アバター会話支援のための音声からのリアルタイム会話モーション生成の検討

A Real-Time Conversational Motion Generation for Avatar Communication Support System

鈴木　颯真[1*]　上乃　聖[1]　李　晃伸[1]
Soma Suzuki[1]　Sei Ueno[1]　Akinobu Lee[1]

[1] 名古屋工業大学
[1] Nagoya Institute of Technology

Abstract: アバター会話では，操作者の動作をトラッキングしキャラクターに反映することで，身体性を伴うリアルなコミュニケーションが可能となるが，専用の機材や環境が必要である．本研究は，音声のみからアバターの上半身の会話動作をリアルタイム生成する Speech-to-Motion 手法を提案する．既存手法の多くは発話全体を参照するため低遅延処理が難しいが，本研究では言語履歴を短縮・省略することでリアルタイム化を図る．本発表では性能の変化を検証した結果を報告する．

1　はじめに

近年，アバターを用いたコミュニケーションが広がりを見せており，遠隔就労や教育分野などでの社会的応用が進んでいる．アバターはロボット型と CG 型など様々な形があるが，本研究では CG アバターに着目する．アバターを用いた会話には，音声のみのやり取りと，身体動作を伴う方法があり，後者は非言語情報の伝達により円滑なコミュニケーションを実現できる．特に，ジェスチャーは会話の理解を助け，共同作業の効率向上にも貢献することが示されている [1].

アバターの身体動作を実現するトラッキングの例として，トラッカーを用いるものやカメラのみを使う手法がある．しかし，トラッカーは装着の手間やプレイスペースが必要という制約があり、カメラによるトラッキングも、明るい場所では精度が低下するため、簡便な手法とは言い難い [2]．この問題を解決するために，本研究では音声からアバターのモーションを生成する「speech-to-motion」の手法を採用する．

speech-to-motion ではモーションの自然性を追求する研究が多く，リアルタイム性に関するものが少ない．また，アバター会話においてリアルタイム性は必要である．本研究ではこれらの点に着目し，既存の speech-to-motion モデルを改良することで，会話での利用に向けた低遅延なモーション生成を目指す．

本研究では，まずアバターを介したコミュニケーションの動向を整理し，次に音声から身体動作を生成する既存研究を紹介する．その後既存モデルを用いた遅延削減手法について述べ，主観評価実験の結果と今後の展望を示す．

*連絡先：名古屋工業大学
愛知県名古屋市昭和区御器所町
E-mail: s.suzuki.663@stn.nitech.ac.jp

2　アバターを介したコミュニケーションの技術

近年，アバターを活用したコミュニケーションが注目を集めており，遠隔就労支援や教育，接客などの分野で社会的応用が進んでいる．アバターにはロボット型と CG 型があり，ロボット型は遠隔操作による対話や業務支援に利用され，CG 型は VR 空間や VTuber などで活用されている．本研究では CG アバターを対象とする．

アバターを介したコミュニケーションには，音声のみのやり取りと，身体動作を伴う方法がある．特に，身体動作を取り入れることで，非言語情報が伝達され，円滑なコミュニケーションが可能になる．非言語情報には，表情やジェスチャー，視線，姿勢などが含まれ，これらをアバターに反映させることで，対話の理解を助け，共同作業の効率向上にも寄与する [1]．例えば，音声とジェスチャーを組み合わせることで，情報伝達がしやすくなることが示されている [3]．また，教育分野ではジェスチャー付きの指導が学習効果を高めることが報告されている [4].

