3D アバターに適用可能な 音声駆動型リアルタイム表情生成システム

市川淳貴 [1]　徳久良子 [1 2]
[1] 愛知工業大学　[2] 理化学研究所
k22010kk@aitech.ac.jp

概要

近年，3D アバターを介したコミュニケーションが普及している．一方で，音声に同期したリアルタイムな表情生成には，遅延や表情の不自然さといった課題が残されている．本研究では，音声のみから表情動作の制御値を逐次推定し，限られた計算資源下でもリアルタイム動作する **Facial Motion Generator** を提案する．異なる音声特徴量と軽量な Cross-Attention モデルを用いた設計により，NVIDIA の Audio2Face-3D と比較して単一話者データで MAE（平均絶対誤差）を 0.147 → 0.111 に低減し，推論時間を 0.62 ms/frame へ高速化した．

1　はじめに

3D アバターを用いた配信や大規模言語モデルに基づく対話エージェントの普及により，近年，人と人（もしくは人と AI）が仮想的なキャラクターを通してコミュニケーションする機会が増えている．本研究で扱う 3D アバターとは，人の姿や顔を 3 次元 CG で表現したキャラクターであり，画面上でユーザやエージェントの「見た目」として振る舞うものである．人と 3D アバターとのコミュニケーションでは，発話内容だけでなく，うなずきや口の動き，表情といった非言語情報が，対話の自然さや相手の理解のしやすさに大きく関わることが知られている [1, 2]. そのため，アバターに対して人間らしい表情を付与する技術は，コミュニケーションの質を左右する重要な要素である．ここで，アバターの表情を適切に制御するためには，口の開閉や眉の上げ下げといった動きを数値として表した「表情動作の制御値」を，各時刻において逐次推定し，アバターに与える必要がある．しかし，これらの制御値を人手で逐次入力することは現実的ではなく，音声などの入力から自動的に推定する仕組みが求められる．

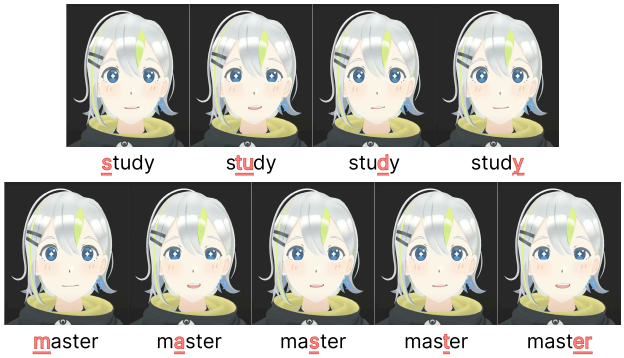

図 1　提案手法の出力例

この課題に対して，例えば NVIDIA は，音声から表情制御に用いる値を推定してアバターを駆動する手法（Audio2Face）を提案しており，高品質な表情生成を実現している [3]．Audio2Face は，入力音声から口元や眉などの動きを表す表情制御値を推定し，その時系列に基づいて 3D アバターの表情を自動的に生成するシステムである．Audio2Face を用いてアバターを構築する方法も考えられるが，Audio2Face の一部は API 経由で動作する構成であるため，通信に起因する遅延が生じる可能性がある．対話のように入力に追従した表情更新が求められる用途では，同期ずれが生じやすく，表情表現の自然さを損ねる恐れがある．また，システムがオープンではない実装に依存する場合，ローカル環境における再現性や改良，運用が困難といった課題も残る．

そこで本研究では，音声入力のみから表情動作の制御値を軽量に推定し，ローカル環境でリアルタイムに動作する音声駆動型表情生成手法 **Facial Motion Generator** を提案する．図 1 に，Facial Motion Generator によって 3D アバターの表情を生成した出力例を示す[1)]．図 1 は，「stusy」，「master」と発音した場合の顔表情を時系列で表示している．

本研究の貢献は以下の 2 点である

1)　本研究で使用した 3D アバターは右記から利用できる：https://github.com/mmdagent-ex/gene

- 軽量でリアルタイムに動作する音声駆動型の表情生成手法 Facial Motion Generator の提案
- Facial Motion Generator のソースコードの公開[2)]

2 関連研究

本節では，音声入力のみから顔表情を生成する既存研究と，3D アバターの表情を動かすために用いられる表情の表現形式について述べる．

音声入力による顔表情生成 音声入力のみを用いた顔表情生成は，発話音声から口元や顔表情の時間変化を推定し，3D アバターの顔の動きを自動生成する技術である．既存研究としては，音声から顔メッシュの頂点変位系列を直接予測する FaceFormer [4], VOCA [5] ，音声から表情動作の制御値を生成する Audio2Face-3D [6, 2] などの手法が提案されている．

