CG アバター遠隔対話のための音声からのモーション生成および CG 特有性の分析

Generation of Conversational Motions from Speech for CG Avatar Communication and Analysis of Their Uniqueness

藤岡　侑貴[1*]　上乃　聖[1]　李　晃伸[1]
Yuki Fujioka[1]　Sei Ueno[1]　Akinobu Lee[1]

[1] 名古屋工業大学大学院工学専攻
[1] Department of Engineering Nagoya Institute of Technology

Abstract: CG アバターを用いた会話において，トラッキングや身体キャプチャーを使ったマルチモーダル会話は操作の負担が大きい．本研究では，音声波形から CG アバターの表情および頭部動作を予測する CG アバターのための Speech2motion システムを構築する．また，CG アバターであることを意識したアバター越しの会話と通常の会話では発話者の身体動作の特徴が異なるという仮説のもと，アバター操演データの収録・分析を行った結果を報告する．

1　はじめに

近年，アバターを介した遠隔コミュニケーションに注目が集まっている．アバターの典型的な操作システムでは，人が話す際の表情や頭，手の動きなどをキャプチャし，別の場所にあるアバターへ音声とともにリアルタイムに転送・反映させることで，まるで発話者自身が乗り移ってそこにいるかのように会話することができる．特に近年，遠隔での教育や接客などの社会的応用も期待されている．本研究では，アバターの中でも CG キャラクターを用いた CG アバターによる対話を扱う．

アバターの利用事例は増加している一方で，普及するには至っていない．その原因として，アバターコミュニケーションにおける課題がいくつか考えられる．類似した問題として Web 会議疲れが挙げられる．Web 会議では，自身をジェスチャを含めてカメラにうまく映るよう気を配る，同意を示すためにより大袈裟に頷いてみせるなど，画面越しの相手に非言語情報を伝えるための意図的な努力が必要であり，これが Web 会議特有の疲れの一因となっている [1]．同じ画面上でのコミュニケーションである CG アバター対話においても，アバターを操作するために身振り手振りを交えながら発話し続けることは，操作者の負担が大きい．また，CG キャラクターを演じて動画コンテンツの作成や配信を行う，いわゆる VTuber（バーチャルユーチューバー）の操作においても，アバターのペルソナ（性格設定）に合わせたふるまいを配信において維持し続けることは困難であることが報告されている [2]．さらには，表情や身体動作をキャプチャするための機材を用意する必要があるという問題もある．今後アバターを介したコミュニケーションが普及していくためには，このようなアバター操作者の負担は障壁となる．

本研究では，話者の音声情報から CG アバター遠隔対話における表情および頭部動作を予測する Speech2motion システムを構築する．音声で会話をするだけでそれに合ったふるまいが自動生成されることで，話者の負担を軽減し，容易なアバターコミュニケーションを可能にする．

また，アバターコミュニケーションにおける CG 特有性についても分析する．従来の Speech2motion に関する研究においては，人の自然な動きをそのまま再現するシステムが主に検討されてきた．しかし，アニメ調のキャラクタや VTuber のような見た目をした CG アバターを想定した際，その見た目からユーザの期待する動きが，人の自然なふるまいとは異なる可能性がある．前述した，画面上での大袈裟な意思表示や，VTuber のペルソナに合わせたふるまいなどもこれに当たり，通常の会話よりも誇張を含んだ動きとなる．本研究では，VTuber らのふるまいから着想を得て，CG アバターであることを意識したアバター越しの会話と通常の会話では発話者の身体動作の特徴が異なるという仮説を立てた．これらの特徴を，本研究では「CG 特有のふるまい」「CG 特有性」と呼ぶ．アバター操演データを収集し，CG 特有性の有無およびその特徴を明らかにする．

*連絡先：名古屋工業大学大学院 工学研究科工学専攻
情報工学系プログラム
〒 466-8555 愛知県名古屋市昭和区御器所町
E-mail: y.fujioka.996@stn.nitech.ac.jp

