大きな役割を担う。本稿では，ロボット周囲の音源位置情報に基づいて，３次元音環境を操作することのできる遠隔操作システムを提案した。ロボット側では、音源定位・分離において複数のマイクロフォンアレイとヒューマントラッキング技術を用いた。操作者側では，操作者の頭部回転をトラッキングし、操作者の動きを補正してロボット側の空間的音環境を再生する。提案システムを用いることによって，従来法よりも高い定位精度と強い臨場感・聞き取り安さが得られることを被験者実験により確認した。また，バーチャル音環境を操作するために，２種類のユーザインタフェースも提案し、検証した。

1 はじめに

近年，ロボット遠隔操作システムにおいて，操作者の存在感をロボット側に伝達する研究が広く行われている。しかし，操作者側へ遠隔地の臨場感を伝達することに注目した研究は少ない[Nishio 2007] [Ishi 2010] [Liu 2012] [Sumioka 2014]。対面コミュニケーションに比べて，遠隔地にいる人物がロボットを介して人とコミュニケーションする場合，空間情報などの欠落によって相手との共有情報が不足する。そのため，操作者側ではコミュニケーションが行われている現場の臨場感を感じることが困難である。

臨場感の伝達に大きな手助けとなるのは，バーチャルリアリティ技術である。現在では多くの遠隔医療・軍事・コミュニケーション目的のアプリケーションなどにおいてバーチャルリアリティ技術が利用されているが[Popescu 2000] [Piron 2009] [Billinghurst 2002] [Ogi 2001]，臨場感の伝達はこれらの一つの大きな目的となっている。しかし，これらバーチャルリアリティに関する研究の大部分は，視覚における臨場感伝達に着目している[Ogi 2001] [Bullinger 1997]。音環境の構築に関するバーチャルリアリティの研究は，ゲームなどのアプリケーションで用いられているものの，未だ少ないのが現状である。リッチな音環境の構築は，遠隔操作ロボットなどのソーシャルメディアにおいても，操作者に遠隔地での自身の存在感や現場の臨場感を伝えるために重要である。

以上の背景から，本研究は遠隔地にあるロボット周囲に分布している複数の音源から構成される音環境（3D 音場）を，操作者（オペレーター）側に再現・加工することで，音の臨場感を伝達する遠隔操作システムの開発を目的とする。提案システムはリアルタイム性を保ちながら，空間的に分布する複数の音源を定位・分離し，正確な位置に再生する能力を備えることが求められる。

3D 音場を再現するため従来広く使われた方法は，バイノーラル（両耳）レコーディングされた音声をステレオで再生することである。この方法は簡便であるという利点があるが，正確なステレオマイクロフォンのセッティングが必要で，尚且つダミーヘッドが動かないためダイナミックに音場を再現することができない。さらに，各音源に対して加工を加えることも不可能である。

サラウンドチャンネルスピーカーは空間的な音場の再現のために開発されており，DirAC（Directional Audio Coding）を用いた音場再現の研究は少なくない[Pulkki 2007] [Laitinen 2011]。だが，サラウンドスピーカーシステムには二つの問題点がある。一つ目は，音場を録音した環境とそれを再生する環境が異なる場合，部屋の大きさや形状などの環境的要素が音響の伝達に影響を与えてしまい，これらの影響を正確に補正することは困難であるという点である。二つ目は，サラウンドスピーカーシステムでは “sweet spot” の位置がシステムの中心付近に限られている[Rumsey 2001]，という点である。即ち，聴者の場所が制限される。

