3K1-R-11-1

遠隔ユーザの音環境理解を支援するユーザインタフェース

A User Interface for Remote User to Support Understanding Sound Environment

植田 俊輔*1 Shunsuke Ueda　今井 倫太*1 Michita Imai　中村 圭佑*2 Keisuke Nakamura　中臺 一博*2 Kazuhiro Nakadai

*1慶應義塾大学 Keio University　*2ホンダ・リサーチ・インスティチュート・ジャパン Honda Research Institute Japan

In a telepresence situation, a remote user has difficulties in catching and joining conversations because of the mixture of sound sources in a remote place. To relax this problem, this paper proposes User Interface for Avatar-based Listenable Telepresence (UI-ALT): A remote user can see scenes and listen to conversations via a real world avatar like a telepresence robot having a camera and microphone array. The user selects a conversation by marking persons of interests as a circle or a line on a UI-ALT display. The user can listen only to the selected conversation even when several conversations occur simultaneously because sound source separation with the microphone array eliminates non-target sound sources.

1. はじめに

人間は，周囲に雑音が多い環境であっても，その場所でどのような会話が行われているか理解することができる．例えば，パーティ会場は音楽やグラスが交わる音，様々な会話など多様な音源が存在し，非常に高雑音な環境であるが，我々はどの方向でどのような会話が行われているか理解することができる．この現象は「カクテルパーティ効果」[1] という心理学の言葉で知られている．しかし，テレプレゼンスロボットがこのような環境下 (図 1) にある場合，遠隔操作者は遠隔地でどのような会話が行われているか理解することは難しい．操作者が操作インタフェース上で遠隔地の様子の画像を見ながらであっても，どの音がどの音源から来ているのかをヘッドフォンを通してインタフェース上で判断することは困難である．

近年，アバタとしてのテレプレゼンスロボットが様々な方法で研究されており [2][3][4]，Anybots 社の QB[4] のように商品化されているものもある．これらのロボットは，遠隔地で存在感を発揮し，人間の代わりに様々なタスクを遂行することが期待されている．しかし，これらのロボットは遠隔地の人間とのインタラクションに必要な音声の情報をうまく処理することができず，高雑音環境において人間とインタラクションを行うことが難しいと考えられる．日常環境において無音という環境はあまり例がなく，実世界におけるアバタロボットと人間のコミュニケーションには，遠隔地の音環境理解が不可欠な要素となり得る．

本稿では，テレプレゼンスロボットの遠隔操作者が遠隔地の音環境を理解しやすくするようなユーザインタフェース UI-ALT(User Interface for Alternative Listenable Telepresence) を提案する．UI-ALT は，マイクロフォンアレイを搭載したテレプレゼンスロボットのための操作インタフェースであり，オープンソースなロボット聴覚ソフトウェア HARK[6][7] が提供する音源定位，音源分離の機能や分離音の音声認識情報により，操作者はインタフェースの画面上で自分が聴きたい方向の分離音を選択的に聴取することが可能である．つまり，

連絡先: 植田俊輔，慶應義塾大学大学院理工学研究科，〒 223-8522 神奈川県横浜市港北区日吉 3 − 1 4 − 1，045-560-1070，ueda@ayu.ics.keio.ac.jp

図 1: アバタロボットが雑音環境にいる様子

UI-ALT を用いることで聴き分けができるテレプレゼンスロボットが実現可能とし，遠隔地が雑音環境であっても，操作者と遠隔地の人とのインタラクションが円滑に行うことができると考えられる．

本稿は以下の通りに展開する．2. 節ではロボット聴覚に関する研究背景および関連研究について述べ，3. 節で UI-ALT のシステム構成全般および主要なモジュールついて詳述する．4. 節で UI-ALT が実際どのような場面で操作者に使用されるか例を挙げ，5. 節で結論と今後の課題をまとめる．

2. 研究背景

2.1 ロボット聴覚

近年，1. 節で述べた「カクテルパーティ効果」を実現させるロボットを目指し，ロボット聴覚に関する研究が進んでいる．ロボットがカクテルパーティ効果を実現するための方法は，音声信号の入力を受け取るマイクの使い方で，バイノーラル方式とマイクロフォンアレイ方式に区別できる．

