第 28 回日本バーチャルリアリティ学会大会論文集（2023 年 9 月）

ソーシャル VR において多人数会話に参加する会話エージェントのための LSTM を用いた自然な身体動作の生成

Behavior generation of social VR conversation agent participating in multi-party conversation using LSTM

加藤圭悟 [1)]，長谷川晶一 [1)]
Keigo Kato, Shoichi Hasegawa

1) 東京工業大学 工学院情報通信系（〒 226-8503 神奈川県横浜市緑区長津田町 4259 R2 棟 624 室 R2-20, info@haselab.net）

概要: ソーシャル VR の利用者が増えているいま, そこでのユーザ同士の会話や交流をより豊かにする存在として, 自然な振る舞いが可能な会話エージェントの存在が期待されている. そのためには, 相手に適応して, 自身の非言語動作を適切に出力できる必要がある. そこで本研究では, ソーシャル VR における実会話事例を LSTM により学習して, リアルタイムかつ動的に身体動作を出力できる会話エージェントを提案する.

キーワード： コミュニケーション, ソーシャル VR, 非言語動作, 多人数会話

1. はじめに

1.1 背景

近年,VR 機器の普及が進み VRChat[1] などを代表としたソーシャル VR の利用者が増えてきている. ソーシャル VR はユーザ同士の交流の場として用いられるが、その利用シーンは雑談や展示会, 研修・セミナーなど多岐にわたる. また利用者が増える分だけ, そこで求められるサービスも多様化していくことが予想される. 例えば, 他ユーザとの雑談の際に隣にいて話を聞いてくれる存在が欲しい, 展示会の際にいちグループごとに質問対応ができる存在が欲しいなどである. しかし, こうしたサービスを提供してソーシャル VR での交流をより豊かなものにしていくには, そうしたサービスが提供可能な存在を投入する必要がある.

そこで以前から注目されているのが,Non Player Character（NPC）である. その中でも会話に参加する NPC は Embodied Conversation Agent（ECA）と呼ばれている.ECA は直接的にユーザと関わり, 前述したサービスなどが提供可能なため, 交流をより豊かにする存在として可能性が高いと考えられている [2]. しかし,ECA はその存在感の弱さが問題になっている.Iwasaki ら [3] は, その問題に焦点をあて,ECA がユーザの行動を理解して応答を返すと, 会話がより促進されることをフィールド実験を通じて紹介している. また神田 [4] は，ECA に人らしさを感じることで,「誰かと関わりたい」という気持ちに似た「ECA と関わりたい」という動機づけがユーザに起きることを，事例を通じて紹介している. つまり ECA は相手に適応して行動できることが重要である.

また ECA の研究では, 相手のエンゲージメントに適応して行動する, エンゲージメントモデリングが注目を集めている. エンゲージメントとは関心や注目の度合いを意味する. 普段, 人は頭や視線, 表情, 姿勢など多くの非言語動作 [5,6,7,8] を通じ, エンゲージメントを暗黙的に理解して円滑な会話を行っている. またそういった非言語動作はユーザ同士だけでなく, ユーザと ECA の間のエンゲージメントを維持するためにも重要な役割を果たすことが分かっている [5]. そのため, ユーザ同士の交流をより豊かにするような ECA を作るには, 相手の非言語動作に適応して自身の非言語動作を動的に変化できることが重要である. そこで本研究では,ECA における非言語動作の出力モデルに注目する.

1.2 目的

非言語動作の出力モデルには大きく分けて, ルールベースと機械学習ベースの二種類ある. 前者は心理学の理論をもとに人手で作成されたモデルで, 一定期間における特定の動作の検出および生成に優れている. しかし普段, 人は相手の連続的に生じる多様な動作を理解して動いている. そこで近年では、連続的に動作の検出および生成が可能な後者のモデルが増えている. しかし, そのアルゴリズムによっては, ある程度決まった非言語動作のみしか検出および生成できない. そこで連続値の時系列データを扱えるアルゴリズムの Long Short Term Memory（LSTM）[9] を用いれば, より多様な非言語動作を検出および生成できると考えた. 本研究では, 動作を連続値として捉えることにより, 多様な非言語動作の検出および生成ができる ECA の実現と, その印象調査を目的とする.

