テレプレゼンスロボットの身ぶりが発話交替に与える影響について

Effects of Telepresence Robot's Gestures on Turn Taking

長谷川 孔明 [1]　中内 靖 [1]

Komei Hasegawa[1], Yasushi Nakauchi[1]

[1] 筑波大学大学院 システム情報工学研究科

[1]Graduate School of Systems and Information Engineering, University of Tsukuba

Abstract: In this research, we propose a telepresence robot for avoiding speech contentions occurred in remote convesations.In face-to-face convesations, humans predict the next speaker by gestures of other participants.However, it is difficult to predict in using 2D video chat services.The failure of prediction causes speech contentions and makes conversations dull.To solve this problem, we propose the telepresence robot which can convey 3D gestures.We use a Kinect as the gesture input device and it enables to convey unconscious gestures.We conducted an experiment to confirm the robot's advantages.

1．はじめに

近年の通信技術の発達により，遠隔地にいる人と会話を行う手段としてビデオチャットが普及している．さらに，現在の通信技術では 1 対 1 ではなく複数人での会議を遠隔地間で行う Web 会議も可能となり広く利用されている．しかしながら，ビデオチャットや Web 会議といった映像と音声のみを用いた遠隔会話の場面では発話交替がうまくいかずに発話衝突という問題が生じる[1]．玉木らによると，Web 会議の場面では対面会議と比較して発話衝突が 30 倍近く起こると報告している．発話衝突が起こると発話をあきらめて中断する傾向が高く，沈黙が発生し消極的な会話になりかねない．また，有意義な意見が発話衝突により妨げられる可能性もある．さらに発話衝突が頻発すれば会話の中断により会議時間が無駄に長くなるという問題にも繋がる．

発話衝突の原因として発話予備動作が認知されにくくなるという点があげられる[2]．対面での会話では，参与者の発話予備動作を読み取ることにより誰がいつ発話し始めるのかを判断している．これにより発話衝突を回避し，円滑な発話交替を実現している．しかしながら，Web 会議の場面では発話予備動作が認知されにくいことにより発話衝突が起こりやすいと示唆されている．

映像では認知されにくくなる発話予備動作を 3 次元的な実体を持つテレプレゼンスロボットを用いて伝達することにより，発話衝突を回避することが出来ると考えられる．さらに我々は意識的身ぶりと無意識的身ぶりというものを定義し，無意識的身ぶりも表出することが発話衝突回避に有効であると考えた．そこで本研究では無意識的身ぶりを表出可能なテレプレゼンスロボットを提案する(図 1)．このテレプレゼンスロボットは発話予備動作を表出する方法として身ぶりの操作方法に着目して設計を行った．そして，無意識的身ぶりが発話衝突回避に有効であることを確かめるための実験を行った．

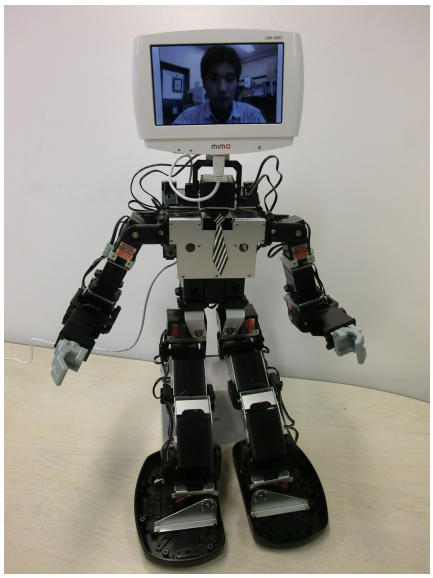

図 1 無意識的身ぶり表出
テレプレゼンスロボット

２．関連研究

既存の遠隔対話を支援するロボットの研究は，ロボットを用いてノンバーバルコミュニケーションを実現することを目的としている．人同士が行うコミュニケーションは，バーバルコミュニケーションとノンバーバルコミュニケーションに分類される[3]．バーバルコミュニケーションは言語を用いたものである．一方，ノンバーバルコミュニケーションは周辺言語や視線，表情，対人距離，身ぶりといった言語以外の手段を用いるコミュニケーションである．さらにバーバルコミュニケーションよりもノンバーバルコミュニケーションによってより多くの情報が伝達されると言われている[4]．そのため遠隔会話でも対面時のようにノンバーバルコミュニケーションを実現することが重要となる．

