日本語 Full-duplex 対話モデルの拡張による発話と身体動作のリアルタイム同時生成に向けた取り組み

Towards Real-time Simultaneous Generation of Speech and Body Movements through Extension of Japanese Full-Duplex Dialogue Models

津田太郎 [1*]　姜菁菁 [1]　東中竜一郎 [1]
Taro Tsuda[1], Jingjing Jiang[1], and Ryuichiro Higashinaka[1]

[1] 名古屋大学大学院情報学研究科
[1]Graduate School of Informatics, Nagoya University

Abstract: 対話ロボットには対話能力だけでなく，自然な身体動作が求められる．しかし，対話内容に適した動作をリアルタイムに生成できるシステムは未だ少ない．そこで本研究では，日本語 Full-duplex 対話モデル J-Moshi を拡張し，発話と身体動作を同時に生成するマルチモーダル対話システムの構築を試みる．具体的には，話者の姿勢情報を音声・テキストと同時に学習させることで，発話と同期した多様な身体動作の生成に取り組み，予備的な結果について報告する．

1　はじめに

近年，人間と自然なコミュニケーションを行う人型の対話ロボットや CG アバターなど，身体性を持つ対話エージェントの研究が盛んに行われている．これらのシステムにおいて，ユーザとの親和性を高め，自然な対話を実現するためには，言語的な応答能力だけでなく，ジェスチャや視線などの非言語的な身体動作が重要な役割を果たす [1, 2]．特に，発話のリズムや内容に同期した自然な身体動作は，対話の臨場感を向上させる上で不可欠である．しかし，対話の文脈や韻律に適した動作を，音声と同期させてリアルタイムに生成できるシステムは未だ少ない．

そこで本研究では，日本語 Full-duplex 対話モデルである J-Moshi [3] を拡張し，発話音声・テキストと身体動作を End-to-End で同時に生成可能なマルチモーダル対話システムの構築を試みる．J-Moshi は音声とテキストをトークンとして扱うことで，低遅延かつ自然な音声対話を実現している．本研究では，この枠組みに話者の「姿勢情報」を新たなモダリティとして統合する．具体的には，独自に収集した大規模な映像付き対話データから身体動作を抽出し，音声・テキスト・姿勢情報を同一の Transformer モデルで学習させることで，文脈を考慮しつつ，発話と高度に同期した多様な身体動作の生成を目指す．本稿では，そのための要素技術として，大規模な対話映像を用いた学習データの構築と，姿勢情報の離散化（トークナイズ）について報告し，発話と身体動作のリアルタイム同時生成の予備的な結果について報告する．

*連絡先：名古屋大学大学院情報学研究科
〒 464-8601 名古屋市千種区不老町
E-mail: tsuda.taro.d7@s.mail.nagoya-u.ac.jp

2　関連研究

2.1　Full-duplex 音声対話システム

従来の対話システムは，ユーザの発話終了を検出してから応答を生成するターンベースの処理が主に用いられていた [4]．しかし，これらは割り込みや相槌といった人間らしい自然な会話のダイナミクスを欠いているという課題があった．一方，近年では音声とテキストをトークンとして統合的に扱い，双方のストリームを同時並列的に処理する Full-duplex 対話モデルが提案されている [5]．例えば，Moshi [6] は，音声を離散的なトークンに圧縮し，テキストトークンと共に Transformer で自己回帰的に生成することで，低遅延な対話を実現している．また，これを日本語に適応した J-Moshi [3] も構築されており，自然な日本語対話が可能となっている．しかし，これらのモデルはあくまで「音声」と「テキスト」の生成に留まっており，ジェスチャや視線のような，対話における非言語情報を生成することはできない．そこで本研究では，J-Moshi の枠組みを拡張し，音声・テキストに加え身体動作も統合的に生成可能なマルチモーダル対話システムの構築を目指す．

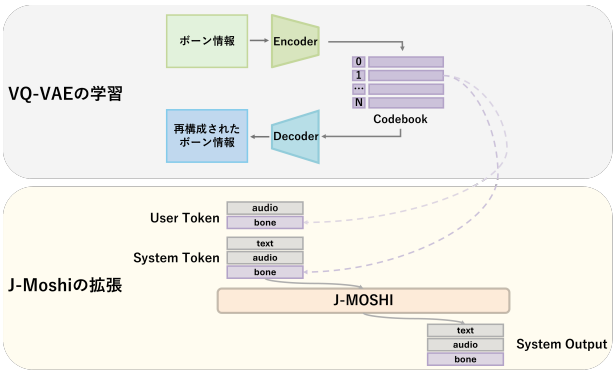


図 1: 提案システムの全体像．J-Moshi のテキスト・音声ストリームに加え，VQ-VAE により離散化された動作トークンを統合し，マルチモーダルな同時生成を行う．