ヘッドフォンを用いた 3D 音場の再現も，これまで広く研究されてきた。日常，人は両耳に到達した音

波の違いによって音源定位を行っている[Meyer 1972]。この違いを再現することで，ステレオヘッドフォンで 3D 音場を合成することが可能になる。頭部伝達関数（HRTF: Head Relative Transfer Function）は空間内の音源から発した音波が人の両耳に到達する時点の違いを表現する関数であって，3D 音場のバイナル再現に多く使われている[Cheng 2001]。しかし，ヘッドフォンを使って空間上に存在する音源を再現する際，バーチャルな音源が聴者の頭部・体の動きと共に動いてしまうという問題点がある。人の日常経験を考えると，外部音源の位置は聴者の体の動きに関連せず，固定されている。ヘッドフォンによる 3D 音場の再現ではこの経験と異なるため，臨場感の伝達にマイナスに働き，不自然な印象の原因となる。さらに，頭部伝達関数を使った場合，前後の誤判断が起こるという問題がある。これは，前方にある音源が後方にあるように聞こえる，もしくはその逆の現象である。日常生活では音源を定位するために意識的・無意識的に頭部を回し，その効果を定位の補助に用いている。また，頭部を回転することで前後の誤判断率が有意に下がったことも報告されている[Iwaya 2003]。

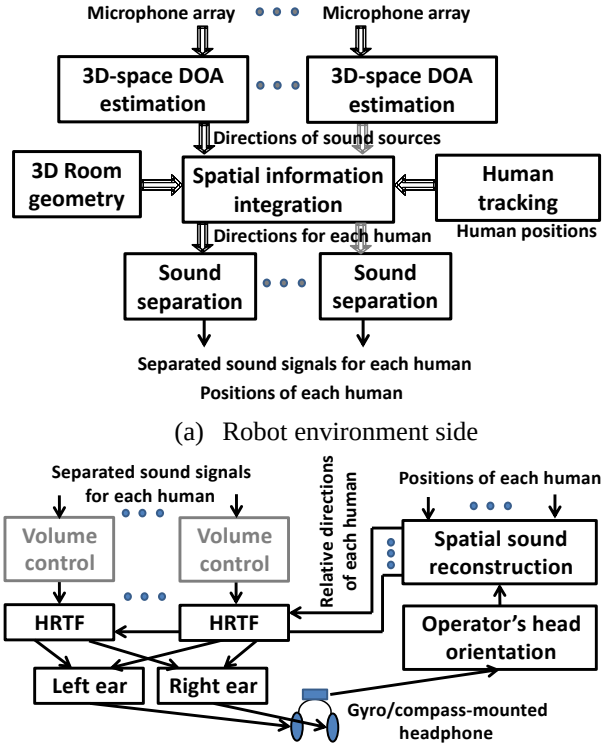
一方で，環境内の音源の空間的特性を保持するために多く使われているのは，マイクロフォンアレイ処理技術である。マイクロフォンアレイを用いた遠隔会議の研究では，音源定位や音源分離，雑音抑圧が応用されているが，多くの場合は分離音をモノラルで再生し、音場を再現している訳ではない。

これらを考慮し，提案システムではオペレーターの頭部回転をトラッキングすることで，頭部の向きに合わせた HRTF を用いてステレオ音声を合成した。正確な HRTF を選択するのに必要な連続的音源位置情報は，複数のマイクロフォンアレイの DOA（Direction Of Arrival）推定結果，および，人位置推定システムから取得する。さらに，合成したバーチャル音場の加工を制御するために 3 つのユーザインタフェースを提案し，被験者実験を通して検証した。

2 提案システム

提案システムは二つの部分から構成されている。一つはロボット側の音源位置推定・トラッキングと複数人の音源分離であり，もう一つはオペレーター側の頭部回転トラッキングとステレオ音声の合成である。Figure 1 に提案システムのブロック図を示す。

ロボット側の処理では，まず，各マイクロフォンアレイによって音の 3 次元到来方向（DOA）が推定される。環境とアレイの位置関係と各音源の DOA を統合することで，3 次元上での人位置情報が得られる。この人位置情報は，ヒューマントラッキングシステムにより，非発声時にも常時追跡されている。次に，推定した人位置情報に基づいて各人の音声を分離し，位置情報と合わせてオペレーター側のシステムに送信する。



(a) Robot environment side

(b) Tele-operator side

Figure 1. Block diagram of the proposed tele-presence system.