発話衝突回避のために重要となる発話予備動作も，何かしらの動作として相手に伝わるためノンバーバル情報の一部と捉えることが出来る．Marjorie が挙げている発話を獲得するための動作例には,「腕組みをほどく」「身を前に乗り出す」「話者の方へ向きを変える」などがある[5]．また，玉木らは発話予備動作を伝達する媒体として，手，頭，頷き，視線，音声を挙げている[2]．これらのことから発話予備動作には，頭部，体幹，腕部の動きが関わっていると考えられる．このことを踏まえた上で関連研究について紹介する．

既存の遠隔会話支援ロボットとして，Paulos らはロボットに搭載したディスプレイに操作者の顔を表示させる PRoP と呼ばれるロボットを開発し，テレプレゼンスについて報告している[6]．このロボットはノンバーバル情報である表情と対人距離の表出が可能である．しかしながら，発話予備動作に関わる頭部，体幹，腕部の動きが表現できないと考えられる．

また，鈴木らは遠隔会議支援ロボットシステムを開発している[7]．これは卓上サイズの移動ロボットにパンチルトカメラを搭載した構成となっている．カメラのパンチルト動作と移動を組み合わせることにより注意喚起能力を高められることを報告している．このロボットはノンバーバル情報である視線方向と対人距離の表出が可能である．そのため，発話予備動作に関連する頭部動作は表現できるが，体幹，腕部の動きが表現できないと考えられる．

Adalgeirsson らは MeBot と呼ばれる卓上サイズの移動ロボットを開発している[8]．MeBot は顔の映像を表示するパンチルトと前後移動が可能なディスプレイを頭部として取り付けている．また，各 3 自由度の腕の機構を有している．このロボットはノンバーバル情報である表情，視線方向，対人距離，身ぶりの表出が可能である．そのため，発話予備動作に関する頭部，体幹，腕部の全動作が表現できる．MeBot は発話予備動作を表出するモダリティとしては十分だと考えられる．しかしながら，MeBot の身ぶりの操作方法は MeBot の腕と同じ機構のコントローラで操作するという方法である．この方法では発話予備動作を操作に反映できないと考えられる．我々がこのように考える理由については次章で詳しく説明する．また，MeBot を評価する実験では，１対１の会話タスクにおいてロボットが動く場合と動かない場合についての印象の違いを主に評価している．そのため発話衝突や発話交替，発話予備動作については一切評価していない．そのため MeBot の操作方法によって表出される身ぶりが発話衝突回避に有効であるかどうかは明らかになっていない．

３．システム設計

3.1 意識的身ぶりと無意識的身ぶり

ここで，本研究で扱う「意識的身ぶり」と「無意識的身ぶり」についての定義を行う．

まず「意識的身ぶり」は，相手に見せるという目的で表出した身ぶりと定義する．例えば,「あそこの」と言いながら指さしをするといったものや「ボールがこう飛んで来て」と言いながら手を握って拳をつくることによりボールを表現し，ボールの飛ぶ起動を手の動きで表すといったものである．

次に「無意識的身ぶり」は．相手に見せるという目的は無く表出した身ぶりと定義する．例えば，頭をかく，口元に手をやる，身を乗り出すと言ったものである．

前項で紹介した MeBot の身ぶりの操作方法は，ロボットと同じ機構のコントローラを動かすことによって行う．そのため，操作者が意識的に表出しようとした身ぶりのみが操作として現れ，ロボットの身ぶりとして表出されると考えられる．そして，この操作方法では人の癖などの無意識的な身ぶりは表出されないと考えられる．

しかし，玉木らが挙げている発話予備動作の例には，「身を前に乗り出す」「手を口および顔周辺へ持っていく，もしくはそこから下ろす動作」といった人が無意識的に行っている動作が多く見受けられる．さらに，Cassell らは，人は対面コミュニケーションの場面において，話し手が意識的に身ぶりを表出している場面でなくとも，聞き手は常に身ぶりから情報を得ていることを示唆している[9]．よって，人が無意識的に行っている身ぶりも相手に影響を与えておりコミュニケーションにおいて何かしらの役