人間は両耳を使って音声情報を取り入れているが，バイノー

ラル方式は両耳に見立ててマイクを２つ配置し，音響信号処理を行うものである．つまり，２本のマイクが両耳の役割を果たすということになる．バイノーラル方式を用いる聴覚ロボットの代表例として Telehead[10] が挙げられる．Telehead は人間の顔を模したダミーヘッドにバイノーラルな２チャネルのマイクを配置したテレプレゼンスロボットであり，これまでに頭部の形状が音声情報処理に与える影響などが検証されてきた．しかし，Telenoid に限らずバイノーラル方式全般に言える問題点として，入力信号が２チャネルでしか処理されないため，音源定位や分離の処理がやや貧弱であるということが挙げられる．音響信号処理においては複数音源における定位情報が重要な役割を果たすため，より精度が高い処理が要求される．

これに対し，複数のマイク（４〜８チャネル）を規則的に配置し，音響信号を様々な方向から取得することにより，音源定位や分離の性能を高めたマイクロフォンアレイ方式でロボット聴覚を実現する方法がある．マイクロフォンアレイを用いたロボット聴覚ソフトウェアとして HARK[6][7] がある．HARK はオープンソースで配布されるロボット聴覚ソフトウェアであり，マルチチャネルマイク入力による音響信号処理によって複数音源の定位，分離，分離音の音声認識までサポートしている．HARK は，ミドルウェア上でモジュール接続によるプログラミングで簡単に聴覚機能を持つロボットを実装可能であり，最近では様々な種類のマイクロフォンアレイに対応している．このように，近年ロボット聴覚は誰でも簡単に実現可能なレベルに近づいている

2.2 HARK on Texai

ここではロボット聴覚ソフトウェア HARK を用いてユーザインタフェースを実装した研究例について述べる．水本ら [5] は Willow Garage 社のテレプレゼンスロボット Texai に音の選択聴取が可能なユーザインタフェースを実装した．本研究では，音がどこから来ているのかという音の定位情報と，来ている音の強さを可視化させて操作者に音を選びやすくしている．しかし本研究では，遠隔操作者が分離された音を聴く際に，音の方向と幅の２つのパラメータを操作しなければならないので，実際に操作者が分離音を聴く際に煩雑な操作が要求される．このため，リアルタイムで遠隔インタラクションを行う際，操作者が自分の聴きたい音を素早く聴き取ることが難しいと考えられる．

本稿では，テレプレゼンスロボットの操作者がより容易に素早く，遠隔地における自分が聴きたい音をインタフェースを通して聴くことができるために，音源選択の方法や操作者に提示する視覚情報が最適なものになるようインタフェースを設計することで，関連研究における問題点の解決を図る．

3. UI-ALT システム構成

UI-ALT のシステム構成図を図 2 に示す．

UI-ALT の操作者は，インタフェースの画面上で複数人が同時に喋っている中で会話を聴きたい方向の人に対してマウスを用いて線や丸を描いたり，その人を直接クリックしたりすることで，選んだ方向の分離音を聴くことができる．この機能は，オープンソースなロボット聴覚ソフトウェア HARK[6][7] を利用した音源定位・分離のモジュールの組み合わせによって実現される．

音源定位や分離，遠隔地のカメラ映像などはすべて ROS（Robot Operating System）[8][9] のメッセージで通信を行う．UI-ALT では音声データとカメラデータを同時に処理するため，処理が重くなってしまう可能性がある．そこで ROS が提