オペレーター側の処理では，まず，人位置情報とオペレーターの顔の向きによって，左右のチャンネルに対応した正確な HRTF をデータベースから選択する。次に，分離した音声に畳み込み演算を行い，ステレオヘッドフォンでオペレーターに再生する。オペレーターの頭部回転トラッキングには，ヘッドフォンの上部に取り付けたジャイロセンサーとコンパスを用いた。また，分離した各音源のボリュームは，ユーザインタフェースにて独立して調節することができる。

2.1 3 次元音源定位

音源定位に関して，まず，各マイクロフォンアレイで DOA 推定を行う。複数のアレイによる DOA 情報と人位置情報を統合することで，音源の 3 次元空間内の位置を推定する。

実環境での音の DOA 推定は広く研究されてきた。MUSIC 法は，複数のソースを高い分解能で定位できる最も有効な手法の一つである[Schmidt 1986]。この手法を使うには事前に音源数が必要であるため，本研究では[Ishi 2009]で提案した解決法を用いる。音源数を固定した数値に仮定し，閾値を超えた MUSIC スペクトルのピークを音源として認識する。この研究で使用した MUSIC 法の実装は 100 ms ごとに 1 度の分解能を有しており，2 GHz のシングルコア CPU でリアルタイムに探索することができる。

コミュニケーションロボットの遠隔操作システム

にとって，最も重要な音源は人の音声である。本研究では人の声を漏れ無く抽出するために，複数の2D-LRF（Laser Range Finder）で構成したヒューマントラッキングシステムを使用した[Glas 2007]。複数のマイクロフォンアレイからの DOA 推定出力と LRF のトラッキング結果が同じ位置で交差すれば，そこに音源がある可能性が高い[Ishi 2013] [石井 2014]。本システムでは 2D の LRF を用いているため、人位置情報は 2D に限られる。ここでは、検出された音源の位置が口元の高さの範囲内にあるかの制限もかけている（z = 1 ~ 1.6m）。 無音区間や音源方向推定が不十分な区間では、最後に推定された口元の高さと最新の 2D 位置情報を用いて，音源分離を行う。

2.2 音源分離

音源分離では，選択された複数の人物をパラレルに分離している。本研究では計算量が少なく且つロバストな Delay-Sum Beamformer を用いて，目的方向の人の声を分離した[Dudgeon 1977]。フレーム長は 20 ms で，シフト長は 10 ms である。

本研究で使用した 16 チャンネルのマイクロフォンアレイ（半球 30cm にマイクを配置した形状）の DS ビームフォーマのレスポンスの特徴として，低周波領域の分解能が低いことが挙げられる。そのため，無指向性雑音の低周波成分が分離音に多く混在してしまい、臨場感の伝達に悪影響を与える可能性がある。

空間に指向性音源 S と無指向性雑音源 N が存在すると仮定すると，DS ビームフォーマの出力は以下の形になる：

$$Y(f) = \boldsymbol{w}_{Sdir}(f) \cdot S(f) + \int_0^{2\pi} (\boldsymbol{w}_\theta(f) \cdot N(f)) d\theta$$

Y は周波数 f に対応したビームフォーマの出力で，Sdir は信号の方向，$\mathbf{w}_{\mathrm{Sdir}}$ は Sdir 方向のビームフォーマレスポンスを指す。式の二つ目の項目は，分離音声に混在する雑音を表している。この雑音成分を低減させるために，各周波数に以下のようなウェイトを掛けた。

$$w_{PF}(f) = \frac{1}{\int_0^{2\pi} \boldsymbol{w}_\theta(f)\, d\theta}$$

$$Y_{PF} = \sum_f w_{PF}(f) \cdot Y_{DS}(f)$$

Y_{PF} はウェイト掛けした後のビームフォーマ出力である。

さらに，各音源とアレイの間の距離による違いを補正するため，分離した各音声に対して距離によって以下のように正規化を行った。

$$g_i = \frac{\sum_{n=1}^{N} dist_n - dist_i}{(N-1) \cdot \sum_n^N dist_n}$$

$$Y_i = g_i \cdot Y_{PF,i}$$

このうち，N は音源の数で，$\mathrm{dist_n}$ は n 番目の音源とアレイの距離を表す。g_i は i 番目の音源に掛ける正規化ファクタで，Y_i は i 番目の音源の分離結果を示している。