身体動作を伴ったアバターの操作には主にトラッキング技術が用いられる．トラッキングには様々な手法があるが，例えば専用機器を装着する方法と，カメラ

のみを用いる方法があげられる．専用トラッカーは高精度だが，手軽に利用するうえで装着の手間やプレイスペースの必要性が課題である．一方，カメラトラッキングは装着不要で手軽だが，環境による精度の低下が問題となる．例えば，明るさによってトラッキング精度が変動し，誤検出が発生する可能性がある [2]．これでは気軽にトラッキングを行えない．

また，アバターを使うことで生じる心理的・身体的負担も問題視されている．対面では無意識に行われるジェスチャーも，アバターを通す場合は意識的に操作する必要があり，負担が増える．Zoom などの Web 会議でも同様の「Web 会議疲れ」が指摘されており，アバターを用いたコミュニケーションでも類似の課題が生じる可能性がある．

このように，身体動作を伴ったアバターの活用には多くの利点がある一方で，操作の負担や環境の制約といった課題が存在し，気軽に利用できない．本研究では，トラッキングを必要とせず，音声のみでアバターのモーションを生成する「speech-to-motion」の手法を用いることで，より手軽なアバターを介したコミュニケーションを実現することを目指す．

3　音声からのモーション生成

音声からモーションを生成する「speech-to-motion」の研究が盛んになっており，近年では頭部動作に加えて上半身や全身のモーションを一括に生成する研究が増加している．

3.1　上半身モーション生成

上半身モーションの生成に関する研究では，ジェスチャーの自然さや多様性を向上させるため，スタイル制御や相互作用の考慮が行われてきた．例えば，ジェスチャーの高さや速度を調整できるモデルや，聞き手のモーションも同時に生成し相互作用を考慮する手法が提案されている [5] [6]．また，LLM（大規模言語モデル）を用いて発話内容に適したモーションを割り当てる研究も進められている [7]．しかし，これらの研究はリアルタイム性よりも自然さや文脈の適合性を重視しており，リアルタイム性を追求した研究は少ない．

3.2　DiffSHEG

リアルタイム性も重視した speech-to-motion モデルとして，DiffSHEG（A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation）がある [8]．DiffSHEG は，表情とジェスチャーを一緒に生成することを目的としたモデルで，文脈情報を考慮しながらを推論速度を向上させるための手法を導入している．例えば，Repaint アルゴリズムを DDIM サンプリングに適用することで，従来の拡散モデルより高速な推論を可能にしている [9]．

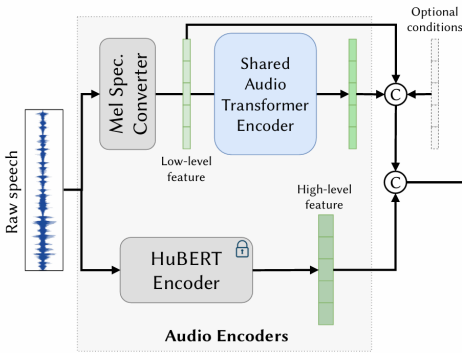

図 1: DiffSHEG: Audio Encoders

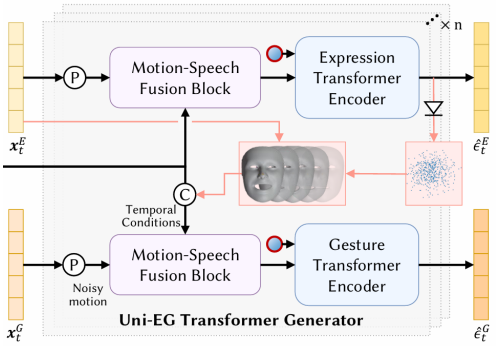

図 2: DiffSHEG: Uni-EG Transfomer Generator

DiffSHEG のモデル構造は Audio Encoder（図 1）と Uni-EG Transformer Generator（図 2）の 2 種に分類される．Audio Encoder ではメルスペクトログラムと HuBERT 特徴量の 2 つの特徴量を音声から抽出する機構である．メルスペクトログラムは音響情報をもった低次元特徴量であり，HuBERT 特徴量は文脈情報をもった高次元特徴量である．

Uni-EG Transformer Genetator では特徴量や話者 ID，サンプリングステップをもとに Diffusion モデルベースの推論を行う．ここでは，生成された表情の特徴量をジェスチャー生成に用いることで生成精度の向上を図っている．