したがって本研究では，以下の 2 つについて扱う．

1. CG アバター対話における表情・頭部動作を予測する Speech2motion システムの構築および評価
2. CG 特有のふるまいについての分析

以下，第 2 章で音声からモーションを生成する先行研究を挙げた上で，第 3 章で本研究で構築したデータセットについて，第 4 章でシステムとその実験的評価について述べる．第 5 章でデータに基づく CG 特有性についての分析を述べ，最後に第 6 章でむすびと今後の展望について述べる．

2 音声からの facial animation 生成

音声信号を入力として，音声に合わせた人の口唇の動きや表情の変化を生成する研究はさまざまに行われている．2D 画像ベースの手法では，顔画像と音声を入力とし，その音声に合わせて顔画像の変化を生成・合成する．あらかじめ作成された動画像を発話内容に合わせて目・口領域に重ね合わせる手法 [3] や，与えられた画像から顔のランドマークを，音声からランドマークの変化を推定し，ランドマーク情報をもとに顔画像を出力する手法 [4] などがある．3D 頂点ベースの手法では，人型 3D モデルの頂点変化を音声から直接推定する．Karras の提案した CNN (Convolutional Neural Network) を用いた End-to-End モデル [5] や，Cudeiro らによる Encoder-Decoder 構造のモデル VOCA [6] などが挙げられる．

3D モデルに対する別のアプローチとして，ブレンドシェイプベースの手法がある．ブレンドシェイプは，3D モデルの目や口の開閉，眉の上下などの動きをあらかじめ定義したシェイプを，重みづけでブレンドすることによってモデルの表情を動かす手法であり，その重みの時系列変化を推定する．その手法の一つとして，Stan らが提案した FaceDiffuser [7] が挙げられる．FaceDiffuser は，拡散モデルに基づく手法であり，同一の発話音声に対して一意に定まらない人の表情の多様性を再現する．本研究では，この FaceDiffuser を CG アバター対話のために拡張する．

3 CG アバター操演データセットの構築

本研究における Speech2motion システムおよび CG 特有のふるまい分析のため，CG アバター操演データセットの構築を行う．データセットは，CG アバターを介した遠隔会話時の様子を収録する．

アバター操演システムとして，音声対話エージェントのシステムである MMDAgent-EX [8] をアバターコミュニケーション用に拡張したシステムを用いる．このシステムは，遠隔地から送られてくるフェイシャルキャプチャ情報や身体動作情報をリアルタイムに反映させたキャラクタを表示し，同時に音声を再生することでアバターコミュニケーションを行うことができる．キャプチャ情報や音声情報を保存する機能を備えており，以降のデータ収集およびシステムはすべて同一の MMDAgent-EX 環境で収集・実行している．

収録するデータは，音声とモーションパラメータである．音声のサンプリング周波数は 16 kHz，ビット数は 16 ビットであり，無音区間を除いた発話のみを記録する．モーションパラメータの取得には iFacialMocap [9] を用い，Apple ARKit [10] のフェイシャルキャプチャ機能によって取得する．収録データは，シェイプ情報 51 種類および頭部の x, y, z 座標 [mm]，頭・左目・右目それぞれの x, y, z 軸回転量 [rad] を合わせた全 63 個のパラメータであり，60 fps で記録される．

本研究では，CG アバターであることを意識したアバター越しの会話と通常の会話では発話者の身体動作の特徴が異なるという仮説を検証するため，それぞれの会話データを収集した．以降，CG アバターであることを意識してアバター越しに会話したデータを「Avatar データ」，CG アバターを使用せず自然に会話したデータを「Natural データ」と呼ぶ．

データ収録は，被験者 1 名と対話相手 1 名による雑談対話を通じて行った．被験者は日本語母語話者であり，対話相手と互いにビデオ通話を介して会話をしながら，前述した収録システムを用いてデータを記録した．Avatar データ収録時は，被験者は自身が操作するアバターを画面上に表示して会話を行い，Natural データ収録時は，アバターを表示する代わりに被験者自身のカメラ映像を映す．被験者 17 名分の Avatar データと Natural データを収録し，収録時間はそれぞれ約 6.5 時間である．