2.3 HRTF による音場合成

一つの音声を特定の方向から聞こえるようにするため，その方向に対応した HRTF によってフィルタリングするステレオ化方法が一般的である。本研究では，一般公開されている KEMAR (Knowles Electronics Manikin for Acoustic Research) ダミーヘッドの HRTF データベースを利用した[Gardner 1995]。KEMAR は HRTF 研究のために一般的な頭部サイズを使って作られたダミーヘッドで，データベースには空間からのインパルス信号に対するダミーヘッドの左右耳のレスポンスとして，仰角-40 度から 90 度までの総計 710 方向のインパルス応答が含まれている。各インパルス応答の長さは 512 サンプルで，サンプリング周波数は 44.1 kHz である。

前述のように, HRTF を用いてダイナミックに音場を合成するには，頭部の向きのリアルタイム検出が必要である。このため，本研究ではヘッドフォンの上部にジャイロセンサーとコンパスを取り付け，頭部回転のトラッキングを行った。角度情報はシリアルおよびブルートゥース経由のいずれかでシステムに送られる。音場の合成に使う方向は音源方向から頭部角度を引いたもので，この方向に対応した左右チャンネルのインパルス応答がデータベースから選出され，分離結果と畳み込み演算を行った音声がオペレーターの両耳に再生される。

3 システム評価

提案システムを評価するため，被験者実験を行った。被験者はロボットを介してロボット側にいる人物と会話をし，ロボット側の視覚情報無しの状態で，その対話相手のいる方向を推定することが求められる。

比較対象として，ロボットの耳に位置するステレオマイクロフォンを用いた。この実験ではミニマルデザインされているヒューマノイドロボット Telenoid-R3（figure 3 左上）を使用した。このロボットは両耳位置にマイクの装着が可能で，且つ，首には 3 自由度があるため，人の頭部動作を線形的にマッピングすることができる。

以下に，比較対象の条件を述べる。この条件では，ロボットの耳にある二つのマイクロフォンから採った音を，そのままオペレーターのステレオヘッドフォンの左右チャンネルで再生する。トラッキングしたオペレーターの首の動きは，線形的にロボットにマッピングされる。

Figure 2. External appearance of the Telenoid R3 (top left), operator environment (bottom left) and the robot environment where interaction experiments were conducted (right).

Figure 2 の左下図にオペレーター側の環境を，右図にロボット側の環境の様子を示す。ロボット側の 3D 音源位置推定は, 3 つのマイクロフォンアレイによって行われた。Figure 2 右図に赤矢印で示してあるように，天井には直径 15 cm で 8 チャンネルのマイクが円形に配置されたマイクロフォンアレイが 2 つ設置してあり，卓上には直径 30 cm で 16 チャンネルのマイクが半球面上に配置されたマイクロフォンアレイが設置してある。

総計 20 名の被験者がこの実験に参加した。全て大学生で，ロボットや音響研究に関わりがない者である。被験者にはオペレーター役として，別室のロボット側にいる話者 1 名（研究補助者）とロボットを介して会話し，その相手のいる方向を判定するように指示した。実験補助者はランダムに方向を選び，その方向から会話を進める。被験者は方向の判定ができたら協力者に知らせ，協力者は次の方向に移動する。この手順を 4 回繰り返した。方向の判定は 8 方向に制限しており，被験者はそのうちのどの方向かを回答するという形式である。

実験の最後に，二つの条件について，臨場感と聞き取り易さに関する主観評価のアンケートを採った。1 から 7 までの七段階評価で，1 は「臨場感が低い/聞き取り難い」で，7 は「臨場感が高い/聞き取り易い」を示す。