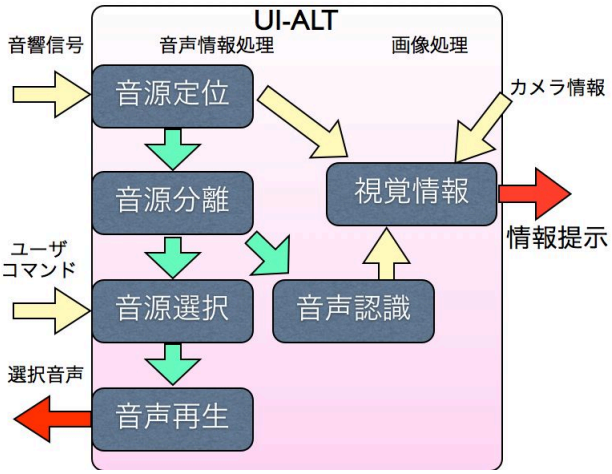

図 2: システム構成図

供するメッセージを用いて通信を行うことにより，音声波形信号や音源 ID，カメラ画像情報など多種にわたるデータを小さい遅延で通信することが可能である．

以下小節では，UI-ALT の主要なモジュールについて詳述する．

3.1 音源定位モジュール

入力である音響信号は最初に音源定位モジュールに送られる．音源定位モジュールではどの音がどの方向から来ているのかを推定することが出来る．音源の定位には HARK で提供されている雑音に頑健で，複数音源の定位が可能な MUSIC(MUltiple SIgnal Classification)[6] を用いる．MUSIC により，複数音源の水平方向の定位が可能となる．定位情報は入力音響信号とともに音源分離モジュールへ渡される．また，定位情報は UI-ALT の画面上で可視化するために画像情報表示モジュールへも渡される．

3.2 音源分離モジュール

定位された音響信号は，音源分離モジュールへ送られる．音源分離モジュールでは，選択的な会話の聴取を実現するために．定位情報と入力音響信号（混合音）から各音源信号を分離する．UI-ALT では HARK で提供されている GHDSS(Geometric-constrained Highorder Decorrelation-based Source Separation)[6] を用いて音源分離を行う．分離された音源情報は音源選択モジュール，画像情報で音声認識結果を表示するために音声認識モジュールへ送られる．

3.3 音源選択モジュール

音源選択モジュールは，操作者のマウスコマンドによって，分離された音源を選択して音声再生モジュールに渡すモジュールである．操作者がどの音源も選択していない場合は入力音響信号がそのまま再生モジュールに渡される．UI-ALT では，選択したいグループを丸で囲う，選択したいグループの上に線を引く，選択したいグループを直接マウスで左クリックして選択するといった方法で音源選択をすることができる．

操作者がマウスモーションにより UI-ALT 上に円もしくは線を描き終わる，もしくは左マウスクリックが行われると，UI-ALT は以下の処理で音源選択を行う．

1. 描画の場合，描かれた円もしくは線の画像内における x, y 座標の最大値および最小値を取得する．クリックで直

接選ぶ場合，クリックされた点から ±5° くらいの点を自動で抽出し，選択範囲を示す楕円を描く．

2. 選択範囲の x 座標の最大値および最小値をあらかじめ決められている USB カメラの画角から以下の式で角度に変換する．画像サイズは 640 × 480 であり，画像の中心が 0° である．

$$\theta = \pm \arctan\left(\frac{|x - 320| \times \tan\left(\frac{\text{カメラ画角}}{2}[deg]\right)}{320}\right) \quad (1)$$

3. 算出された角度範囲と音源の角度を比較して範囲に含まれていれば音源が選択されたと判断する．
4. 音声再生モジュールへ選択された分離音情報を送る．

UI-ALT では複数の音源を選択することも可能であり，複数選択された場合には選択された音源の数の分の混合分離音が再生される．また音源選択を解除することも可能である．操作者がマウスを右クリックすることで，音源選択状態をリセットして何も選択していない状態に戻すことが出来る．

3.4 視覚情報提示モジュール

画像情報提示モジュールでは，3.1 節で述べた各音源の定位情報，3.2 節で述べた分離音声の音声認識情報を，カメラ映像とともにインタフェース上に表示するモジュールである．操作者が自分で聴きたい音を選択する際，視覚的に操作者を補助するために必要な情報を画面上で行う．これにより操作者はよりスムーズに音源選択を行うことが可能である．音声認識の結果は画面外のインターフェース上に方向毎に別で表示される．

以上で示したモジュール群により，UI-ALT は遠隔操作者が遠隔地の音環境を理解しやすくなるようなインタフェースの実現が可能となっている．

4. 応用例

本節では，UI-ALT が実際どのように操作者に使用されるのか，あらかじめ録音録画したデータをオフラインで再生することによる動作例を示すことで明らかにする．

4.1 シナリオ

今回動作例を示すために，あらかじめ実世界で起こりうる雑音環境の中で会話を行うという場面を録音録画した．今回録音録画するにあたり，4 人の大学生に部屋に集まってもらい，アバタロボットと同じ部屋でパーティに参加するという想定をした．集まった 4 人を 2 人 1 組のペアに分け，図 3 で示すようにアバタロボットの正面の向きから ±30° の位置に 2 組を配置した．そしてお互い初対面という想定で背景音楽が流れる室内で自己紹介を 1 分間行ってもらった．操作者が遠隔地の会話内容を理解しているかを検証するため，お互いの自己紹介の中で，名前・所属・出身・趣味の 4 つの内容に関して必ず触れてもらった．