Avatar データを収録する際は，被験者の思う CG らしい動きをするように指示をした．このときの「CG らしさ」については被験者に委ねており，明確なインストラクションは行っていない．CG アバターは，ムーンショット研究開発「アバター共生社会」の CG アバター「ジェネ」 [8] を採用し，被験者全員が同一のもので操演した．

4 CG アバター遠隔対話のための Speech2motion システム

本研究では CG アバターを用いた遠隔対話において，操作者の発話音声から CG アバターの表情および頭部動作を予測する Speech2motion システムを構築する．バストアップで表示された CG アバターとの会話を対

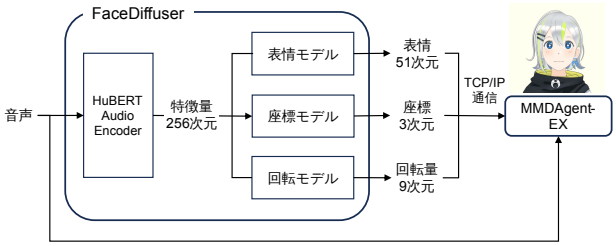

図 1: Speech2motion システム構成（分離モデル）

象とし，腕や全身を用いたジェスチャーなどの身体動作は含まない．顔全体の表情，音声対話に伴う頭部および上半身の自然な動作を発話音声のみから生成することを目指す．

4.1 Speech2motion システム

音声を入力し，CG アバターにモーションを出力するシステムについて説明する．システム構成を図 1 に示す．

学習モデルには，FaceDiffuser [7] を使用し，3 章で収録した Avatar データでモデルを学習する．FaceDiffuser では，入力音声波形から HuBERT [11] を用いて抽出した特徴量 256 次元に基づき，拡散ステップ 1000 で表情 51 次元のブレンドシェイプ重み系列を出力する．本研究では，この出力に頭部動作 12 次元を追加して予測するように拡張する．拡張方法として，表情 51 次元，頭部座標 3 次元，頭部回転量 9 次元をそれぞれ個別にモデル化する「分離モデル」と，全パラメータ 63 次元を結合してモデル化する「結合モデル」の 2 つを学習した．なお，図 1 では分離モデルの構成を示している．各モデルのハイパーパラメータは，結合モデルおよび分離モデルの表情，頭部回転量に関しては FaceDiffuser のデフォルト値を使用している．分離モデルの頭部座標については予備実験の結果より，デフォルト値からエポック数を 100，GRU 層を 64 次元に変更した．モデルから出力された表情，頭部座標，頭部回転量を，音声と同時に MMDAgent-EX に伝送することで，CG アバター上で音声に合わせたモーションが再生される．

頭部の座標は，MMDAgent-EX が最初に受け取った値との相対値分だけ CG アバターが動く仕様となっている．したがって訓練時には，記録した全フレームの頭部座標から各データごとの座標の平均値を引くことで正規化した値を使用する．また，60 fps で記録したモーションデータは，FaceDiffuser のデフォルト値に合わせ，30 fps にダウンサンプリングして使用した．

データは訓練データ，検証データ，テストデータに 0.8, 0.1, 0.1 の割合で分割して使用した．また，入力するデータの長さが 1 データあたり 5〜10 秒 (150〜300 frame) となるように，5 秒未満のデータは結合し，10 秒以上のデータは分割して使用する．

表 1: 各モデルの MSE ↓

	BEAT	Avatar		
	表情	表情	頭部座標	頭部回転量
結合モデル	-	1.039	13.498	0.236
分離モデル	0.651	0.630	13.491	0.226

表 2: 各モデルの MDD ↓

	BEAT	Avatar		
	表情	表情	頭部座標	頭部回転量
結合モデル	-	9.620	22217	1.559
分離モデル	5.157	3.404	22146	0.856

4.2 客観評価

モデルの再現度を評価するため，客観評価を行った．先行研究 [7] を参考に，Speech2motion システムを用いて生成したモーションと Ground Truth との MSE (Mean Squared Error), MDD (Motion Dynamics Deviation) を算出した．MDD は，標準偏差の差を計算した値であり，モーションのばらつきの再現度を評価する．