Figure 3 上図に，提案システム条件と比較条件での方向定位の精度の平均値とその標準偏差を示す。T-test の結果，両者の精度差に有意差がみられた（$t = 0.59, p < 0.001$）。

主観評価アンケートでは，臨場感と聞き取り易さの評価で類似した結果が得られた。Figure 3 下図にその結果を示す。臨場感と聞き取り易さの両方において，提案システム条件での評価は，比較条件よりも有意に高い（$t = 6.68, p < 0.001$ と $t = 4.86, p < 0.001$）。

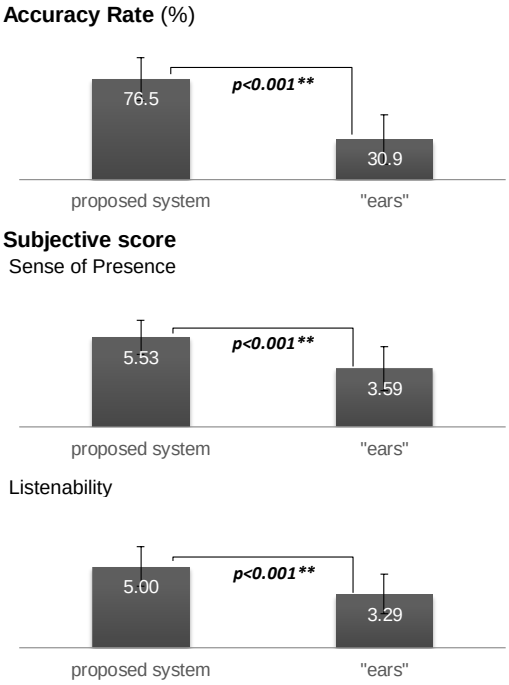


Figure 3. Accuracy rates for direction perception and subjective scores (1 to 7 scale) for sense of presence and listenability, in two conditions: “proposed system” and “robot’s ears”.

両条件で聞き取り易さに差が出た理由として，ロボットの両耳位置に埋め込まれたマイクロフォンの SNR が考えられる。このマイクロフォンはロボット内部のモーターに物理的に近いためモーターノイズの影響を受けやすく，これが SNR の低下に繋がったと考えられる。臨場感の評価にも両条件で有意差が見られたが，可能な理由としては，ロボットの首と人間の首の可動範囲が違うことが挙げられる。人間の首の可動範囲はロボットより広いため，オペレーターが首を回している途中でもロボットの首はすでに最大角度にヒットしている可能性がある。このオペレーターとロボットの頭部オリエンテーションのミスマッチが臨場感の評価に影響した可能性がある。

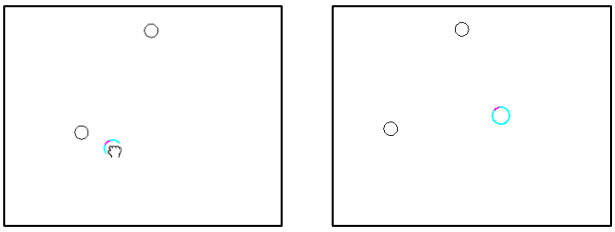
4 バーチャル音場における音源ボリュームの調整

提案システムでは，選択されたすべての音源に対して，位置情報を反映したステレオ音声を合成し，足し合わせて，バーチャル音場を表現する出力が再生される。しかし，これでは選択された各音源のボリュームが予測できない。もし，オペレーター側で各音源のボリュームを各々独立して操作することができれば，自分にとって最も快適な音環境を作ることができる。このことに注目して，オペレーターがバーチャル空間上にある音源や自分の位置を変えることができるように，二つのインタフェースを提案した。

4.1 提案のユーザインタフェース

このセクションでは，バーチャル音場をコントロールするための 2 つの異なる操作パターンのユーザインタフェースについて説明する。

Figure 4 に二つのインタフェースのスクリーンショットを示す。



(a) “drag-and-drop” (b) “face dir”

Figure.4. Screen shots of the displays for different user interfaces.