2. 関連研究

2.1 ルールベース

Aburumman ら [10] は, ソーシャル VR における二者会話において、適切なタイミングで頷くことのできる ECA を

作成した. この ECA には Hale ら [11] が発見した,「会話において人は, 相手の頷きの 600ms 後に, それを模倣する形で頷き返す」という心理学の理論が実装してある. その結果, 頷きの大きさやタイミングについては適切で, 人からより好感と信頼を得ることができている. しかし, 視線や表情などの他の箇所も同様に心理学の理論を用いて共に実装したとき, 動きやタイミングがかみ合わずに不自然さを感じることがある.

2.2 機械学習ベース

近年では, 機械学習ベースのモデルを採用した ECA が増えている. その中でも, 相手のエンゲージメントや次の行動の予測ができる ECA[12,13] は多いが,ECA 自身のアニメーションの生成まで視野に入れている研究は少ない.

Zou ら [14] は、ソーシャル VR における三者会話において, フレームごとに相手と自身の注視状態を検出し、自身の次のフレームでの注視状態をリアルタイムで生成する ECA を作成した. この ECA は時系列や, 自分の行動も顧みて反応できる, いわゆるインタラクションループも考慮しているため, より適切な反応ができている。しかし用いているアルゴリズムが Hidden Markov Model（HMM）で, 非言語動作を離散値として捉えているため, ある程度決まったものしか検出および生成できない.

そこで最近では、非言語動作を離散値ではなく連続値として捉えて, 多様な動作の検出および生成を可能にした研究 [15] もある. この研究では現実世界における ECA との二者会話において,LSTM を用いて時系列とインタラクションループを考慮して, フレームごとに相手と自身の頭の回転, 視線の方向, 笑顔を検出して, 自身の次のフレームでの同特徴量の生成する ECA が作成された. しかし, 想定している環境が二者会話のためか, 被験者は ECA から注目をずらすことはほとんどなく, 非言語動作を連続値として捉えたときの効果を十分に測ることができなかった.

そこで本研究では、ソーシャル VR の多人数会話において、時系列とインタラクションループを考慮してフレームごとに非言語動作を連続値として検出および生成する ECA の実現と印象調査を目的とする.

3. 提案手法

人同士が会話で行う自然な非言語動作のやりとりを再現するため,ECA の非言語動作の出力モデルを実会話事例から学習することで得る.

本研究では, 三者会話における傍参与者の立場にたつ ECA を対象とする. 傍参与者とは、やりとりには加わっていないが, 会話は聞いていて, 将来, 話者または受話者になる可能性がある人を指す [16]. この条件にした理由は大きく三つある. 一つ目は非言語動作に焦点を当てるためである. 二つ目は, 人をつないで交流をより豊かにする ECA を想定した場合, 傍参与者の立場でも貢献できると考えたからである. 三つ目は, 三者会話は二者会話よりも複雑な相互作用が行われて、非言語動作を連続値として捉えたときの効果を十分に調査できると考えられるからである.