データによる比較として，先行研究でも用いられている BEAT [12] の表情データセットで実験した．BEAT の表情データセットは，本研究の表情データと同一の形式であり，そのうちの英語母語話者 10 名の合計約 16 時間分のデータを使用した．MSE および MDD の結果を，それぞれ表 1, 2 に示す．さらに，頭部座標，回転量については生成結果と Ground Truth の時間変化をパラメータごとにプロットした結果を図 2 に示す．

いずれのパラメータについても，MSE, MDD ともに分離モデルの方が誤差が小さい結果となった．BEAT データでの結果と比較しても，分離モデルでは BEAT と同程度あるいはより高い精度が得られた．頭部動作に関しては，分離モデルと結合モデルに MSE 差は現れなかったものの，図 2 の結果から，頭部回転量に関しては特に分離モデルの方が Ground Truth の動きを再現できていることがわかる．このことから，以降の実験では分離モデル，すなわち表情，頭部座標，頭部回転量をそれぞれ別にモデル化した結果を用いる．

4.3 主観評価

生成されたモーションの再現度が実際にどう知覚されるかを測るため，主観評価を実施した．被験者は大学生・大学院生 28 名であり，音声に合わせて CG アバ

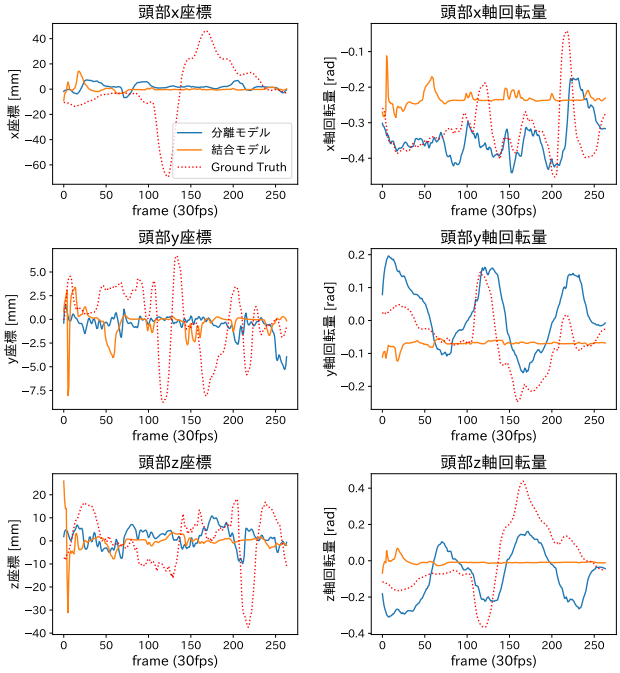

図 2: 生成結果の例（頭部動作）

表 3: 生成結果の再現度評価（5 段階評価）

(a) リップシンク	3.85	(e) 俊敏さ	3.55
(b) 表情	4.05	(f) 幅	3.29
(c) 自然性	3.85	(g) 多様性	3.71
(d) タイミング	3.78		

ターが動く様子を記録した動画を視聴し，主観評価アンケートに回答する．

被験者は，同じ音声に対して異なる動きをするアバター A と B について，その動きを比較して評価をする．動画は，アバター A, B が画面上に並んでおり，アバター A は Ground Truth の動きを，アバター B は Speech2motion システムで生成した動きを再生する．以下の (a)〜(g) について，アバター A の動きに対してアバター B の動きがどの程度似ているかを評価した．(a), (b) は表情，(c)〜(g) は頭部動作についての項目である．全ての項目について「全く違う（1 点）」〜「似ている（5 点）」の 5 段階で評価した結果の平均値を表 3 に示す．

(a) 音声と口の動きの同期性（リップシンク）
(b) 表情
(c) 自然性（動きのぎこちなさ・滑らかさ）
(d) タイミング（音声に対する動きの同期性）
(e) 機敏さ（前後左右の動きの速さ・緩やかさ）
(f) 幅（前後左右の動きの大きさ）
(g) 多様性（動きの単調さ・多様さ）