Figure 4 (a)に示す 1 つ目のインタフェースでは，オペレーターがスクリーン上の青い円（これはバーチャル空間上でのオペレーターの位置を表す）を任意の場所にマウスでドラッグ&ドロップすることによって，各音源のボリュームを調整する。希望の場所へ自身のバーチャルな位置を移動させることによって各音源との距離・角度が再計算され，音源のボリュームがその距離に従って変更される（特定の音源に接近させると，その音源のボリュームが大きくなる）。このインタフェースを“drag-and-drop”と表記する。実環境での会話シーンでは，会話参加者間の物理的距離は環境や相手との社会的関係に影響される。“drag-and-drop”は，この観点に注目したバーチャル音場コントロール法である。

Figure 4 (b)に示す 2 つ目のインタフェースでは，オペレーターの顔の向きによって各音源のボリュームが調整される。オペレーターの顔方向を利用して音源の音量を操作するため，両手が解放される。オペレーターの顔の前方にある音源は強調され，後方にある音源は減衰される。ボリュームを調節するファクタは角度と比例する。このインタフェースを“face dir”と表記する。顔の向きや視線方向は現時点における人の注意を示すだけでなく，次のターゲットやそのゴールをも示す[Langton 2000] [Yokoyama 2012]。“face dir”はこの観点に注目したバーチャル音場コントロール法である。

4.2 提案ユーザインタフェースの評価

提案のユーザインタフェースを評価するための被験者実験を行った。比較対象として，従来のモノラルマイクロフォンを使ったインタフェースを用いた。

前セクションで述べた実験被験者が，この実験にも参加した（大学生 16 名。前セクションの 20 名中最初の 4 名は従来法との比較を行っていないため除外）。実験のデザインは被験者内比較を採用した。被験者は提案インタフェース及び従来のインタフェースを使って，ロボット側の環境にいる対話者 2 名（研究補助者）と会話をする。会話トピックに制限はない。用いたインタフェースごとに会話のセッションを分けた。セッションの長さは 3 分間で，各セッション終了後にインタフェースの「使い易さ」「臨場感」「聞取り易さ」に関して前実験と同じく 1 から 7 まで 7 段階の主観評価アンケートを採った。

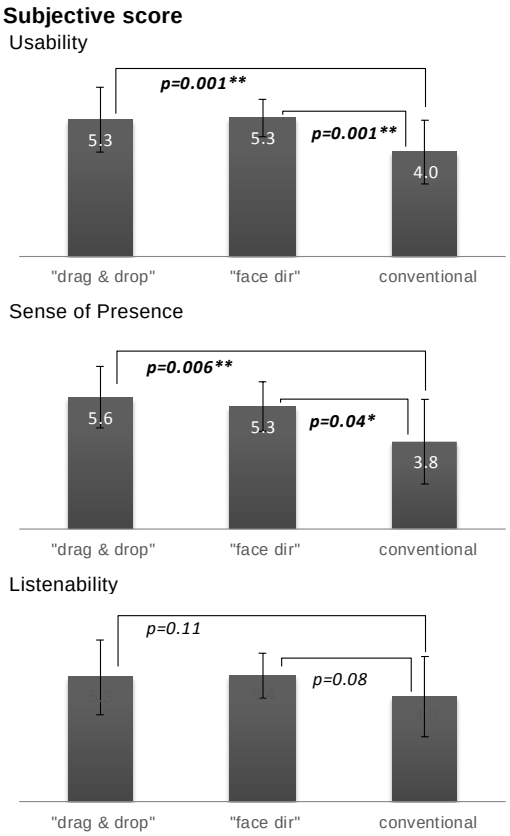


Figure 5. Subjective scores (1 to 7 scale) for three types of user interface: “drag and drop”, “face dir” and “conventional”.