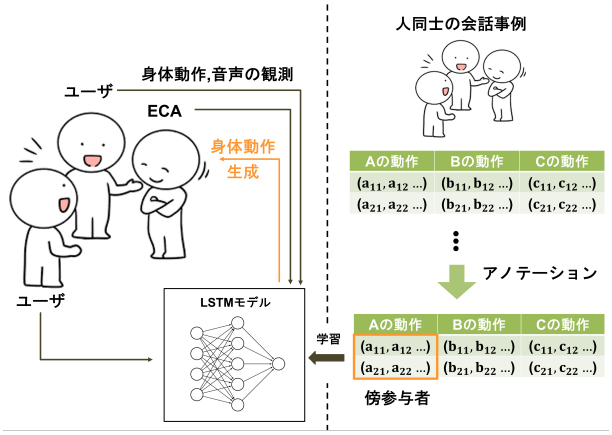

図 1: システム全体像

そして今回の目的を満たす ECA の作成方法は以下の通りである. まず実会話事例における頭部動作や音声等の時系列データを収集し, 次にそれを学習に利用できるように, 誰が傍参与者であるかというアノテーションを行う. その様にして得られた学習データを LSTM で学習し，獲得した学習済モデルを用いることでインタラクティブに非言語動作を決定する. 全体像を図 1 に示す.

3.1 実会話事例の収集

今回の研究ではソーシャル VR における ECA を対象としているため、会話データもソーシャル VR におけるものを収集する. これはソーシャル VR における非言語動作のニュアンスが、現実世界のそれと比較して強調されるという調査結果 [17] があるためである. そこで VRChat[1] を利用して, 会話の様子を記録する実験を行った. このとき, 雑談会と称して参加者を募り, ワールド上での三者会話の様子を一回あたり 30~60 分程度記録した. 参加者は研究室内から募り, 会話において優位性が生じないようにした. また実験では, 三点トラッキング以上かつヒューマノイドアバター使用した. データ収集は東京工業大学倫理審査委員会の承認を経て, 説明同意を得られた方を対象に行った.

3.1.1 実験環境

実験参加者には各々, 自宅や研究室などのインターネットがつながる環境からワールドに参加してもらった. この際, 特に環境は指定しなかった.

3.1.2 会話

会話中の様子を図 2 に示す. この際, 三人のうちの誰かに傍参与者の役割をするように指示してはおらず、自然な会話を行ってもらった. 会話内容はタスクを設定せず, また特定のオブジェクトを対象にする会話は行っていない. そして会話中にワールド内を移動するのは自由とした.

3.1.3 収集データ

会話は 0.02 秒ごとに記録し, 現時点で約 3 時間, すなわち約 540000 フレームの時系列データを得た. 現在のコンシューマー向け HMD によって課される制限のため, 会話参加者からは頭, 左手, 右手の位置と傾き, 音声の音量とピッチを記録した. またアバターの位置と回転も記録した. それぞれ

図 2: 会話中の様子

表 1: 記録内容

記録対象	記録内容
位置 (頭, 左右の手)	(x,y,z)
傾き (頭, 左右の手)	(w,x,y,z)
音量	RMS（0~1）
ピッチ	AMDF（50~500 [Hz]）

の対象の記録内容を表 1 に示す. 位置は三次元座標, 傾きはクォータニオン, 音量は Root Mean Square（RMS）を正規化した値, ピッチは Average Magnitude Deference function（AMDF）[18] を用いて 50~500 Hz の範囲で推定した値を記録した. 音声情報に関しては, 頭の動作と強く関係されるとされる音量とピッチ [19] を選択している. またピッチは短時間計算が可能な AMDF を選択した. 範囲も人声におけるピッチ範囲 [20] を参考にした. このとき, 音声以外は VRChat から, 音声は別個で開発した音声記録ウェブアプリを通じて収集を行った.

3.2 収集データへのアノテーション

収集した時系列データのうち, 誰が傍参与者であるかを算出してアノテーションを行い, 学習可能なデータへと変換する必要がある.

ここでは, 矢野ら [21] が定義した多人数会話における近似的な参与役割の算出方法を参考にした. 音量が 0.1 以上かで発話の有無を決め, 有無を 1 と 0 の値とみて 100 個ずつガウス加重移動平均をとり, そのなかで時間当たりの発話量が 0.25 以上かつ最も発話量が多い者を話者とした. 現話者の受話者は次話者の者とし, その他を傍参与者とした.