4　Speech-to-motion におけるリアルタイム性の検討

アバターを用いた会話では，ジェスチャーや表情などの非言語的要素が重要だが，リアルタイムで自然なモーションを生成することは技術的な課題が多くあまり

取り組まれていない．また，既存の speech-to-motion 研究では，音声全体を入力としてモーションを生成する手法が主流であり，リアルタイム性を考慮した研究は限られている．そこで本研究では，DiffSHEG をベースに，会話に向けたリアルタイムなモーション生成手法を検討する．

リアルタイム性として，生成速度や低遅延なことが挙げられる．生成速度は DiffSHEG で取り組まれており，RTX3090 環境下では音声と同等またはそれ以下の時間で生成が可能である．一方遅延について，DiffSHEG では音声特徴量の抽出に HuBERT を使用しており，20 秒単位で音声を処理する仕組みとなっている．そのため，バッファを貯めるために 20 秒の遅延が発生する．この遅延は，会話における許容範囲（一般的に 200 ミリ秒未満）を大幅に超えており，アバター会話にはそのまま応用できない [10].

本研究では，HuBERT Encoder の処理遅延を削減するため，音声区間の分割を細かくし，より短い時間で特徴量を抽出する方法を提案する．具体的には，20 秒単位ではなく，5 秒や 2 秒などの短い区間で HuBERT を適用することで，遅延を削減する．また，HuBERT を使用せず，メルスペクトログラムのみで特徴量を取得する方法も検討する．これにより，入力音声の収集待ち時間を削減し，リアルタイム性を向上させる．

ただし，音声区間を短縮すると，文脈情報の欠落によりモーションの自然さが低下する可能性がある．特に，長期的な音声依存関係が重要な場合，生成モーションの一貫性が損なわれるリスクがある．そのため，低遅延化とモーションの品質のバランスを取ることが重要であり，本研究ではこれらのトレードオフについて検証を行う．

5　主観評価実験

本研究では，提案した低遅延 speech-to-motion モデルの評価を行うため，主観評価実験を実施した．低遅延化により文脈情報が減少し，ジェスチャーの自然さや発話との整合性に影響を与える可能性があるため，実験を通じてその影響を検証する．

5.1　学習条件

提案手法の評価のため，HuBERT の入力音声区切り長を 20 秒，5 秒，2 秒 に設定し，それぞれの条件で学習を行った．また，音声の前処理に HuBERT を用いないものも学習を行った．データセットには BEAT の英語話者データ（計 8 時間）を使用し，そのうち 1 時間を検証用，1 時間をテスト用，残りの 6 時間を学習用 に分割した．学習は エポック数 1000，バッチサイズ 1200 で実施し，モーションデータは 15fps にリサンプリングし，軸角度表現 で学習を行った．

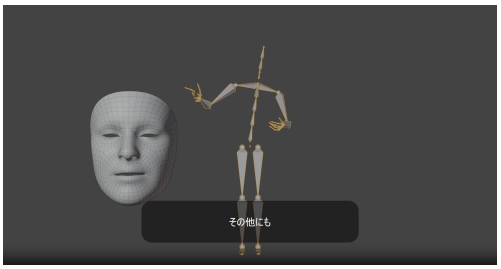

図 3: 実験に使用した動画

表 1: 評価項目の一覧

番号	評価項目
A-1	ジェスチャーの違和感
A-2	表情の違和感
B-1	ジェスチャーの動きの大きさ
B-2	ジェスチャーの動きの滑らかさ
B-3	ジェスチャーの動きの多様性
C-1	発話タイミングとジェスチャーの動きの同期性
C-2	ジェスチャーの発話内容の表現度
C-3	口の動きと音声の同期性
C-4	口以外の表情の発話内容の表現性

5.2　実験条件

評価には 20 代の男女 21 名を対象とし，Google Forms を用いたアンケート形式で実施した．参加者は タブレットやモニターを使用し，音声が遅延なく明瞭に聞こえる環境で動画を視聴した．評価対象の動画は以下の 6 種類で，音声は同一のものを使用した．