NVIDIA の Audio2Face-3D は音声から顔表情アニメーションを生成し，後述する BlendShape という形式の表情動作の制御値を出力する．また Audio2Face-3D のツールキットには，学習フレームワークの例示用データとして，音声ファイルとそれに対応する表情係数や形状データ等を含むサンプルデータセット（Audio2Face-3D-Dataset-v1.0.0-claire）も含まれている．本研究では，Audio2Face-3D をベースラインとして比較評価を行い，学習には NVIDIA が公開しているデータセットを使用する

表情の表現手法と標準的な形式 3D アバターの顔表情は時間ごとに更新される制御パラメータの系列として与えられる．表情表現の代表的な制御手法には，(i) 顔メッシュの頂点座標を直接変位させる方法 [4, 7]，(ii) あらかじめ用意した複数の表情の動きを係数で制御する方法 [8, 9], (iii) ボーン（関節）の回転や平行移動で顔部位を制御する方法 [10, 11] などがある．このうち (ii) では BlendShape と呼ばれる係数を出力する手法が一般的である．BlendShape [12] は表情の動きを 52 個の要素に分解して表現する方法であり，各係数は口元や眉などの動きの大きさに対応し，中立状態を基準に 0 から 1 の範囲で表情の動きの大きさを表す．表情の動きを係数値として扱えるため学習モデルの出力として扱いやすく，ツール やゲームエンジン とも接続可能な設計となっている．本研究では，既存アバターへの適用容易性とベースライン手法との互換性の観点から，BlendShape の係数列を表情制御値として用い，音声から各時刻における BlendShape の値を推定する．

2) Facial Motion Generator は右記から入手できる：https://github.com/cl-ait/facial-motion-generator

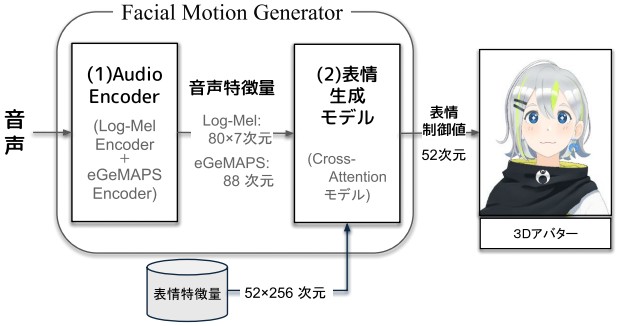

図 2 提案手法の推論時アーキテクチャ

3 提案手法

図 2 に提案手法 **Facial Motion Generator** の構成を示す．Facial Motion Generator は音声のみを入力とし，3D アバターの表情を駆動するための表情動作の制御値（52 次元）をリアルタイムに推定して出力することができる．本モデルは **(1)Audio Encoder** と **(2) 表情生成モデル**から構成され，入力音声から抽出した音声特徴量（Log-Mel および eGeMAPS）を用いて表情全体の動きを逐次出力する．出力された 52 次元の制御値を 3D アバターに適用することで，発話音声に合った表情をリアルタイムに生成できる．

Audio Encoder 図 2(1) に示すように， Audio Encoder は音声から表情推定に有用な音声特徴量を抽出する．本研究では音のスペクトル情報を表す Log-Mel 特徴量と，話し方の特徴である声の高さや強さ等を表す eGeMAPS 特徴量を併用し，発音に由来する口の形と抑揚に由来する表情変化の両方を表現する．また，リアルタイム性を確保するために音声特徴量は固定の窓長と固定のフレーム数に制限することで, ストリーミング入力に対して一定でかつ少ない計算量で逐次推定を行う. 具体的には Log-Mel 特徴量は窓長 25ms，シフト 10 ms で推論実行タイミングに対して最も近い 7 フレームを出力とし，eGeMAPS 特徴量は窓長 5 s，シフト 300 ms で推論実行タイミングに対して最も近い eGeMAPS ベクトルを選び，出力としている．

表情生成モデル 表情生成モデルは，図 2(2) に示すように，音声特徴量を入力として表情動作の制御値である BlendShape（52 次元）を逐次推定する．本研究では，52 個の BlendShape の各要素に対応した表情特徴量を用意する．この表情特徴量は入力データから抽出する特徴ではなく，学習を通して自動的に調整されるパラメータである．推論時は表情特徴

