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

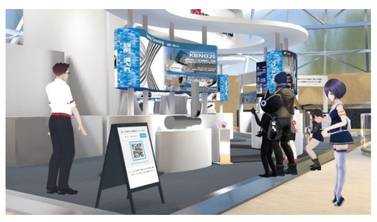
東京工業大学 工学院 情報通信系　加藤圭悟,長谷川晶一

1.背景

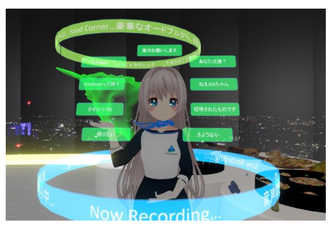
- **ソーシャルVRの利用シーン**
  - ユーザ同士の交流の場
  - 雑談や展示会,研修,セミナーなど多岐にわたる
- **求められるサービスの多様化**
  - Ex1.他ユーザとの雑談の際に隣にいて話を聞いてくれる存在
  - Ex2.展示会の際に一グループごとに質問対応ができる存在
    - コストを考えると,安易なサービス提供人材の追加は難しい
- **ソーシャルVRとNPC**
  - 会話に参加するNPCはEmbodied Conversational Agent (ECA)
  - ユーザの行動を理解した応答による会話の促進 [1]
  - ECAに人らしさを感じることで,ECAと関わりたいという動機の発生 [2]
- **ECAと非言語動作**
  - 人は頭や視線,表情,姿勢などを通じて,円滑な会話を実現
  - これらの動作は人とECAの間でも重要 [3,4,5,6]
    - 本研究では,ECAの**非言語動作の出力モデル**に注目

雑談の様子

展示会の様子 [7]



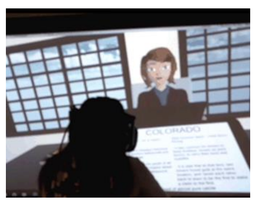
ECA [8]



2.関連研究

- **ルールベース**
  - 自然な反応のためのルールを心理学の理論等を用いて人手により作成する
  - VRにおいて,相手の頷きの600ms後,模倣して頷き返す理論が実装されてるECA [9]
    - 頷きは好感を得られているが,視線や表情と組み合わせると不自然
    - 自然な反応のための細かいルールを人手で作成することは難しい
- **機械学習**
  - 自然な反応のための細かいルールを学習により機械が作成する
  - 相手の次の動作やエンゲージメントなどの予測までの研究が多い [10,11]
  - VRにおいて,フレームごとに相手の注視方向と発話,自身の注視方向を検出して,自身の次のフレームでの注視方向を生成するECA [12]
    - 離散型HMMで動作を**離散値**として捉えているため,**決まった動作**の検出および生成に限定される
  - 現実において,フレームごとに相手と自身の視線,頭部回転,笑顔を検出して、自身の次のフレームでの同特徴量を生成するECA [13]
    - LSTMで動作を**連続値**として捉えているため,**多様な動作**の検出および生成が可能
    - 二者会話のために連続値の効果の不十分な検証
    - VRでなく現実

ルールベースECA [9]



ML(離散) ECA [12]

ML(連続) ECA [13]

3.目的

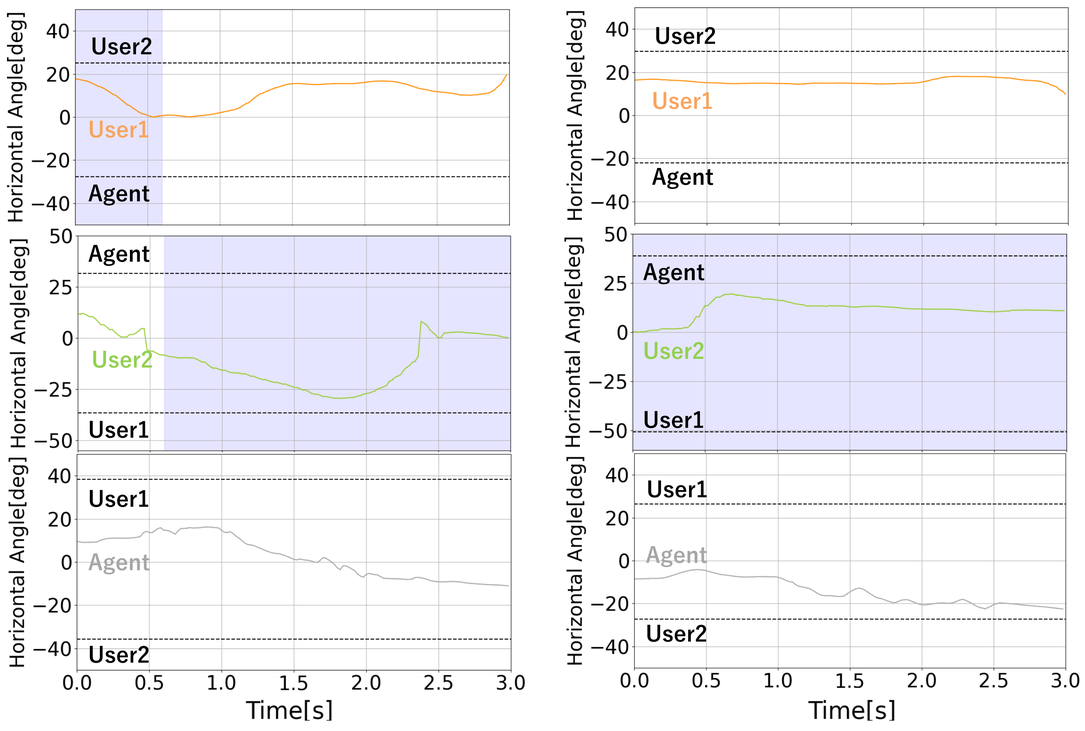
- **VRの多人数会話において、動作を連続値として検出・生成することで多様な動作に対応可能な人間らしいECAの実現**
  - 三者会話における傍参与者を対象とする
    - 非言語動作に焦点をあてるため
    - 人をつないで交流をより豊かにするために十分に貢献可能なため
    - 複雑な相互作用が行われて連続変数の効果を十分に検証可能なため