Figure 5 に各インタフェースに対する主観評価の平均値と標準偏差を示す。実験結果に対して分散分析（ANOVA, with-in participants, Bonferroni’s posttest）を行った。

「使い易さ」（Figure 5 上図）と「臨場感」（Figure 5 中図）では，主観評価の平均値に有意差が見られた（*F(2,13)=16.03, p<0.001 and F(2,13)=6.74, p=0.009*）。多重比較（Bonferroni 法）の結果，提案法である“drag-and-drop”と“face dir”は従来法よりも使い易く（*“drag-and-drop” vs. “conventional”: p=0.001; “face dir” vs. “conventional”: p=0.001*），臨場感が高い（*“drag-and-drop” vs. “conventional”: p=0.006; “face dir” vs. “conventional”: p=0.04*）と評価された。しかし，「聞取り易さ」では有意差が見られなかった（*F(2,13)=3.67, p=0.052*）。

以上の結果は，提案インタフェースの有効性を示

している。

4.3 考察

ユーザインタフェースの評価実験は，興味深い結果を示している。通常、オペレーターとロボットは連動することで臨場感を感じるが、“drag-and-drop”インタフェース使用時では，被験者のみが自分（ロボット）の位置をバーチャル空間で変えるだけで、ロボットは実際に移動していないにも関わらず,「臨場感」の評価が高かった。

「聞取り易さ」の評価結果に関しては，提案インタフェースに対する評価スコアの平均値は従来法より高いものの，有意差が見られなかった。この可能性として，以下の理由が考えられる。今回の実験ではロボット側にいる対話者が２名のみであるため，多人数対話環境と比較して音の収録状況が良好である。そのため，従来法でも難なく音声を聴き取ることができたと考えられる。音源が増えるに連れて聞取り易さにも差が出る可能性があるが，これについての検証は今後行なう予定である。

また，今回の実験ではダミーヘッドの HRTF データベースを利用したが，被験者の頭部の形状に対応した HRTF を合成できれば，システムの効果の向上が期待できる。

5 おわりに

本稿では，操作者の頭部の動きに合わせて遠隔ロボットの環境の 3D 音場を合成する遠隔コミュニケーションロボット操作システムを提案し，被験者実験によってこれを評価した。

マイクロフォンアレイを用いて音源を収録し音場を合成する提案法は，ロボットの両耳にマイクを装着させて音源を収録した手法よりも，音源位置の同定実験では有意に高い精度を示し，臨場感と聞き取り易さの主観評価実験では、いずれも有意に高い評価が得られた。

また，バーチャル音場における音源のボリュームを操作するために 2 種類のユーザインタフェースを提案し，これを被験者実験によって評価した。

その結果，オペレーターがスクリーン上で音源に対する自身のバーチャルな位置を変更させてボリュームを調整する方法，及び，オペレーターの顔の向きに応じてボリュームを調整する方法は，従来法よりも「使い易さ」と「臨場感」の評価において有意に高く評価された。

謝辞

本研究は JST/CREST の委託研究により実施したものである。音源定位に関するシステムの一部は、総務省 SCOPE の委託研究により開発されたものを利用している。評価実験にご協力いただいた森田美香氏、波多野博顕氏に感謝する。

参考文献

[Nishio 2007] Nishio, S., Ishiguro, H., Hagita, N. Can a Teleoperated Android Represent Personal Presence? - A Case Study with Children. *Psychologia*, 50(4): 330-342. 2007.

[Ishi 2010] Ishi, C.T., Liu, C., Ishiguro, H., Hagita, N. 2010. Head motion during dialogue speech and nod timing control in humanoid robots. *In Proceedings of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010)*. OSAKA, JAPAN. 293-300.

[Liu 2012] Liu, C., Ishi, C. T., Ishiguro, H., Hagita, N. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. *In Proceeding of ACM/IEEE International Conference on Human Robot Interaction (HRI 2012)*. Boston, USA. 285-292, March, 2012.

[Sumioka 2014] Sumioka, H., Nishio, S., Minato, T., Yamazaki, R., Ishiguro, H. Minimal Human Design Approach for Sonzai-kan Media: Investigation of a Feeling of Human Presence. Cognitive Computation, 2014.