量を手がかりとして，音声特徴量のどの部分に注目するべきかを Cross-Attention で計算し，得られた音声特徴量から各 BlendShape の値を推定する．短期 Log-Mel と長期 eGeMAPS を統合し，BlendShape を Query とする構造により，口元の高速な動きと感情表情の緩やかな変化を単一ネットワークで扱えるようにした．この仕組みにより，音声特徴量の成分が口元，眉，目など，どの表情要素の動きに対応するかをモデルが学習する．

学習・評価用データと教師係数列の生成　学習と評価には，NVIDIA が公開する *Audio2Face-3D-Dataset-v1.0.0-claire* を用いる[3]．本データセットは単一話者 Claire の音声と，各時刻の 3D 顔形状からなる顔メッシュ列で構成される．一方で，BlendShape 係数はデータセットに含まれない．そこで，本データセットに付属している BlendShape 基底ファイルを用いて，各フレームの顔メッシュ系列を BlendShape52 次元空間へ変換し，30fps にリサンプリングした係数列を教師データとして生成した．最終的に約 12 分に相当する 41 クリップから成るデータセットを構成し，Train/Val/Test のデータをそれぞれ 32/2/7 クリップ用意した．

以上により，提案モデルは音声から BlendShape 係数（52 次元）を推定し，3D アバターを表情駆動するシステムを構成する．

4　実験

本節では提案手法が音声から BlendShape 係数をどの程度正確に推定できたか，および，リアルタイム運用に十分な計算効率を持つかを評価する．

4.1　BlendShape の精度

テストクリップ（7 クリップ．合計約 106 秒の音声）における全 BlendShape・全フレームの MSE/MAE を表 1 に示す．MSE（Mean Squared Error; 平均二乗誤差）は予測値と実際の値との差の二乗平均で誤差の大きさを表す指標であり，値が小さいほど誤差が少ないことを表す．また，MAE（Mean Absolute Error; 平均絶対誤差）は予測値と実際の値との差の絶対値を平均した指標で，MSE と比べて外れ値の影響を受けにくく，平均的な誤差量を表す．いずれの指標も値が小さいほど推定精度が高い．表 1 より，提案手法は Audio2Face-3D と比較して MSE が 0.074 から 0.035 へ低下し（差 0.039），MAE も 0.147 から 0.111 へ低下した（差 0.036）．本テスト条件では，提案手法が BlendShape の平均誤差を低減する傾向が確認できた．

3）　本研究におけるモデルの学習と評価に利用したデータセットは右記からダウンロードできる：https://huggingface.co/datasets/nvidia/Audio2Face-3D-Dataset-v1.0.0-claire

表 1　Audio2Face-3D と提案手法の係数レベル性能比較

モデル	MSE↓	MAE↓
A2F-3D	0.074	0.147
Facial Motion Generator	0.035	0.111

A2F-3D: Audio2Face-3D.
Facial Motion Generator: 提案手法.

表 2　モデル規模と推論時間の比較

モデル	パラメータ数	MACs	時間 [ms/frame]
A2F-3D	181M	2.84G	21.5
Facial Motion Generator	0.723M	0.0151G	0.62

A2F-3D: Audio2Face-3D.
Facial Motion Generator: 提案手法.

4.2　モデル規模と推論効率

モデル規模と推論時間を表 2 に示す．Audio2Face-3D は 181 M パラメータ，2.84 GMACs を要し，推論時間は約 21.5 ms/frame であった．これに対し，提案手法は約 0.72 M パラメータ，0.0151 GMACs と 2 桁以上軽量であり，推論時間は約 0.62 ms/frame であった．これらの結果から，提案手法は Audio2Face-3D に比べて，BlendShape の係数誤差を削減しつつ，パラメータ数・計算量・推論時間のいずれにおいても約 2 桁の削減を達成していることが分かった．計算資源が限られた環境において多数のアバターを同時駆動するシステムや，スマートフォンやヘッドセットなどの端末上でのリアルタイム運用において，大きな利点が期待できる．

4.3　結果の考察

表 1 および表 2 の結果，提案手法は Audio2Face-3D より小さい MAE/MSE を示しつつ，パラメータ数・MACs・推論時間を 1〜2 桁削減できている．一方で，生成された動作に対する主観的な自然さという観点では，発話に同期した顎開閉や口唇形状など口周りの係数は変動するものの，眉・頬・眼瞼といった顔の上半分の計数変動が小さく，結果として口元しか動かない単調な表情になりやすい傾向が観察された．このため，瞬きや視線の変化，頬・眉の微細な揺らぎなどの非言語表現が十分でなく，表情全体の豊かさという点では Audio2Face-3D の方が優れて見える場面があった．