- Ground Truth (gt) - 実際の人間のジェスチャー
- 20 秒モデル (20s) - HuBERT 20 秒区切り
- 10 秒モデル (10s) - HuBERT 10 秒区切り
- 5 秒モデル (5s) - HuBERT 5 秒区切り
- 2 秒モデル (2s) - HuBERT 2 秒区切り
- HuBERT なし (off) - 文脈情報を使わないモデル
- ミスマッチ (miss) - 音声と異なるジェスチャー (話者は同一)

評価は 7 段階のリッカート尺度および SD 法 で実施し，以下の項目について評価した．(表 1 および 2)

表 2: 評価基準の一覧

番号	評価基準
A-1	1（全く違和感がない）←→ 7（とても違和感がある）
A-2	1（全く違和感がない）←→ 7（とても違和感がある）
B-1	1（小さい）←→ 7（大きい）
B-2	1（ガタガタだ）←→ 7（滑らかだ）
B-3	1（単調だ）←→ 7（多様だ）
C-1	1（適切でない）←→ 7（適切だ）
C-2	1（全く表していない）←→ 7（とても表している）
C-3	1（全く対応していない）←→ 7（とても対応している）
C-4	1（全く表していない）←→ 7（とても表している）

参加者は事前に 英語圏のジェスチャーに慣れるための動画を視聴し，実験の概要を静止画で説明された．また，評価の影響を抑えるため，2 つのグループ（A/B）に分けて動画の提示順序を変更した．図 3 は実験動画の様子である．

5.3 実験結果及び考察

HuBERT なしの言語文脈情報を用いない off について，言語情報を用いる他の手法より大きく低い評価となり，ミスマッチなモーションを再生する miss と比べてもほぼ同等となった．言語情報を用いない生成は，会話内容に対して一貫した意味を持つモーションを生成することは難しく，リアルタイムのモーション生成においても何らかの言語情報は利用すべきであることが分かった．

一方，HuBERT で用いる言語文脈情報の長さを 20 秒から 5 秒，2 秒と下げた場合，ほとんどの項目で 2 秒のモデルが最も高い評価を得た．原理的には，使用する言語情報の長さを短くするほど，会話の文脈に沿ったモーションが生成できなくなるため，モーションの評価も下がると予測されていたが，結果はこの予測と反するものとなった．この要因について，以下の検証を行った．

まず当初の予測に関して，言語文脈情報の長さとモーションの質の関連を改めて検証を行った．実験用の各動画について目視で検証を行った結果，言語文脈情報を短くしたモデルでは，たしかにモーションが文脈と関係ないものになっていく様子が観測された．例えば「a full of」と発話するとき，20 秒のモデルでは円を描くジェスチャーをしていたが 2 秒や off のモデルでは右手を左右に動かすだけであった．また，20 秒のモデルでは発話中のみジェスチャーが動くが，2 秒や off では，発話していないにもかかわらずジェスチャーが動くときがあった．このことから，当初の予測についてはモデルは正しく機能していることが分かった．

次に，同じく言語情報の長さと，今度はモーションのダイナミズムとの関連について検証した．目視での検証の結果，言語情報を長くしたモデルでは，結果として抑揚の大きいモーションが生成されていた．例えば，20 秒や 5 秒のモデルでは手を瞬間的に広げるような加速度の大きいモーションが生成されたが，2 秒のモデルにはそのようなモーションは生成されなかった．このことから，本研究のモデルは，言語情報が長いモデルでは変化量の大きいモーションが生成され，言語情報が短いモデルでは，生成モーションの動きの幅が比較的小さくなる傾向があることが分かった． また，gt のモーションと比較したところ 2 秒のモデルのダイナミズムが最も近いことが分かった．そのため，2 秒のモデルにおける評価を押し上げたものと推察される．