4.提案手法

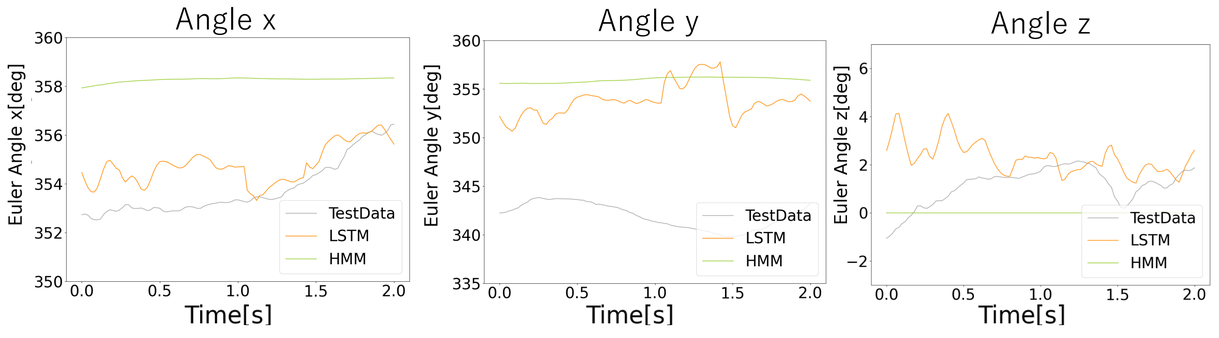
- **Step1: 会話データの収集**
  - ソーシャルVRのデータを収集(約234万フレーム)
    - 現実よりも動作のニュアンスが強調される [12]
    - 1回60分程の雑談会.参加者は研究室から集め優位性はない
    - VRモードかつヒューマノイドが対象
    - 0.02秒毎,頭の位置(x,y,z)と傾き(w,x,y,z),音量(rms(0~1))を収集
- **Step2: ラベリング**
  - 矢野らの近似的な算出方法を参考 [14]
    - 音量0.1未満は0、それ以上は1とラベリング⇒100個ずつガウス加重平均⇒時間当たり発話量が0.25を超え最大の者を話者
    - 次話者を現話者の受話者、最後は前話者を現話者の受話者
    - 話者と受話者以外を傍参与者とする
- **Step3: LSTMによる学習**
  - 時系列,インタラクションループを考慮できる
  - 以下を満たすデータを学習データとして採用(約120万フレーム)
    - 傍参与者が10秒以上は不変（ラベリングを確かにするため）
    - 各々、1.8 m以内かつ正面から90度以内（要は三角型）
  - 入力: 話者/受話者の頭の位置(x,y,z)と傾き(x,y,z),音量(rms(0~1))
  - 正解: 傍参与者の頭の傾き(x,y,z)
  - 3分割検証を行ってチューニング（以下,検証用データの値）
    - 平均R2: 約0.55(提案), 約-1.05(5フレーム平均のベースライン)
    - [13]（約62万フレーム）の平均R2: 約0.61 **※この計算は間違えている。上記より低い値となる。**
      - [13]は二者会話のために少ない学習量で汎化,更に好印象のECA
      - 本研究は倍以上のデータ量だが,R2スコアが悪い
      - 訓練用,検証用のR2の差の平均は0.1のためあまり過学習は起きてない
      - モデル自体は良いと考えられるが、データ量が少ないと考えられる

5.評価

- **学習データの再現** **※注視方向が変化する動作は確認できたものの、これは稀にのみ発生して、かつ大概は小さい動作**
  - 注視方向の変遷(時間[s]-顔向きの水平角度 [deg]) ※背景色の塗り潰し:話者
    - 基本的に中間を注視しつつ,細かな動作をする（データの質の問題）
      - 話者が変わった際に,話者または受話者を注視する（左図,話者）
      - 話者がエージェントを注視すると,注視し返す（右図）



  - 細かな動作表現（時間[s]–頭のオイラー角[deg]）
    - 提案モデル(LSM)では値が動くため,その間の頭の揺れや首の傾げを実現できている



6.参考文献

[1] M.Iwasaki et al., 2019 [2] 神田崇行., 2013 [3] T.Fukuda et al., 2004
[4] Y.Nakanoetal., 2010 [5] G.Castellanoetal., 2009 [6] S.Mota et al., 2003
[7] CNET Japan : https://japan.cnet.com/article/35195668/
[8] Advanced Media, Inc : https://www.advanced-media.co.jp/newsrelease/23695
[9] N.Aburumman et al., 2022 [10] S.Dermouche et al. 2019 [11] G.Cristina et al. 2022
[12] S.Zou et al., 2017 [13] S.Dermouche et al., 2019 [14]矢野正治ら., 2009