Hierarchical SQ-VAEによる発話にともなうジェスチャの離散表現の獲得

Learning Discrete Representation of Speech-Accompanied Gestures Using Hierarchical SQ-VAE

長谷川大[1*] 金子直史[2] 白川真一[3]
Dai Hasegawa[1] Naoshi Kaneko[2] Shinichi Shirakawa[3]

[1] 北海学園大学工学部生命工学科
[1] Department of Life Science and Technology, Faculty of Engineering, Hokkai Gakuen University
[2] 東京電機大学未来科学部情報メディア学科
[2] Department of Information Systems and Multimedia Design, School of Science and Technology for Future Life, Tokyo Denki University
[3] 横浜国立大学大学院環境情報研究院
[3] Faculty of Environment and Information Sciences, Yokohama National University

Abstract: 近年，深層学習を用いて音声やテキストから発話にともなうジェスチャを生成する試みが行われているが，ジェスチャのデータ表現を事前に工夫することでより良好な生成結果が得られることが予想される．本稿では，Hierarchical SQ-VAE を用いて，ジェスチャ生成に利用するためのジェスチャ表現を獲得することを目的とする．約 210 分の日本語発話音声と姿勢データをペアとしたデータから，20fps にダウンサンプリングした姿勢データの両腕 8 関節分のデータを抽出し，本研究のデータセットとした．また姿勢データは 1 フレーム 72 次元の姿勢ベクトルとする前処理を行った．本データに対して，20 フレーム分の姿勢ベクトル系列を 504 次元の離散表現に変換し，離散表現から元の姿勢系列を再構築するよう Hierachical SQ-VAE の学習を行なった．学習後のモデルにより，離散表現への変換およびジェスチャ系列の再構成を試みた結果，元のジェスチャの特徴を捉えた再構築結果が確認された．

1 はじめに

人間同士のコミュニケーションにおいて，発話にともなうジェスチャは，話者がスムーズに発話を行うために必要なだけでなく，聞き手にとっても発話内容の補完や強調をする機能や，話者の印象や伝達内容の信頼性に影響を与えるなど様々な役割を果たしている．そのため，バーチャルエージェントやヒューマノイドロボットに代表される人型のインタフェースには，発話の意味内容に則した適切なジェスチャを産出することが求められている．とりわけ，近年の大規模言語モデルの発展により人間のような自然な応答文を生成できるようになったことから，人型の対話インタフェースの実用化にあたり，応答文生成と同時に自動的にジェスチャも生成されることが望ましい．

*連絡先：北海学園大学工学部生命工学科
〒 062-0911 北海道札幌市豊平区旭町４丁目１－４０
E-mail: dhasegawa@hgu.jp

ジェスチャ生成手法の既存研究では，これまでにルールベースのアプローチが提案されている．ルールベースの手法では事前に準備されたドメイン知識を与えることで，限定的なタスクにおいては発話の意味内容に応じた適切なジェスチャを生成できることが示されている [3, 4]．また近年では，ドメインを強く限定しないジェスチャ生成手法として，音声やテキストとジェスチャ動作をペアとしたデータセットから深層学習を用いて生成モデルを構築する試みが行われている．

深層学習を用いた初期的な試みとして，日本語の発話音声と 3 次元ジェスチャデータを学習した Bi-Directional LSTM による生成モデル [1] や，英語の発話音声と 2 次元ジェスチャデータを学習した Conditional GAN による生成モデル [5] が提案されていたが，これらは発話音声に対して自然な動作は生成できるものの，意味内容との対応は十分ではなかった．しかし現在では，Diffusion Probabilistic Model による生成モデル [6] や，