割を果たしていると考えられる．

これらのことから，無意識的身ぶりには発話予備動作となるものが含まれていると推察できる．そして，無意識的身ぶりを表出することにより発話衝突の回避を行うことが期待できる．そこで本研究では，テレプレゼンスロボットに無意識的身ぶりを表出させることにより，発話衝突を回避するシステムを実現する．

3.2 無意識的身ぶりの取得方法

無意識的身ぶりを表出するためには，まずその身ぶり自体を取得する方法が必要となる．身ぶりを意識的か無意識的かに分別し，無意識的身ぶりのみを抽出するような方法は困難である．

しかしながら，無意識的身ぶりも何かしらの体の動作である．そのため操作者のすべての動作を取得すれば，意識的か無意識的かの分別は出来ないが，無意識的身ぶりを含んだ動きをロボットの操作に反映することが可能である．

そこで，操作者の動作を取得する方法としてモーションキャプチャの技術を利用する．これにより無意識的なものも含めた身ぶりを取得することが可能である．

4． システム実装

4.1 システム全体構成

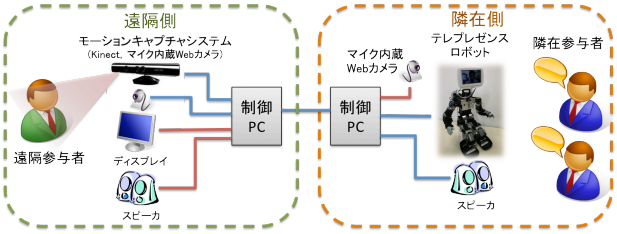

図 2 システム構成図

実装したテレプレゼンスロボットのシステム構成図を図 2 に示す．

遠隔側では，遠隔参与者の頭部，体幹，腕部の動作をモーションキャプチャで取得し，遠隔参与者の表情と音声は Web カメラとマイクにより取得する．遠隔参与者側で取得した頭部，体幹，腕部の動作，表情，音声の情報は制御 PC によりインターネットを介し隣在側の制御 PC へと送られる．この通信の実装は SkypeAPI を用いて行った．隣在側では受け取った情報をもとにテレプレゼンスロボットにより遠隔参与者の身ぶりと表情を隣在参与者へと提示する．また，音声は制御 PC のスピーカを用いて提示する．

隣在側から遠隔側への情報伝達は既存の Web 会議と同じ構成である．隣在側では，テレプレゼンスロボットの背後に設置した Web カメラとマイクからそれぞれ隣在参与者の映像と音声を取得する．映像と音声は制御 PC によりインターネットを介して遠隔側の制御 PC へと送られる．遠隔側では，映像はディスプレイを，音声はスピーカを用いて操作者へと提示される．

4.2 テレプレゼンスロボット

実装したテレプレゼンスロボットの外観を図 1 に示す．ロボットのプラットフォームとして近藤科学株式会社製ヒューマノイドロボット KHR-3HV を用いた．身ぶりの表現度を高めるためサーボモータの追加を行った．本研究では着座状態での会話で用いることを想定し，脚部の動きは表現しないものとした．よって実際に身ぶりの表現に用いるのは，頭部 3 自由度，腕部各 4 自由度，腰部の左右の捻り 1 自由度，身の乗り出し仰け反り 1 自由度である．これにより遠隔参与者の頭部と体幹，腕部動作を隣在参与者へと表出する．

遠隔参与者の表情を表出については，テレプレゼンスロボットの頭部に取り付けた小型ディスプレイを用いる．このディスプレイに遠隔参与者の顔の映像を映し出すことにより，表情の表出を行う．本研究では，テレプレゼンスロボットの顔として違和感ができるだけ少なくなるよう大きさを考慮した．

4.3 操作方法

4.3.1 モーションキャプチャ操作

体幹と腕部の動作を取得するためのモーションキャプチャデバイスとして Microsoft 社製の Kinect を用いた．Kinect は同社のソフトウェア開発キット Kinect for Windows SDK を用いることにより深度センサーからの情報をもとに人物のボーンを認識し，体の動きを各関節や手先位置などの 3 次元情報として取得することが可能である．Kinect はモーションキャプチャデバイスとしては比較的安価である．また，一般的なモーションキャプチャシステムは動作を取りたい人の体に専用のマーカを取り付ける必要がある．それに対し Kinect は，体にマーカをつけることなく動作を取得でき，手軽に運用できるという特徴がある．そのため今回はモーションキャプチャデバイスとして Kinect を用いた．