5　まとめと今後の課題

本稿では，音声のみを入力として 52 次元の BlendShape 係数をリアルタイムに推定する軽量 Cross-Attention モデルである **Facial Motion Generator** を提案した．

Audio2Face-3D が学習に用いたデータの一部である Claire 話者のデータセットから生成した BlendShape を用いて比較した結果，提案手法は約 0.7 M パラメータ・0.0151 GMACs という軽量な構成で，Audio2Face-3D（181 M パラメータ・2.84 GMACs と比べて係数誤差が小さい値を示した．また，推論時間は 0.62 ms/frame（Audio2Face-3D は 21.5 ms/frame）であり，測定環境（RTX 5070 Ti）において約 35 倍（21.5/0.62 ≈ 34.7）の高速化を確認した．これにより，係数推定を多数同時に実行するようなリアルタイム運用において，計算資源の観点から有効であることが示された．

一方で，本研究には以下の制約がある．第一に，評価が単一話者に限られており，多話者・多ドメイン条件への一般化性能は未検証である．第二に，動作に対する主観評価を実施していないため，係数誤差の違いが知覚的な自然さに影響するかどうかは確認できていない．第三に，実用上の最大の課題として，現状の提案手法は口周りの動きは生成できる一方で，眉・頬・眼瞼といった顔の上半分の変化が乏しく，結果として口元しか動かない単調な表情になりやすい．以上の点は，今後のデータ拡張，評価設計，および上半顔運動の生成強化により改善する必要がある．

今後は，多話者・多言語・多様な話し方を含むデータセットへの拡張に加え，話者や発話スタイルを条件とした表情生成手法の検討に取り組む．さらに，音声認識・言語理解・応答生成を担う対話システムや大規模言語モデルとの統合を進める．また，生成された表情に対する主観評価実験を実施し，提案手法がエージェント対話に与える影響を検証する．これらを通じて，発話内容と同期した表情制御を実現し，音声のみでも自然な対話体験を提供可能なマルチモーダル対話エージェントの基盤技術としての有効性を明らかにしたい．

参考文献

[1] Catherine Oh Kruzic, David Kruzic, Fernanda Herrera, and Jeremy Bailenson. Facial expressions contribute more than body movements to conversational outcomes in avatar-mediated virtual environments. **Scientific Reports**, Vol. 10, No. 1, p. 20626, 2020.

[2] Nadine Aburumman, Marco Gillies, Jamie A. Ward, and Antonia F. de C. Hamilton. Nonverbal communication in virtual reality: Nodding as a social signal in virtual interactions. **International Journal of Human-Computer Studies**, Vol. 164, p. 102819, 2022.

[3] Ike Nnoli and NVIDIA. Nvidia open sources audio2face animation model. https://developer.nvidia.com/blog/nvidia-open-sources-audio2face-animation-model/. Published: 2025-09-24, Accessed: 2025-12-23.

[4] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, 2022.

[5] 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 (CVPR)**, 2019.

[6] NVIDIA. Overview — audio2face-3d. https://docs.nvidia.com/ace/audio2face-3d-microservice/1.2/text/getting-started/overview.html. Accessed: 2025-12-23.

[7] Dmitriy Pinskiy. Sliding deformation: Shape preserving per-vertex displacement. In **Eurographics 2010 - Short Papers**, 2010. Accessed: 2025-12-25.

[8] Pushkar Joshi, Wen C. Tien, Mathieu Desbrun, and Frédéric Pighin. Learning controls for blend shape based realistic facial animation. In **Eurographics/SIGGRAPH Symposium on Computer Animation (SCA)**, 2003. Accessed: 2025-12-25.

[9] John P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. Practice and theory of blendshape facial models. In **EUROGRAPHICS 2014 - State of the Art Reports (STAR)**, 2014. Accessed: 2025-12-25.

[10] Ladislav Kavan. Siggraph course 2014 — skinning: Real-time shape deformation, part i: Direct skinning methods and deformation primitives. SIGGRAPH Course Notes, 2014. Accessed: 2025-12-25.

[11] Blender Foundation. Introduction — bones (blender manual 3.6). Blender Manual, 2025. Last updated 2025-03-29, accessed 2025-12-25.

[12] Apple. Face tracking with arkit blendshapes. https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapes?language=objc&utm_source=chatgpt.com. Accessed: 2025-12-23.