音声の録音は頭部に 8 チャネルのマイクロフフォンアレイを搭載したアバタロボットを使用し，映像の録画には USB カメラを使用した．背景音楽はロボットの後方にスピーカを置き，様々な曲をランダムに再生を行った．会話はお互いランダムにやってもらったため，録音データは音楽とランダムな会話の混成音データとなり，生のデータだと遠隔操作者がヘッドフォンを通して遠隔地の音環境を理解することは非常に難しいものになっている．

操作者は録音録画されたシーンを UI-ALT で再生する場合と，そのまま再生する場合の 2 通りで遠隔地の会話内容をそれぞれ書き起こしてもらった．

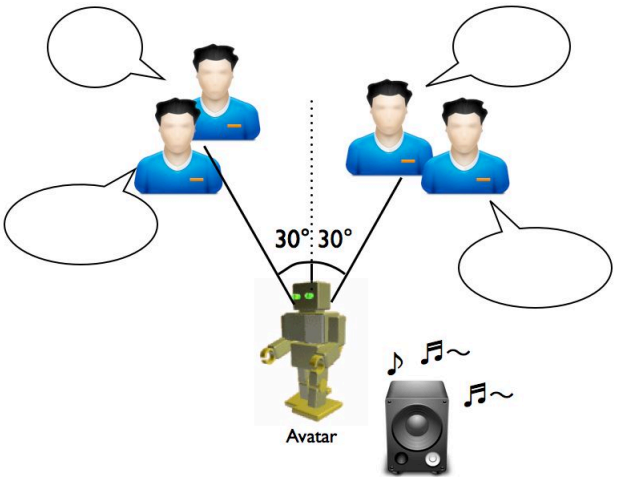

図 3: 録音録画場面の室内配置

4.2 UI-ALT の動作例

操作者は 3. 節で述べたように，UI-ALT の操作画面上で自分が音声を聴きたい人を選んで分離音を聴くことができる．操作者はそれぞれ図 4，5，6 で示すように，

- 音を聴きたい方向を丸で囲う
- 音を聴きたい方向に線を引っ張る
- 音声を聴きたい人を直接クリックする

といった選択方法がある．操作者は自分が選択しやすいようどの選択方法も好きなように利用できる．

また，図 4，5，6 の上部には音源の定位情報を可視化し，カメラの画角の範囲で音がどの方向から来ているのか，という情報を四角形で画面上部に表示することで操作者に提示している．これにより操作者はリアルタイムでどこから音が来ているのかを知ることができる．

4.3 操作者による意見

実際に 8 名にオフラインで UI-ALT の操作者になってもらい，UI-ALT を用いて遠隔地の会話を選択聴取することで何も用いない場合と比べて

- 会話は聴こえやすくなったか
- 会話を書き起こす際にどれくらい役立つと思うか
- 会話を選びやすくなったと思うか

の 3 つの項目について 5 段階評価でアンケートを実施した．結果を図 7 に示す．

アンケートの結果より，3 つの項目全てにおいて UI-ALT を用いて優れた結果が得られた．このことから，UI-ALT が操作者にとって遠隔地の音環境理解に有用なユーザインタフェースであることが示されていることがわかる．また，数名の操作者の自由記述により，UI-ALT の視覚情報提示部分でより効果的な情報を提示すれば，操作者にとってより自分が聴きたい音源を選びやすくなるのではないか，といった意見が得られた．このことより，操作者が音源を選択聴取する際に，遠隔地の音環境の視覚情報が不可欠であることが伺える．

5. まとめ

本稿では，ロボットの遠隔操作者が遠隔地の音環境を理解しやすくするようなユーザインタフェース UI-ALT を提案した．UIALT は人とアバタロボットとのインタラクションにおいて欠かすことのできない音声情報を扱えるインタフェース