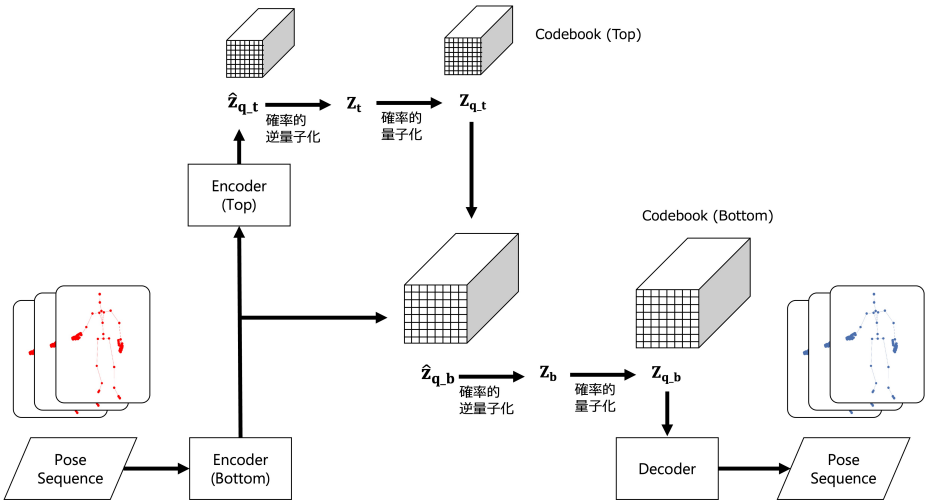

図 1: Hierarchical SQ-VAE

Transformer による生成モデル [7] など多数の生成モデルが提案されており，生成されるジェスチャの自然性が向上しているだけでなく，音声に加えてテキストの埋め込み表現を入力に用いることで発話意味内容とのマッチングも良好になってきている．

様々な生成モデルが検討される一方で，ジェスチャのデータ表現をみると，各関節の位置情報や 3 次元回転角度など素朴な表現が使われることが多いため，ジェスチャの特徴をより的確に表現できる表現系を考案することでジェスチャ生成の精度も向上すると考えられる．Yazdian ら [8] は，VQ-VAE を用いてジェスチャの離散表現を獲得し，その離散表現を用いて音声からジェスチャを生成する手法を提案している．しかしながら，VQ-VAE は Encoder の出力が単一もしくはごく少数のコードブックにマッピングされてしまう現象（コードブック崩壊）が起き，学習が安定しない問題が知られている．そこで VQ-VAE のコードブック崩壊問題を改善し，安定した学習が可能な SQ-VAE[9] が提案されている．VQ-VAE が Encoder の出力に最も近いコードブックを決定的に選択するのに対して，SQ-VAE は確率的にコードブックを選択することでコードブック使用率を上げることができると考えられている．本研究では，SQ-VAE を用いてジェスチャ生成に利用するためのジェスチャ動作の離散表現を獲得することを目的とする．

2 提案方法

2.1 データセット

本研究のデータセットとして，約 210 分の日本語発話音声と bvh 形式の姿勢データをペアとしたデータセットを用いた．このデータセットには，1,047 発話，約 210 分の音声データと対応する全身 64 関節の 3 次元姿勢データが収録されており，そのうちの 765 発話をトレーニングセット，192 発話をバリデーションセット，残りの 90 発話をテストセットとした．

また本研究では，このデータセットからジェスチャデータのみを利用した．またジェスチャデータは両腕 8 関節の情報のみを利用し，すべて 20fps にダウンサンプリングを行った．さらに，各関節の 3 軸回転角度を 3 x 3 回転行列表現（同時座標系は含まない）に変換することで，1 フレームの姿勢を 72 次元ベクトルとして表現し，この姿勢ベクトルを 5 フレームおきに 20 フレームずつ取り出すことで，ジェスチャ系列のデータセットを作成した．

2.2 ネットワーク構成

図 1 に，本研究で提案する Hierarchical SQ-VAE の構造を示す．Encoder には注意機構を含む複数の 1 次元畳み込み層からなるニューラルネットワーク（bottom: 6 層, top: 4 層），Decoder には全結合層と GRU 層からなる 4 層のニューラルネットワークを用いた．また bottom 層のコードブックは 504 次元，コードブックサイズは 500 とし，top 層のコードブックは 72 次元，サイズは 256 とした．