全ての項目において「やや似ている」という評価を最も多く獲得する結果となり，平均値としても 3 点以上を獲得した．表情は特に得点が高い結果が得られた．評価動画にて顕著な表情変化としては笑顔があり，笑い声に合わせた笑顔の生成ができていることも確認できた．ハイコンテクストな感情表現については確認できないが，簡単な感情表現が可能であることがわかった．一方で，(f) 幅は他の項目に比べて点数が低い．これは，図 2 の頭部座標のプロットからもわかるとおり，頭部動作における座標の大きな変化を再現できていないことが起因していると考えられ，妥当な結果だと言える．

5 CG 特有のふるまいについての分析

本研究の目標の 2 つ目である，CG 特有のふるまいについての分析を行う．3 章にて収集した Avatar データおよび Natural データを比較することによって，CG 特有性の有無およびその特徴について検証する．特にここでは，CG アバターを用いることによって，通常の対話よりも大袈裟な，誇張を含むふるまいが現れるという仮説を立て，それについて検証する．

5.1 客観評価

仮説に基づき，収録データの頭部動作の変化量の違いについて分析を行った．収録データを 1 秒間隔で区切り，その間の最大値と最小値の差を計算することによって，1 秒間での最大変化量を各モーションパラメータについて算出した．最大変化量の平均値を話者ごとにプロットした結果を図 3 に示す．また，全話者の平均値を表 4 に，標準偏差を表 5 に示す．なお，座標軸とアバターの動きの対応は図 4 のとおりである．

個人差はあるものの，全体として Avatar データの方が変化量が大きいことがわかる．図 3 において特に Avatar と Natural の差が大きい話者 6, 11, 12, 17 らに注目すると，どの話者も全てのパラメータにおいて Avatar データの方が平均値が大きい．Natural データの方が平均値が大きい話者もいるが，全てのパラメータにおいて Natural の方が大きい結果となった話者は

表 4: 頭部動作 1 秒間における最大変化量の平均値

	頭部座標 [mm/sec]			頭部回転量 [rad/sec]		
	x 座標	y 座標	z 座標	x 軸	y 軸	z 軸
Avatar	10.15	5.24	15.80	0.074	0.069	0.063
Natural	6.60	4.46	12.72	0.057	0.054	0.048
差	3.55	0.78	3.08	0.017	0.015	0.015

表 5: 頭部動作 1 秒間における最大変化量の標準偏差

	頭部座標 [mm/sec]			頭部回転量 [rad/sec]		
	x 座標	y 座標	z 座標	x 軸	y 軸	z 軸
Avatar	7.75	3.34	9.96	0.046	0.057	0.049
Natural	4.66	3.01	7.89	0.034	0.036	0.034
差	3.09	0.33	2.07	0.012	0.021	0.016

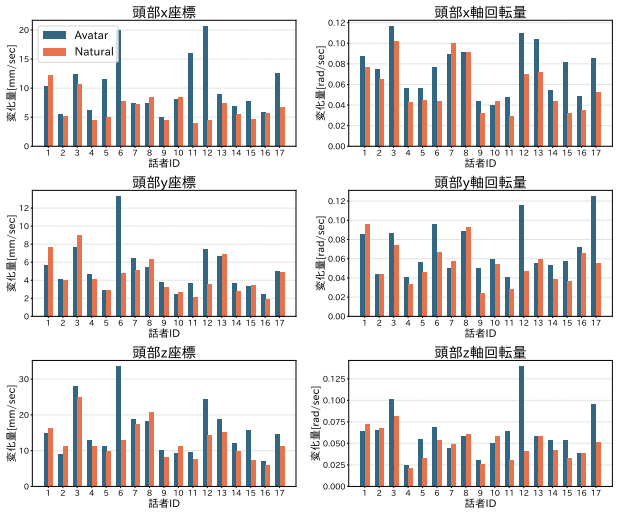

図 3: 話者ごとの頭部動作 1 秒間における最大変化量の平均値

おらず，またその差についても話者 6, 11, 12, 17 のものと比較して小さい．

表 4 より，パラメータごとに注目をすると，y 座標の変化量については x, z 座標の結果よりも Avatar と Natural の差が小さい．しかし，y 軸方向の移動，すなわち画面縦方向の移動は，画面上での対話においては基本的に固定されるものであることを考慮すると，y 座標の値に差が現れにくい傾向にあることは妥当な結果であると考えられる．