2.2 身体動作の生成

発話に同期した身体動作の生成は，音声やテキストを条件として動作系列を生成する課題として広く研究されている．例えば，音声から話者のジェスチャを推定する手法 [7] や音声と言語特徴を統合する手法 [8]，拡散モデルを用いる手法 [9] などが提案されている．しかし，これらの多くは与えられた音声やテキストに追従して動作を生成する設計が中心であり，発話と非言語行動を同時にリアルタイムで生成する枠組みは限定的である．

本研究は，動作を離散トークン化し顔運動を自己回帰的に生成する CodeTalker の設計 [10] から着想を得て，顔運動に代えて動作トークンを学習し，J-Moshi に統合することで，発話生成と身体動作生成のリアルタイム同時生成を目指す．

3 提案手法

本研究では，音声・テキスト・姿勢情報を統合的に扱うため，J-Moshi のアーキテクチャを拡張する．提案手法は図 1 に示すように，(1) 姿勢情報を離散トークン化するための Vector Quantized Variational AutoEncoder (VQ-VAE [11]) の学習，(2) 獲得した動作トークンを用いた J-Moshi の拡張，の 2 段階から構成される．

3.1 VQ-VAE による姿勢情報のトークン化

J-Moshi のような Transformer ベースのモデルで連続的な身体動作を扱うためには，ボーン座標データを離散的なトークン列に変換する必要がある．そこで本研究では，まず姿勢情報のみを用いて，動作の圧縮・復元を行う VQ-VAE を学習する（図 1 上段）．具体的には，時刻 t における全身の関節座標（33 点 $\times$3 次元）をエンコーダに入力し，潜在表現を得る．これをサイズ $N = 128$ のコードブックを用いて量子化し，デコーダを経て元の座標を復元する．この学習済みの VQ-VAE は，連続的な姿勢情報を，音声トークンと同様の離散的な ID 列（動作トークン）へと変換する役割を担う．

3.2 J-Moshi の拡張

J-Moshi の RQ-Transformer は，大規模な Temporal Transformer（7B 規模）と，小規模な Depth Transformer から構成される．本研究では，音声と同期した身体動作を入出力するために，RQ-Transformer に動作トークンを扱うモジュールを追加する（図 1 下段）．Temporal Transformer は，12.5 Hz のレートで時間方向のトークン列をモデル化する．元の J-Moshi では，テキスト（1 階層），システム音声（8 階層），ユーザ音声（8 階層）からなる計 17 階層のトークン列を扱う．本研究ではさらに，システムの動作トークン（8 階層）とユーザの動作トークン（8 階層）を追加し，合計 33 階層のトークン列として扱う．

各時間ステップ s において，Temporal Transformer は直前のステップ $s-1$ までの履歴トークン（33 階層）を入力として，埋め込みベクトル $\mathbf{z}_s$ を出力する．テキストトークンは $\mathbf{z}_s$ に基づき生成される．Depth Transformer は，$\mathbf{z}_s$ およびテキストトークンを条件として，同一ステップ s におけるシステム音声トークン 8 個とユーザ音声トークン 8 個を自己回帰的に生成する．

さらに，身体動作を生成するための予測モジュールを追加する．具体的には，システム用およびユーザ用に，それぞれ 8 個のトークンに対応した MLP を備える．$\mathbf{z}_s$ とテキストトークンおよび生成された音声トークンを入力として，対応する動作トークンの確率分布を出力し，トークンをサンプルする．これにより，同一ステップ内で複数の動作トークンを低遅延に推定できる．

4 実装と評価実験

提案手法の有効性を検証するため，大規模な対話データセットから姿勢情報の抽出を行い，学習された VQ-VAE による再構成誤差の評価実験を行った．

4.1 姿勢情報の抽出

学習データの構築にあたり，本研究では Zoom を用いて独自に収集した対話映像データセットを使用した．被験者は計 20 名であり，各自が Web カメラ等を用いて自身を撮影しながら，様々なトピックに関する雑談

図 2: 対話の様子の例.