1 バッチ毎の学習時の再構成誤差およびコードブック使用率の推移を図 2 に示す．図 2 の perplexity(bottom) および perplexity(top) は，コードブック使用率の指標の推移を示している．学習が進むにつれて使用率が増加しており，Encoder 出力がすべて 1 つのコードブックにマッピングされてしまうコードブック崩壊は生じていないことがわかる．また，perplexity(bottom) は

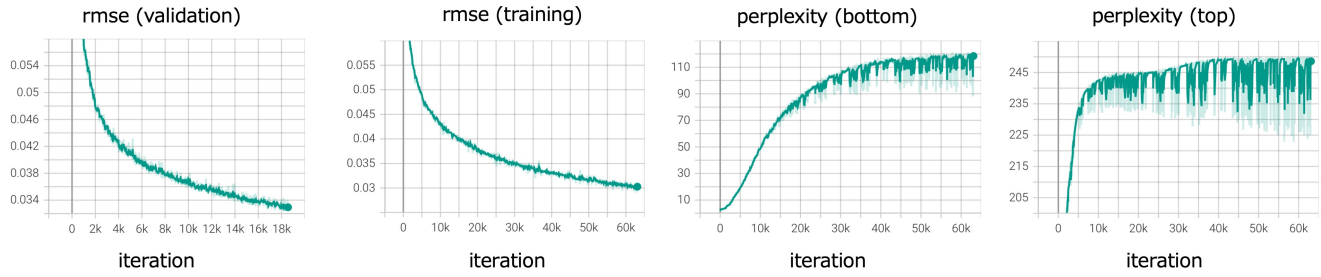

図 2: 再構成誤差およびコードブック使用率の推移

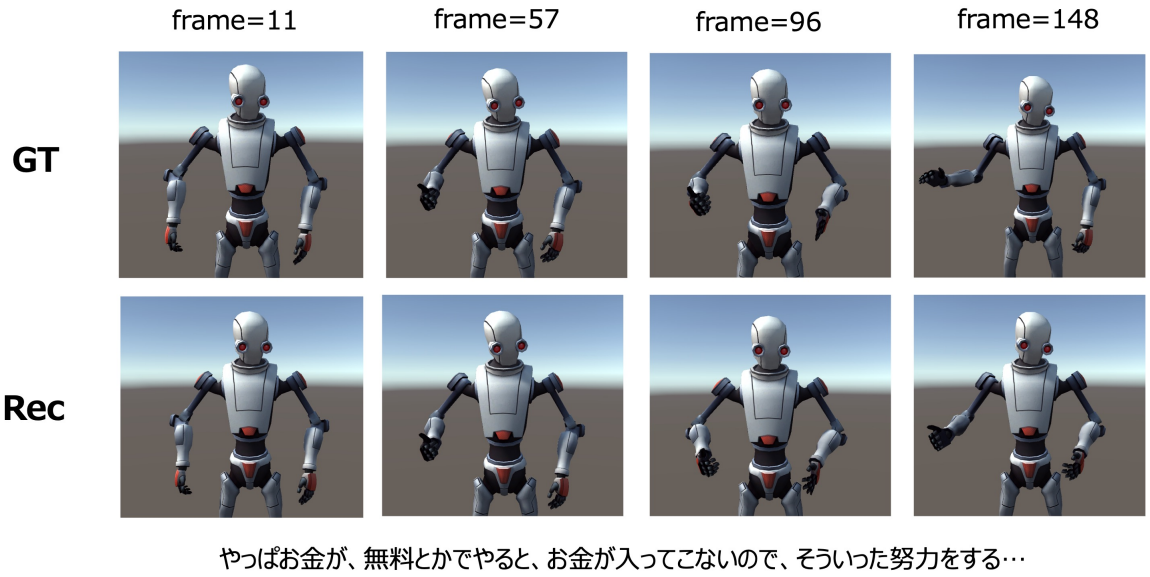

図 3: 再構成結果

およそ 120，perplexity(top) はおよそ 250 が最大値となって収束していることから，bottom 層では 120 種類程度のコードブックが使用されており，また top 層では 250 種類程度のコードブックから使用されていることがわかる．top 層と比較して bottom 層のコードブック使用率が低いのは，量子化ベクトルの次元が大きいため，少ない種類でも表現力が十分であったためと考えられる．