3.3 LSTM による学習

会話における非言語動作は, 相手のそれを受けて行われるため, 会話データは時系列データとして扱う. またその動作は, 過去の行動を顧みて変容させると考えられる. そこで本研究では, 時系列データを扱え, インタラクションループも考慮でき、その上で今回の目的である, 入出力ともに連続変数を扱える LSTM を用いた. そして前節で生成した学習データのうち, 各会話参加者の距離が 1.8 m 以内で正面から 90 度以内の位置関係の会話形態を抜き出し, その上で傍参与者が 10 秒以上変化しない期間のデータを学習させた. 入力データは話者・受話者のデータ, 正解データは傍参与者の

表 2: 学習条件

記録対象	記録内容
損失関数	MSE
最適化手法	SGD（lr=0.01）
隠れ層の次元	10
入力長	50
バッチサイズ	1024
エポック数	1000

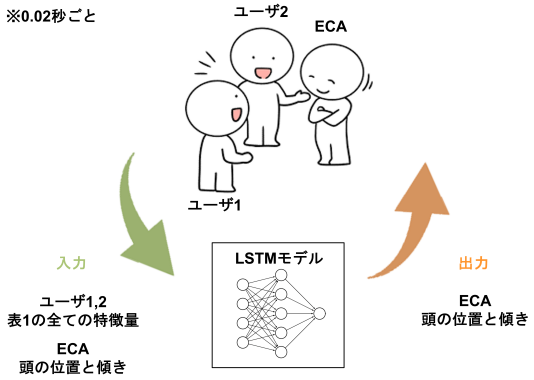

図 3: LSTM による身体動作生成

データとして学習させた. また入力データには表 1 で示した全ての特徴量が, 正解データには頭の位置と傾きが含まれる. なおこのときの学習条件を表 2 に示す. 学習は Pytorch[22] を使用して行ったが, 表 2 以外の条件は全てデフォルトの設定で行った.

3.4 学習済モデルによる身体動作生成

前節で得られた学習済モデルを用いて, インタラクティブに ECA の身体動作を生成する手法の概要を図 3 に示す. まず, 学習に用いた時系列データと同じ時間刻みごとに,ECA 以外の二者から表 1 で示した全ての特徴量, さらに ECA 自身の頭の位置と傾きを取得して, これを学習済モデルに入力する. モデルはその値に応じて, 次時刻での ECA の頭の位置と傾きを生成する. その値を,ECA のそれに代入することでインタラクティブな身体動作を実現する.

4. 実装

Unity[23] を使用し, 提案手法による傍参与者としての ECA を交えて会話を行うデモを作成した. データ収集時と同様に 3 体のキャラクタモデルを配置し, そのうち 1 体に前章で作成した LSTM を使ったインタラクションシステムを組み込んだ. 残りの 2 体は, テストデータとして設定していた収集データのうち, 傍参与者以外の二者のものを用いて動作させる. 動作の様子を図 4 に示す.

5. 今後の課題

本研究では, ソーシャル VR において三者会話に参加する ECA がより多様な動作の検出および生成を可能にするため

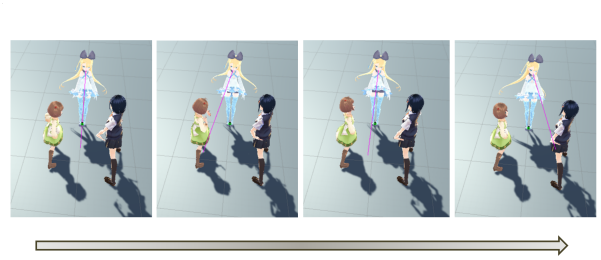

図 4: 動作の様子

に, フレームごとに動作を連続値として捉えることができる LSTM を用いたインタラクションシステムを提案した. 今後は, 学習モデルの精度を高めるための作業として, 学習条件の調整や学習量の増加, アノテーション方法の改良を行いたいと考えている. またその後, 今回の提案手法による ECA と, それ以外での ECA での比較実験を行うことで印象調査を行いたいと考えている.