を行った．対話の様子の例を図 2 に示す．なお，立位での収録が難しいケースがあることから，一部のデータには着座状態の話者が混在している．

データセットは合計 500 対話からなり，総録画時間は約 148 時間に及ぶ．1 対話当たりの収録時間は平均約 18 分（最大 26 分，最小 16 分）であり，映像のフレームレートは 25fps である．データの前処理として，録画の開始・終了時における対話外の動作や人物の不在を除去するため，全データに対して開始から 140 秒，および，終了前の 20 秒をトリミングした．なお，録画が正常に終了しなかった 2 対話については，例外的に終了前の 5 分，および，2 分をそれぞれ削除した．姿勢情報の抽出には Python の MediaPipe ライブラリ[1] を用い，各フレームの映像から全身 33 点の 3 次元ランドマーク（pose_world_landmarks）を取得した．この座標系では，腰の中心を原点としたメートル単位の相対座標で表現されており，カメラの位置や画角の影響を受けにくくなっている．

4.2 VQ-VAE の学習

取得したランドマークを用いて，動作の離散表現を獲得するための VQ-VAE モデルを学習した．学習にあたり，全 500 対話を対話単位で学習用（400），検証用（50），テスト用（50）に分割した．コードブックサイズは全身動作の主要なパターンを捉えるため $N = 128$ とし，埋め込み次元は $C = 64$ とした．学習は 100 エポック行い，最適化手法には Adam（学習率 8×10^{-4}, $\beta_1 = 0.5$, $\beta_2 = 0.999$）を採用した．また，学習率のスケジュールとして，20 エポックごとに減衰率 0.7 のステップ減衰を適用した．

[1] https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker

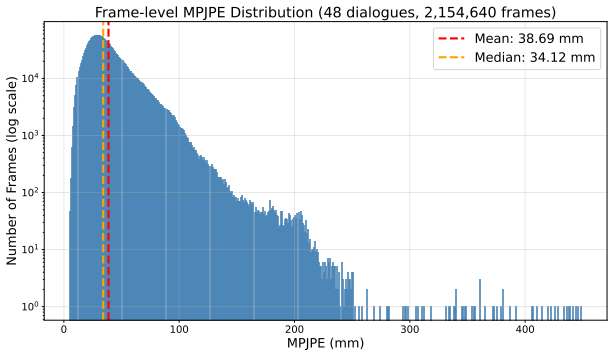


図 3: MPJPE のヒストグラム．縦軸が対数尺度となっていることに注意されたい．

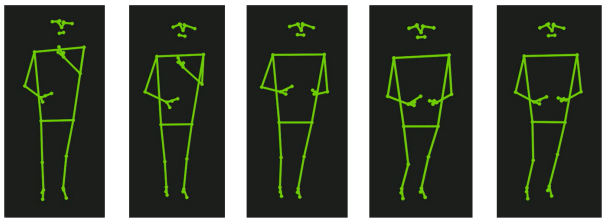
4.3 学習した VQ-VAE の評価

学習済みモデルの性能を定量的に評価するため，評価指標として平均関節位置誤差（Mean Per Joint Position Error: MPJPE）を用いた．これは，学習に使用していないテストデータの各フレームに対し，元のランドマーク座標と，VQ-VAE により一度トークン化してから再構成されたランドマーク座標とのユークリッド距離を平均したものである．評価にあたり，話者がカメラに対して横向きとなっていた 2 対話については，本モデルが想定する入力分布外のデータと見做し，除外した．その結果，テストデータにおける MPJPE は平均 38.7mm，中央値 34.1mm となった．

図 3 にフレームごとの MPJPE の分布のヒストグラムを示す．ヒストグラムのピークは低誤差領域に位置しており，大半のフレームにおいて高精度な再構成が可能であることを示している．一方で，250mm を超える大きな誤差が観測されたフレームについて確認を行ったところ，話者がカメラに極端に接近している場面や，伸びをする動作，頭を深く下げる動作が確認された．このことから，学習データに含まれる頻度が低い特異な姿勢や，遮蔽によるランドマーク抽出の不安定さが要因で誤差が高くなったと考えられる．また，定性的な確認においても，対話中の手振りのような主要な動作は，おおむね正確に再構成されていることが確認された．これらの結果は，提案手法による離散表現が，対話モデルの学習データとして十分な表現能力を有していることを示唆している．