3　結果と考察

学習したモデルによりテストセットの 1 発話分のジェスチャを Encoder により離散ベクトル化し，また Decoder で再構成を試みた結果を図 3 に示す．1 発話分のジェスチャの離散ベクトル化と再構成にあたっては次の手順をとった．まず先頭 20 フレームの動作系列を Encoder により離散ベクトルに変換し，これを Decoder（GRU）の初期ステートとして与えることで，20 フレームの再構成系列を生成した．また以降は重なりがないように次の 20 フレームのデータを入力として与えるが，Decoder には離散ベクトルと 1 つ前のフレームの予測姿勢を入力した．最終的に 20 フレームずつ生成された動作系列を連結することで，1 発話分のジェスチャを生成した．

図 3 の入力姿勢系列（上段）と再構成された姿勢系列（下段）を比較すると，元の姿勢データと再構成された姿勢データは完全に一致はしていないものの，元の動作の特徴が捉えられた再構成結果になっていることがわかる．一方で，20 フレーム毎に動作を生成するため，一部に動作の連続性が保たれていない部分も認められた．

4　むすび

本研究では，Hierarchical SQ-VAE を用いてジェスチャ生成に利用するためのジェスチャ表現を獲得することを目的として，72 次元 ×20 フレームの姿勢系列を，504 次元の離散ベクトルに変換し再構成するモデ

ルを構築した．学習時には安定したコードブック使用率の増加がみられた．また学習後のモデルを使用してテストセットに対してジェスチャデータの離散ベクトル化および再構築を試みた結果，元のジェスチャの特徴を捉えた再構築結果が得られることを確認した．今後，定量的な評価を行った上で，モデルの改善，および生成方法の検討を行う．

謝辞

本研究は JSPS 科研費 21K12160 の助成を受けたものです.

参考文献

[1] Hasegawa, D., Kaneko, N., Shirakawa, S., Sakuta, H., Sumi, K.: Evaluation of speech-to-gesture generation using bi-directional LSTM network, *In Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp.79–86 (2018)

[2] 金子 直史, 竹内 健太, 長谷川 大, 白川 真一, 佐久田 博司, 鷲見 和彦: Bi-Directional LSTM Network を用いた発話に伴うジェスチャの自動生成手法, 人工知能学会論文誌, Vol. 34, No. 6, p. C–J41_1–12 (2019)

[3] Cassell, J., Vilhjálmsson, H. H., Bickmore, T.: Beat: the behavior expression animation toolkit, *In Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp.477–486 (2001).

[4] Cassell, J., Kopp, S., Tepper, P., Ferriman, K., Striegnitz, K.: Trading spaces: How humans and humanoids use speech and gesture to give directions, Conversational Informatics, pp. 133–160 (2007)

[5] Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning Individual Styles of Conversational Gesture, *Computer Vision and Pattern Recognition (CVPR)*, IEEE (2019)

[6] Ng, E., Romero, J., Bagautdinov, T., Bai, S., Darrell, T., Kanazawa, A., Richard, A.: From Audio to Photoreal Embodiment: Synthesizing Humans in Conversations. arXiv preprint arXiv:2401.01885 (2024)

[7] Pang, K., Qin, D., Fan, Y., Habekost, J., Shiratori, T., Yamagishi, J., Komura, T.: Bodyformer: Semantics-guided 3d body gesture synthesis with transformer, *ACM Transactions on Graphics (TOG)*, Vol. 42, No. 4, pp.1–12 (2023)

[8] Yazdian, P. J., Chen, M., Lim, A.: Gesture2Vec: Clustering gestures using representation learning methods for co-speech gesture generation. *In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3100-3107 (2022)

[9] Takida, Y., Shibuya, T., Liao, W., Lai, C., Ohmura, J., Uesaka, T., Murata, N., Takahashi, S., Kumakura, T., Mitsufuji, Y.: SQ-VAE: Variational Bayes on Discrete Representation with Self-annealed Stochastic Quantization, *International Conference on Machine Learning*, (2022)