頭部動作の取得には Web カメラと Seeing Machines 社の faceAPI を用いた．faceAPI は，Web カメラからの画像をもとにリアルタイムな 3 次元フ

ェーストラッキングを行うことが可能である．

4.3.2 コントローラ操作

提案するシステムの比較対象として，コントローラ操作により身ぶりを表出するシステムの実装を行った．システム全体の構成と用いるテレプレゼンスロボットはモーションキャプチャ操作を行うシステムと同一のものであるが，体幹と腕部の操作方法のみをコントローラ操作に変更する．コントローラ操作を用いたシステムは，関連研究の章で取り上げた MeBot の操作方法を参考にして実装を行った．具体的には，テレプレゼンスロボットの腕の動きについては，ロボットと同じ機構を持つコントローラによりマスタ・スレーブ方式で操作し，頭部の動きについては faceAPI を用いた操作を行う．

コントローラとしては，テレプレゼンスロボットと同様にサーボモータを追加し頭部のサーボモータのみを取り除いたもう一台の KHR-3HV を用いた．

５．実験

5.1 実験目的

この実験の目的は，テレプレゼンスロボットの表出する無意識的身ぶりが発話衝突回避に有効であるかどうかを明らかにすることである．

5.2 実験方法

5.2.1 比較条件

本実験では，「コントローラ操作条件」と「モーションキャプチャ操作条件」の２種類を比較条件とする．「コントローラ操作条件」では意識的身ぶりのみを表出すると考えられる．一方，「モーションキャプチャ操作条件」では意識的身ぶりに加えて無意識的身ぶりも表出されると考える．これらの比較により無意識的身ぶりが発話衝突や発話交替に与える影響を検証する．

5.2.2 実験環境

発話衝突は複数人が参加する遠隔会話で発生しやすい．そのため本実験では，３人参加の多人数会話を想定し，そのうちの１人が遠隔参与者として遠隔地からロボットを操作して会話に参加する．

また，システムの動作テストを行った際に，ネットワークの不安定によりロボットの動作にラグが発生することが分かっていた．ラグが会話に与える影響を出来る限り少なくし，操作条件の違いが会話に与える影響を捉えやすくする実験設定が望ましい．そのため，本実験ではネットワークを介しての通信は行わず，隣在参与者と同一の部屋にて遠隔参与者がロボットの操作を行うものとした．その際に，隣在参与者と遠隔参与者の間に衝立を設け，互いの姿が直接見えないように配慮した(図３)．

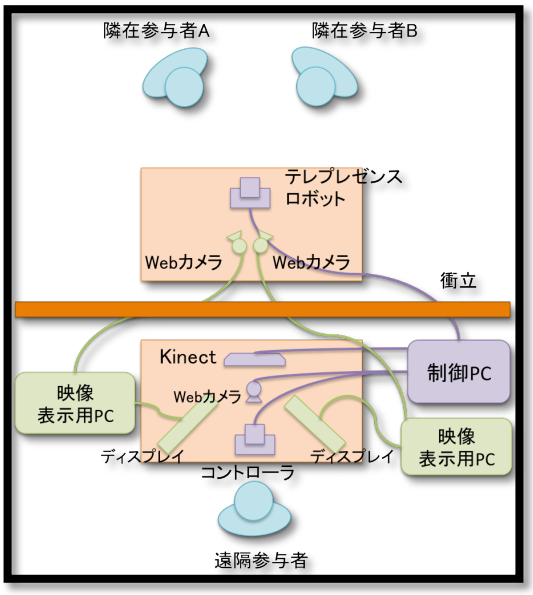

図 3 実験環境図

5.2.3 会話タスク

本実験では，会話内容を統制するために会話タスクとして砂漠生き残り問題を用いた．砂漠生き残り問題とは，砂漠で遭難しているという状況を想定し，リストアップされた道具について生き残るために必要なものの優先順位を議論により決定するというタスクである．複数の道具の組み合わせを用意し，４分間の短い会話タスクを複数回行う．また，遠隔参与者となる被験者と操作条件をそれぞれ変更して各会話タスクを行う．