また，表 5 より，動きの最大変化量の標準偏差に関しても，Avatar の方が大きい傾向が見られた．これらの結果から，アバターを意識して用いることによる動きの変化はあると考えられ，そこでは通常の対話よりも動きが大きく，多様なふるまいが現れる傾向があることがわかる．

5.2 主観評価

客観評価にて明らかになった Avatar データと Natural データの差が，実際にどう知覚されるかについて検証するため，主観評価実験を行った．4.3 節の主観評価実験と同様にして，収録した音声とモーションを CG アバターに適用した動画を視聴し，主観評価アンケートに回答する．被験者は，大学生・大学院生 27 名である．

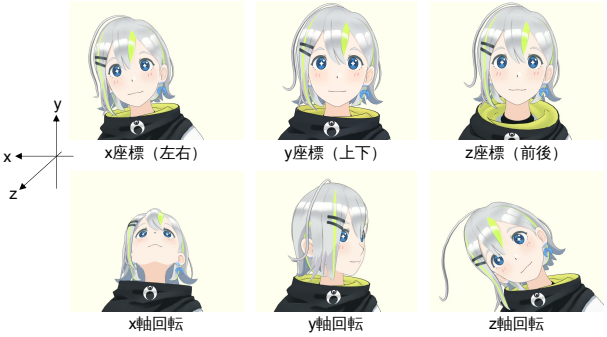

図 4: 座標軸とアバターの動きの対応

同一話者の Avatar データと Natural データについて，それぞれアバター A または B として再生する．Avatar と Natural の発話内容は異なり，被験者はアバター A，アバター B の順でその動きを観察し，その動きについて比較評価をする．アバター A, B と Avatar, Natural データの対応は，動画によってランダムに割り当てており，被験者はアバター A, B が Avatar と Natural どちらのデータであるかを知らない状態で評価をする．

実験に使用したデータは，5.1 節の結果をもとに選んだ．客観指標での結果と主観的な知覚の関係を調べるため，客観的指標において特に差が大きく現れた話者 6, 12, 17，客観指標の一部において差が現れた話者 11, 15，差があまり現れなかった話者 1 の計 6 名のデータについて評価をした．

「アバター A・B の動きはどの程度似ているか（質問 1）」,「アバター A・B の頭・首の動きを比較して、どちらのアバターの方が動きの幅が大きい/多様だと感じたか（質問 2）」という質問をし，質問 1 については以下の (a)〜(e)，質問 2 は (d), (e) の要素について評価した．それぞれ 5 段階で評価をし，話者ごとの平均値を算出した結果を表 6 に示す．

(a) 表情
(b) 頭・首の動きのタイミング
(c) 頭・首の動きの機敏さ
(d) 頭・首の動きの幅
(e) 頭・首の動きの多様性

表情，タイミングについては，どの話者も Avatar と Natural に違いがない結果となった．アバターを意識することにより，誇張された表現として豊かな表情が現れることも考えられたが，今回の結果ではその影響は現れなかった．俊敏さでは話者 17 が 3 点以上を獲得し，僅かに差が出たものの，全体としては同様の傾向は見られない．

幅，多様性の項目では，差が知覚される結果が得られた．話者 6, 12, 17 が 3 点以上を獲得しており，Avatar と Natural には差があるという評価が多くされた．同

表 6: Avatar データと Natural データの主観比較結果

話者 ID	質問 1（動きの相違度） (a) 表情	(b) タイミング	(c) 俊敏さ	(d) 幅	(e) 多様性	質問 2（比較） (d) 幅	(e) 多様性
6	2.07	2.44	2.96	**3.67**	**3.15**	**4.41**	3.96
12	2.30	2.30	2.63	**4.56**	**3.33**	**4.67**	**4.00**
17	2.70	2.48	**3.22**	**4.37**	**3.89**	**4.67**	**4.26**
11	2.19	1.96	1.96	1.85	2.33	3.33	3.19
15	2.59	2.00	2.41	2.78	2.63	3.96	3.96
1	1.85	1.81	1.78	1.74	2.11	3.30	3.30
平均	2.28	2.17	2.49	**3.16**	2.91	**4.06**	3.78