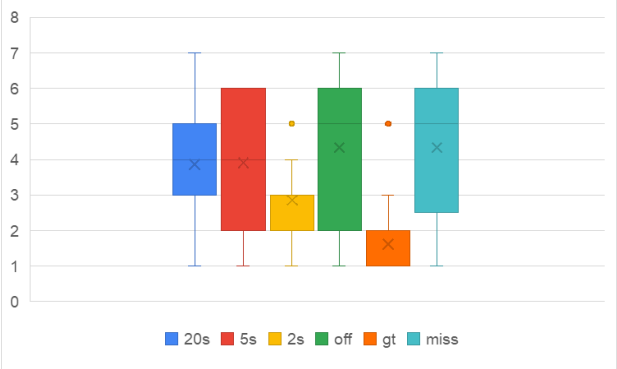

図 4: A-1：ジェスチャーの違和感↓

6 まとめ

本研究では，気軽なアバター操作を実現するために，音声のみで CG アバターを操作可能とする speech-to-motion を活用した．文脈情報を考慮した上半身生成モデル DiffSHEG をベースラインとしてモーションに含まれる言語情報量が漠然とした違和感にどのような影響を及ぼすか調査した．具体的には DiffSHEG で使われていた HuBERT Encoder 内の処理を改変し，HuBERT への入力フレーム長を従来手法より短く区切ることで，低遅延なモーション生成を可能にした．

主観評価では，低遅延化したモデルによって生成されたモーションの違和感について評価し，その結果，HuBERT の入力の単位時間が 2s のときに違和感の少ないモーションが生成できた．これを説明するために，「文脈の考慮時間増加に伴い，前後の発話を考慮したモーション生成が行われる」ことと「2 秒と Ground Truth のダイナミズムが近しかった」ことの 2 つの可能性を提示した．しかし，これらはあくまで可能性であり，具体的にこれらが正しいか更なる検証が必要である．また言語性に依存しないモーションの差については今回の実験では不明な点が多く，より詳細な分析が必要である．

参考文献

[1] Harrison Jesse Smith and Michael Neff. Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–12, New York, NY, USA, 2018. Association for Computing Machinery.

[2] Konstantinos Konstantoudakis, Kyriaki Christaki, Dimitrios Tsiakmakis, Dimitrios Sainidis, Giorgos Albanis, Anastasios Dimou, and Petros Daras. Drone control in ar: An intuitive system for single-handed gesture control, drone tracking, and contextualized camera feed visualization in augmented reality. *Drones*, Vol. 6, p. 43, 02 2022.

[3] Ryan Khushan Ghamandi, Ravi Kiran Kattoju, Yahya Hmaiti, Mykola Maslych, Eugene Matthew Taranta, Ryan P. McMahan, and Joseph LaViola. Unlocking understanding: An investigation of multimodal communication in virtual reality collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[4] Amy L. Baylor and Soyoung Kim. The effects of agent nonverbal communication on procedural and attitudinal learning outcomes. In Helmut Prendinger, James Lester, and Mitsuru Ishizuka, editors, *Intelligent Virtual Agents*, pp. 208–214, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[5] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum*, Vol. 39, No. 2, pp. 487–496, 2020.

[6] Mingze Sun, Chao Xu, Xinyu Jiang, Yang Liu, Baigui Sun, and Ruqi Huang. Beyond talking – generating holistic 3d human dyadic motion for communication, 2024.

[7] Qingrong Cheng, Xu Li, Xinghui Fu, Fei Xia, and Zhongqian Sun. Siggesture: Generalized co-speech gesture synthesis via semantic injection with large-scale pre-training diffusion models, 2024.

[8] Junming Chen, Yunfei Liu, Jianan Wang, Ailing Zeng, Yu Li, and Qifeng Chen. Diffsheg: A diffusion-based approach for real-time speech-driven holistic 3d expression and gesture generation, 2024.

[9] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022.

[10] International Telecommunication Union (ITU). One-way transmission time. Recommendation G.114, ITU-T, May 2003. Series G: Transmission Systems and Media, Digital Systems and Networks.

7 謝辞

本研究は，JST ムーンショット型研究開発事業，JPMJMS2011 の支援を受けたものです.