5.3 評価方法と仮説

実験タスクの会話の様子を観察することによりモーションキャプチャ操作条件とコントローラ操作条件での会話の評価を行う．評価の指標としては，会話中の発話衝突の回数と参与者の発話ターン取得率を用いる．発話ターン取得率は，会話中における全発話ターンのうち，ある参与者が獲得した発話ターンの割合とする．

コントローラ操作条件では意識的身ぶりしか表出されないが，モーションキャプチャ操作条件では意

識的身ぶりに加えて無意識的身ぶりも表出され，その身ぶりが発話予備動作の役割を有していると考えられる．そこで，以下の仮説を立て検証を行う．

仮説 1：モーションキャプチャ操作条件はコントローラ操作条件に比べて発話衝突の回数が少ない

仮説 2：モーションキャプチャ操作条件はコントローラ操作条件に比べてロボットを介した参加者の発話ターン取得率が高い

これらの仮説が検証されれば，モーションキャプチャを用いた操作方法により無意識的な身ぶりを表出でき，その身ぶりが発話衝突の回避と円滑な発話交替に有効であることが明らかになると考える．

5.4 実験結果

図 4 実験の様子(隣在側)

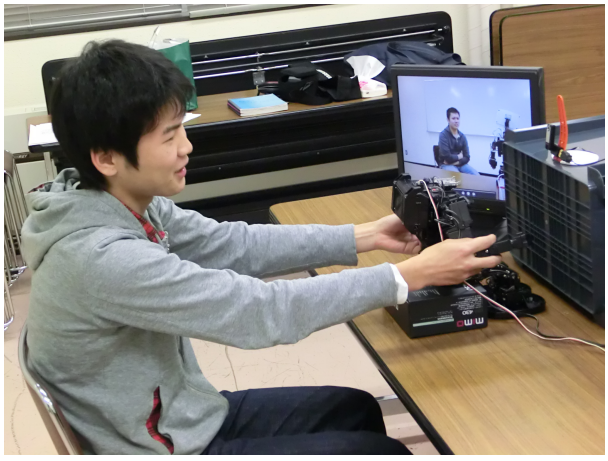

図 5 実験の様子(遠隔側)

3 組（合計 9 人）の被験者に対して実験を行い，ビデオカメラを用いて実験の様子を録画した．今回実験を行った被験者は 20 歳から 25 歳までの 9 人(男性 8 人，女性 1 人）であった．実験の様子を図 4，図 5 に示す．実験中のシステムトラブルにより，想定した実験条件での会話が行えなかった会話タスクが 1 度だけあった．そのため，その被験者 1 人のデータを除いて実験結果の解析を行った．

各操作条件における遠隔参与者に関わる発話衝突回数を表 1 に示す．この表は，被験者 A〜H が遠隔参与者としてそれぞれの操作条件で実験タスクを行った際に，発話衝突が起こった回数を示している．モーションキャプチャ操作条件とコントローラ操作条件における発話衝突回数に有意な差があるのかどうかを確かめるため Mann-Whitney の U 検定を行った．検定結果，有意水準 5%で 2 つの操作条件における発話衝突回数間に有意な差があることがわかった．

各操作条件における遠隔参与者のターン取得率を表 2 に示す．ターン取得率に関しても Mann-Whitney の U 検定を行ったが，有意な差はみられなかった．実験の結果からターン取得率は個人による差が大きいことがわかる．ターン取得率の個人差を排除するため，ロボットを介さずに 3 人とも対面で実験タスクを行った対面条件でのターン取得率を基準とし，各操作条件でのターン取得率との差を算出した．その算出結果を表 3 に示す．これについても同様に検定を行ったが，有意な差はみられなかった．

表 1 各操作条件における発話衝突回数

	遠隔参与者								
	A	B	C	D	E	F	G	H	中央値
モーションキャプチャ操作条件における発話衝突回数[回]	3	1	1	5	1	2	1	0	1
コントローラ操作条件における発話衝突回数[回]	8	4	4	3	1	4	3	3	3.5

表 2 遠隔参与者のターン取得率

	遠隔参与者								
	A	B	C	D	E	F	G	H	中央値
モーションキャプチャ操作条件におけるターン取得率	0.308	0.133	0.411	0.268	0.271	0.429	0.13	0.39	0.29
コントローラ操作条件におけるターン取得率	0.327	0.256	0.476	0.244	0.2	0.314	0.104	0.292	0.274