参考文献

[1] VRChat.Inc, “VRChat”, https://hello.vrchat.com, (参照 2023-07-03)

[2] Unity Technologies, “自律キャラクタ研究から考えるソーシャル VR と NPC”, Unity Learning Materials, https://learning.unity3d.jp/8906, (参照 2023-07-03)

[3] M.Iwasaki et al., ““That Robot Stared Back at Me!”: Demonstrating Perceptual Ability Is Key to Successful Human–Robot Interactions”, Frontiers in Robotics and AI, vol6, September 2019

[4] 神田崇行, “ロボットに「人らしさ」を感じる人々 –フィールド実験での事例 – ”, 日本ロボット学会誌, 31 巻, 9 号, pp.860-863, 2013

[5] T.Fukuda et al., “Facial Expressive Robotic Head Systemfor Human–Robot Communication and ItsApplication in Home Environmen”, Proceedings Of The IEEE, VOl92, No11, pp.1851-1865, Novenber 2004

[6] Y.Nakano et al., “Estimating user’s engagement from eye-gaze behaviors in human-agent conversations”, Proceedings of the 15th international conference on Intelligent user interfaces, pp.139–148, February 2010

[7] G.Castellano et al., “Detecting user engagement with a robot companion using taskand social interaction-based features”, ICMI-MLMI’09, pp.119–126, Novenber 2009.

[8] S.Mota, R.Picard, “Automated posture analysis for detecting learner’s interest level”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol5, pp.49-49, June 2003

[9] S.Hochreiter, J.Schmidhuber, “Long Short-Term Memory”, Neural Computation, vol9, pp.1735-1780, November 1997

[10] N.Aburumman et al., “Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions”, International Journal of Human-Computer Studies, vol164, August 2022

[11] J.Hale et al., “Are you on my wavelength? Interpersonal coordination in dyadic conversations”, Journal of Nonverbal Behavior vol44, pp.63–83, October 2019

[12] S.Dermouche, C.Pelachaud, “Engagement Modeling in Dyadic Interaction”, 2019 International Conference on Multimodal Interaction, October 2019

[13] G.Dobre et al., “Immersive machine learning for social attitude detection in virtual reality narrative games”, Virtual Reality, vol26, pp.1519–1538, April 2022

[14] S.Zou et al., “Conversational agent learning natural gaze and motion of multi-party conversation from example”, Proceedings of the 5th International Conference on Human Agent Interaction, pp.405–409, October 2017

[15] S.Dermouche, C.Pelachaud, “Generative Model of Agent’s Behaviors in Human-Agent Interaction”, 2019 International Conference on Multimodal Interaction, pp.375–384, October 2019

[16] E.Goffman, “Forms of talk”, 1981

[17] S.Dermouche, C.Pelachaud, “Generative Model of Agent’s Behaviors in Human-Agent Interaction”, Proceedings of the ACM on Human-Computer Interaction, vol4, pp.1–25, October 2020

[18] M.Ross et al., “Average magnitude difference function pitch extractor”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol22, October 1974

[19] C.Busso et al., “Natural head motion synthesis driven by acoustic prosodic features: Virtual Humans and Social Agents”, Computer Animation and Virtual Worlds vol16, pp.283-290, July 2005

[20] C.Busso et al., “Vocal Attractiveness Increases by Averaging”, Current biology, vol20, pp.116-120, January 2010

[21] 矢野正治, 中田篤志, 福間 良平, “非言語マルチモーダルデータを用いた会話構造の分析のための環境構築”, 情報処理学会研究報告 1 巻, pp.1-8, 2009 年

[22] PyTorch, “PyTorch”, https://pytorch.org, (参照 2023-07-03)

[23] Unity Technologies, “Unity のリアルタイム開発プラットフォーム — 3D/2D、VR/AR のエンジン”, Unity, https://unity.com/ja, (参照 2023-07-03)