図 4: 丸で囲って選択する場合

図 5: 線を引っ張って選択する場合

図 6: マウスクリックで選択する場合

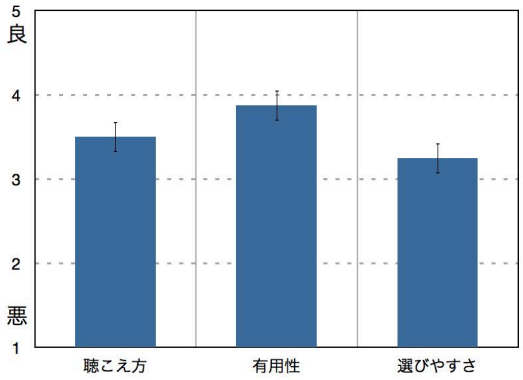

図 7: アンケート結果

であるため，実世界の様々な環境に適用可能であると考えられる．本稿では実際に UI-ALT の応用が可能であると考えられる３つのインタラクションシナリオを示した．

今後の課題として，まずインタフェースの改善が挙げられる．現在，あらかじめ録音したデータを用いるオフライン実験を通して操作者にとってインタラクションを行うために有用な情報は何か，ということを調査中である．インタフェース上で聴覚情報だけでなく，画像情報も用いて操作者が音を選びやすい環境を提供しているため，どの情報が操作者にとって本当に役に立つのかを見極めていく必要がある．現在，ロボットの方向に話しかけていると思われる人を顔認識してインタフェースに認識情報を反映させる，画面外の話者情報をインタフェースに表示させて画面外の情報にも注意を向けさせる，といったことを実装中であり，オフライン実験で実用性の検証を行う予定である．また，オフライン実験から得られた知見によりインタフェースを改善し，再びオフライン実験でその有用性を確かめていく予定である．

UI-ALT を用いたオンライン実験も計画している．今後行うオフライン実験で得られた知見を基に，アバタロボットを遠隔操作出来るようインタフェースを改良し，実際にパーティにアバタロボットを参加させて遠隔でユーザに参加してもらう．その際，遠隔地の人と遠隔操作者がどのようなインタラクションを行うのかということを検証する実験を行っていく予定である．

参考文献

[1] Cherry E. Colin: Some Experiments on the Recognition of Speech, with One and with Two Ears, in *The Journal of the Acoustical Society of America, vol.25, pp.975-979, 1953.*

[2] Sigurdur Orn Adalegeirsson, Cynthia Brezeal: Mebot a robotic platform for socially embodied telepresence. in *Proc. of ACM/IEEE International Conference on Human-Robot Interaction(HRI), pp.15-22, 2010.*

[3] Nishio, S, Ishiguro, H., Anderson, M., Hagita, N.: Repesentating personal presence with a teleoperated android: A case study with family. in *Proc. of AAAI 2008 Spring Symposium on Emotion, Peronality, and Social Behavior, pp.96-103, 2008.*

[4] Anybots : Your Personal Avatar `http://www.anybots.com`.

[5] Takeshi Mizumoto, Takami Yoshida, Kazuhiro Nakadai, Ryu Takeda, Takuma Ohtsuka, Toru Takahashi, Hiroshi G. Okuno: Design and Implementation of Selectable Sound Separation on a Texai Telepresence System Using HARK in *Proc. of IEEE-RAS International Conference on Robotics and Automation(ICRA), pp.2130-2137, 2011.*

[6] Kazuhiro Nakadai, Toru Takahashi, Hiroshi G. Okuno, Hirofumi Nakajima, Yuji Hasegawa, Huroshi Tsujino: Design and Implementation of Robot Audition System ”HARK” in *Advanced Robotics, vol.24 pp.739-761, 2010.*

[7] HARK Main Page: `http://winnie.kuis.kyoto-u.ac.jp/HARK/`.

[8] Morgan Quigley, Brian Gerkey, Ken Conley, Josh Faust, Tully Foote, Jeremy Leibs, Eric Berger, Rob Wheeler, Andrew Ng: ROS: an open-source Robot Operating System in *Proc. of IEEE-RAS International Conference on Robotics and Automation (ICRA) Workshop on Open Source Software in Robotics, 2009.*

[9] ROS: `http://www.ros.org/wiki/`.

[10] Iwaki Toshima, Sigeaki Aoki, Tatsuya Hirahara: An acoustical tele-presence robot: TeleHead ll in *Proc. of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS), pp2105-2110, 2004.*