[Popescu 2000] Popescu, V. G., Burdea, G. C., Bouzit, M., Hentz, V. R. A virtual-reality-based telerehabilitation system with force feedback. *IEEE transactions on Information Technology in Biomedicine*. 4(1): 45-51. 2000.

[Piron 2009] Piron, L., Turolla, A., Agostini, M., Zucconi, C., Cortese, F., Zampolini, M., Zannini, M., Dam, M., Ventura, L., Battauz, M., Tonin, P. Exercises for paretic upper limb after stroke: a combined virtual-reality and telemedicine approach. *J. of Rehabilitation Medicine*. 41(12): 1016-1020(5). 2009.

[Billinghurst 2002] Billinghurst, M., Cheok, A., Prince, S., Kato, H. Real world teleconferencing. *IEEE Computer Graphics and Applications*. 22(6): 11-13. 2002.

[Ogi 2001] Ogi, T., Yamada, T., Tamagawa, K., Kano, M. Immersive telecommunication using stereo video avatar. *Proceedings of Ieee Virtual Reality*. Yokohama, Japan. 45-51. 2001

[Bullinger 1997] Bullinger, H., Riedel, O., Breining, R. Immersive Projection Technology- Benefits for the Industry, *International Immersive Projection Technology Workshop*, 13-25, 1997.

[Pulkki 2007] Pulkki, V. Spatial sound reproduction with directional audio coding. *J. Audio Eng. Soc*. 55(6): 503-516. 2007.

[Laitinen 2011] Laitinen, M., Kuech, F., Disch, S., Pulkki, V. Reproducing applause-type signals with directional audio coding. *J. Audio Eng. Soc*. 59(1/2): 29-43. 2011.

[Rumsey 2001] Rumsey, F. *Spatial Audio*. Focal Press, 2001.

[Meyer 1972] Meyer, E., Neumann, E. *Physical and Applied Acoustics: An Introduction*. Academic Press, New York, 1972. ISBN 0124931502.

[Cheng 2001] Cheng, C. I., Wakefield, G. H. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. *J. Acoust. Soc. Am*, 49(4):231-249, April 2001.

[Iwaya 2003] Iwaya, Y., Suzuki, Y., Kimura, D. Effects

of head movement on front-back error in sound localization. *Acoustical Science and Technology*. 24(5): 322-324. 2003.

[Schmidt 1986] Schmidt, R. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34, 276-280, 1986.

[Ishi 2009] Ishi, C. T., Chatot, O., Ishiguro, H., Hagita, N. Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments. Proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 09)*. 2027‑2032. 2009.

[Glas 2007] Glas, D.F. et al, 2007. Laser tracking of human body motion using adaptive shape modeling. In Proceedings of the *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2007)*, 602-608. 2007.

[Ishi 2013] Ishi, C., Even, J., Hagita, N. (2013). Using multiple microphone arrays and reflections for 3D localization of sound sources. In Proc. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013)*, 3937-3942, Nov., 2013.

[石井 2014] 石井カルロス寿憲, Jani EVEN, 萩田紀博, (2014) "複数のマイクロホンアレイと人位置情報を組み合わせた音声アクティビティの記録システムの改善", 第３２回日本ロボット学会学術講演会, Sep. 2014.

[Dudgeon 1977] Dudgeon, D. E. Fundamentals of digital array processing. *Proceedings of the IEEE*. 65(6): 898-904. 1977.

[Gardner 1995] Gardner, W. G., Martin, K. D. HRTF measurements of a KEMAR. *J. Acoust. Soc. Am*. 97(6):3907-3908, Jun. 1995.

[Langton 2000] Langton, S. R., Watt, R. J., Bruce, I. I. Do the eyes have it? Cues to the direction of social attention. *Trends Cog. Sci*. 4, 50–59, 2000.

[Yokoyama 2012] Yokoyama, T., Noguchi, Y. Kita, S. Attentional shifts by gaze direction in voluntary orienting: evidence from a microsaccade study. *Exp. Brain Res.* 223, 291–300, 2012