表 3 対面条件でのターン取得率との差

	遠隔参与者								
	A	B	C	D	E	F	G	H	中央値
モーションキャプチャ操作条件と対面条件とのターン取得率の差	−0.041	−0.03	−0.077	−0.045	0.042	−0.031	−0.03	0.01	−0.031
コントローラ操作条件と対面条件とのターン取得率の差	−0.022	0.093	−0.012	−0.069	−0.029	−0.146	−0.056	−0.088	−0.043

5.5 考察

仮説 1 の「モーションキャプチャ操作条件はコントローラ操作条件に比べて発話衝突の回数が少ない」については実験結果から，モーションキャプチャ操作条件の発話衝突回数が有意に少ないことが確認されたため，仮説は立証されたといえる．このことか

ら，モーションキャプチャ操作条件は発話衝突の回避に有効であるといえる．ただし，今回の実験は被験者数が少なかったため，より信頼のおける結果を得るためにはより多くの被験者を対象に実験を行う必要がある．

また，仮説 2 の「モーションキャプチャ操作条件はコントローラ操作条件に比べてロボットを介した参加者の発話ターン取得率が高い」については，実験結果から仮説の立証はされなかった．実験結果から，モーションキャプチャ操作条件とコントローラ操作条件のターン取得率はどちらも，対面条件のターン取得率より減少する傾向にあることがわかった．そのため，ターンの取得率は操作条件だけではなく他の要因が考えられる．要因の１つとして，遠隔参与者の得る隣在側の情報が映像と音声のみの既存の遠隔会話と同じものであることがあげられる．そのため遠隔参与者が隣在参与者の発話予備動作を見逃している可能性が高い．このことが要因となり操作者がターン取得の機会を逃していたと考えられる．

６．おわりに

音声と映像のみの遠隔会話の場面では発話予備動作が認知されにくくなり，発話衝突という問題が発生する．そこで，３次元的な実体を持つロボットを介して身ぶりを表出することにより身ぶりが認知されやすくなると考えた．また，身ぶりの中でも無意識的に行っている身ぶりが発話予備動作を有し話者交替において重要であると考え，無意識的身ぶりを表出可能なテレプレゼンスロボットを提案した．

意識的身ぶりのみを表出するコントローラ操作条件と，提案手法である無意識的身ぶりも表出するモーションキャプチャ操作条件を比較する実験を行い，提案手法が発話衝突を減少させ話者交替を円滑にすることを確認した．

今後は被験者数を増やし，より信頼のおける実験を行う予定である．また，表出される身ぶりの傾向や頻度が操作条件によってどのよう異なるのかについて実験の映像より解析を行う予定である．

参考文献

[１] 玉木秀和, 中茂睦裕, 東野豪, 小林稔: 人のコミュニケーションリズムに着目した Web 会議円滑化手法, IEICE Technical Report MVE2009, pp.101-106, (2009)

[２] 玉木秀和, 東野豪, 小林稔, 井原雅行: Web 会議における話者交替円滑化手法の検討, 画像電子学会ＶＭＡ研究会, Vol.29, pp.9-18, (2011)

[３] マジョリー・F・ヴァーガス 著, 石丸正 訳: 非言語コミュニケーション, 新潮社, (1987)

[４] アルバート・マレービアン 著, 西田司 他共訳: 非言語コミュニケーション, 聖文社, (1986)

[５] 高橋誠: 会議の進め方,日本経済新聞出版社,(1987)

[６] E. Paulos, J. Canny: Social Tele-embodiment: Understanding Presence, Autonomous Robots, Vol.11, No.1, pp.87-95, (2001)

[７] 鈴木雄介, 福島寛之, 深澤伸一, 竹内晃一: 遠隔会議支援ロボットシステムの注意喚起能力評価, 情報処理学会論文誌, Vol.51, No.1, pp.25-35, (2010)

[８] S.O. Adalgeirsson, C. Breazeal: MeBot: A Robotic Platform for Socially Embodied Telepresence, HRI2010, pp.15-22, (2010)

[９] J. Cassell, D. McNeill, K. E. McCullough: Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information, Pragmatics and Cognition, Vol.7, No.1, pp.1-33, (1999)