質問 1 は「似ている（1 点）」〜「全く違う（5 点）」を 5 段階で評価し，値が高いほど動きに差があることを示す．質問 2 は「Natural の方が幅が大きい/多様（1 点）」〜「Avatar の方が幅が大きい/多様（5 点）」を 5 段評価し，値が高いほど Avatar の方が動きの幅が大きい/多様であることを示す．

話者は，質問 2 の結果でも「Avatar の方が幅が大きい/多様」と点数が高く評価されていることから，その差がアバターを用いた際のふるまいの方が動きの幅が大きく，多様であることによるものであることがわかる．話者 6, 12, 17 は，客観評価において特に差が大きく現れた話者であり，これらの結果は客観評価とも対応付く．これらの結果より，CG アバターをよく意識しながら会話を行うことによって，通常の対話よりも大袈裟なふるまいが現れる傾向があると言える．

6 むすび

本研究では，CG アバターを介したコミュニケーションを音声のみで行うことを目標として，発話音声から CG アバターの表情および頭部動作を予測する Speech2motion システムを構築した．実際に CG アバターを操演しながら遠隔での雑談対話を行った際の音声とフェイシャルキャプチャデータを収録し，Speech2motion の先行研究である FaceDiffuser モデルを CG アバターのために拡張してこれらのデータを学習した．分離モデル，結合モデルの 2 つについて客観評価した結果，分離モデルの方が良い精度が得られた．また，分離モデルの生成結果について主観評価を実施したところ，実際に人がアバター操演する動きに近い自然性や同期性のある動きが生成可能であると示された．

さらに，CG アバターを意識したアバター越しの会話と通常の会話では発話者の身体動作の特徴が異なるという仮説のもと，それぞれの会話データを収集し，そのデータについての分析を行った．客観評価では，動きの変化量を算出することによって 2 種のデータの違いを検証した．個人差はあるものの，アバターを意識することによって，動きが大きなふるまいをするようになるユーザがいることがわかった．主観評価では，その違いが人にどう知覚されるかを実験した．表情やタイミングに違いは無かったが，動きの幅や多様性には違いがあると評価され，アバターを介することによるふるまいの大きさへの影響がわかった．

本研究では，複数名分のアバター操演データを収録し，それをもとにモデル化したが，その個人性については考慮していない．今後は，アバターのペルソナや操演者の違いによってそのふるまいを変化させる Speech2motion システムも考えられる．その際，CG 特有のふるまいについての詳細をさらに明らかにすることによって，より CG らしい動きを生成する Speech2motion システムへと繋がることが期待される．

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ある．

参考文献

[1] Jeremy N Bailenson. Nonverbal overload: A theoretical argument for the causes of zoom fatigue. 2021.

[2] Man To Tang, Victor Long Zhu, and Voicu Popescu. Alterecho: Loose avatar-streamer coupling for expressive vtubing. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 128–137. IEEE, 2021.

[3] 北岡教英, 西村良太, 太田健吾. フォトリアル cg エージェントとのマルチモーダル対話. 日本音響学会誌, Vol. 78, No. 5, pp. 257–264, 2022.

[4] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, Vol. 39, No. 6, pp. 1–15, 2020.

[5] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (ToG)*, Vol. 36, No. 4, pp. 1–12, 2017.

[6] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10101–10111, 2019.

[7] Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. Facediffuser: Speech-driven 3d facial animation synthesis using diffusion. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–11, 2023.

[8] 李晃伸, 石黒浩. 自律・遠隔融合対話システムのための高生命感・高存在感 cg エージェントの開発. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 第 96 回 (2022.12), p. 27. 一般社団法人 人工知能学会, 2022.

[9] HOME - iFacialMocap. `https://www.ifacialmocap.com/`.

[10] ARFaceAnchor — Apple 開発者向けドキュメント. `https://developer.apple.com/documentation/arkit/arfaceanchor/`.

[11] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, Vol. 29, pp. 3451–3460, 2021.

[12] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. *arXiv preprint arXiv:2203.05297*, 2022.