4.4 J-Moshi による生成の試行

最後に，J-Moshi の拡張実装を行い，マルチモーダル生成の動作検証を行った．図 4 に，構築したシステムにより生成された身体動作（システム側）の例を示す．この実装により，システムがテキスト，音声，身体動作を同時生成すること自体は確認できた．しかし，



「実はお金の話で…」「お金がほらそう，なんだろう…」

図 4: 拡張した J-Moshi により生成された身体動作および発話の例.

いくつかの課題も明らかになった．第一に，身体動作モダリティを追加した影響により，生成される音声の品質が低下する現象が確認された．具体的には，生成初期は日本語として成立するものの，時間の経過とともに言語構造が崩れる傾向が見られた．第二に，生成される身体動作が学習データに含まれる静止状態に引きずられ，動きが乏しくなるケースなどが観測された．これらの結果から，マルチモーダル同時生成の実現には，単純なトークンの追加だけでなく，モダリティごとの損失関数の重みづけ調整や，モデルサイズの拡張など，詳細なチューニングが必要であることが示唆された．

5　おわりに

本研究では，Full-duplex 対話モデル J-Moshi を身体的対話へと拡張するための取り組みについて述べた．具体的には，大規模対話データから姿勢情報の抽出を行い，VQ-VAE を用いて姿勢情報の離散トークン化を実現した．さらに，得られた動作トークンを用いた J-Moshi の拡張実装を行い，生成実験を試みた．その結果，テキスト・音声・身体動作の同時生成は実現できたものの，音声品質の低下や，動作生成の不安定さといった課題が明らかになった．この予備的な結果を受け，今後の課題は，これらの問題を解決するための学習ハイパーパラメータの調整およびモデルの最適化をしたいと考えている．特に，高品質な音声を生成する能力を維持しつつ，自然な動作を生成するための，マルチモーダル学習手法の確立を目指す．

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．本研究で利用した対話コーパスは，株式会社アイシンとの共同研究において構築しました．また，本研究では，名古屋大学のスーパーコンピュータ「不老」を利用しました．

参考文献

[1] Justine Cassell, Yukiko I. Nakano, Timothy W. Bickmore, Candace L. Sidner, Charles Rich.: Non-verbal cues for discourse structure, *In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 114–123 (2001)

[2] Nikolaos Mavridis.: A review of verbal and non-verbal human–robot interactive communication, *Robotics and Autonomous Systems*, pp. 22–35 (2015)

[3] Atsumoto Ohashi, Shinya Iizuka, Jingjing Jiang, Ryuichiro Higashinaka.: Towards a Japanese Full-duplex Spoken Dialogue System, *In Proceedings of the 26th Interspeech Conference*, pp. 14570–24839 (2025)

[4] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, Xipeng Qiu.: SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities, *arXiv preprint arXiv:2305.11000*, (2023)

[5] Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, Xie Chen.: Language model can listen while speaking, *In Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 24831–14580 (2025)

[6] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, Neil Zeghidour.: Moshi: a speech-text foundation model for real-time dialogue, *arXiv preprint arXiv:2410.00037*, (2024)

[7] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, Jitendra Malik.: Learning individual styles of conversational gesture, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3497–3506 (2019)

[8] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, Hedvig Kjellström.: Gesticulator: A framework for semantically-aware speech-driven gesture generation, *In Proceedings of the 2020 International Conference on Multimodal Interaction*, pp. 242–250 (2020)

[9] Anna Deichler, Shivam Mehta, Simon Alexanderson, Jonas Beskow.: Diffusion-Based Co-Speech Gesture Generation Using Joint Text and Audio Representation, *In Proceedings of the 25th International Conference on Multimodal Interaction*, pp. 755–762 (2023)

[10] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, Tien-Tsin Wong.: CodeTalker: Speech-Driven 3D Facial Animation With Discrete Motion Prior, *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12780–12790 (2023)

[11] Aaron van den Oord, Oriol Vinyals, Koray Kavukcuoglu.: Neural Discrete Representation Learning, *In Proceedings of the 31st Conference on Neural Information Processing Systems*, pp. 6309–6318 (2017)