Bi-Directional LSTM を用いた 発話に伴うジェスチャの自動生成手法の検討

An Approach in Speech-to-Gesture Generation with Bi-Directional LSTM

竹内健太[1*]，長谷川大[2]，白川真一[3]，
金子直史[4]，佐久田博司[4]，鷲見和彦[4]
Kenta Takeuchi[1]，Dai Hasegawa[2]，Shinichi Shirakawa[3]，
Naoshi Kaneko[4]，Hiroshi Sakuta[4]，and Kazuhiko Sumi[4]

[1] 青山学院大学大学院理工学研究科
[1] Graduate School of Science and Engineering, Aoyama Gakuin University
[2] 東京工科大学メディア学部
[2] School of Media Science, Tokyo University of Technology
[3] 横浜国立大学大学院環境情報研究院
[3] Faculty of Environment and Information Science, Yokohama National University
[4] 青山学院大学理工学部
[4] Collage of Science and Engineering, Aoyama Gakuin University

Abstract: 人型の CG キャラクタをインタフェースとするシステムにおいて，発話に伴うジェスチャアニメーションを付与するために大きな労力を要している．本稿では，ディープラーニングを利用して，発話音声からジェスチャアニメーションを自動生成することを検討する．音声およびモーションキャプチャデータを同期したデータセットを作成し，Bi-Directional LSTM を含む 5 層のネットワークにより，音声特徴ベクトルから 51 関節の回転角度を推定する．

1 はじめに

発話に伴うジェスチャは，提示される内容の理解を補助する重要な役割を担っている．近年，Embodied Conversational Agent（ECA）と呼ばれる人に似た身体特徴を持つバーチャルなキャラクターをシステムやアプリケーションに実装し，ジェスチャなどの非言語情報の取り扱いを可能にすることで，人対コンピュータのインタラクションを改善する試みが多く見られる．現状，ECA の発話に伴うジェスチャを作成する手段として，手動でアニメーションを定義する，またはモーションキャプチャで取得した実際のジェスチャのデータを使用するなどが挙げられる．しかし，これらの手法は専門の知識や技術が必要となる点や設備上の制約が大きい点から，非常にコストの高い作業となっている．

一方，発話とそれに伴うジェスチャの関係性を機械学習によって学習し，その結果を利用し発話からジェスチャを自動生成する手法はユーザにとって手軽に運用できる点でメリットがある．本研究では，Bi-Directional Long Short-Term Memory（LSTM）[1] ユニットを用いたディープニューラルネットワークを使用し，ディープラーニングを用いて音声特徴からジェスチャを自動生成する手法を検討する．

通常，機械学習を用いて音声からジェスチャを自動生成する際はジェスチャの種類を定義し対応する音声に対しタグ付けを行うなど，データの前処理が必要となる場合が多い．また，ジェスチャを出現させるタイミングや順番をアルゴリズムまたは手動で決定することが必要となる．そういった作業は，特に大きなデータセットを作成する場合に非常に手間のかかるものとなる．

そこで，本研究では，発話音声の Mel-Frequency Cepstral Coefficients（MFCC）による音声特徴を入力とし，BioVision Hierarchy（BVH）形式のモーションデータを元にした，全身 51 関節の 3 次元オイラー回転角の時系列データを正解データとして 5 層からなるニューラルネットワークで学習を行うことで，音声特徴から直

*連絡先：青山学院大学大学院理工学研究科知能情報コース
〒 252-5258 神奈川県相模原市中央区淵野辺 5-10-1
E-mail: c5616160@aoyama.jp

接ジェスチャのモーションデータを生成する手法を検討する．

2　関連研究

発話音声に伴う適切なジェスチャを生成することを試みた研究は他にいくつか挙げられる．一つのアプローチとして，人間の発話時の行動の性質等から解析的に発話に対するジェスチャを割り当てる方法が存在する．Cassell ら [2] は，テキストを入力として受け取り，音声合成による音声とジェスチャに加え，視線や表情などの非言語情報を伝達する動作を生成する Behavior Expression Animation Toolkit（BEAT）を実装した．ジェスチャと非言語情報の割り当ては，人の因習的な行動についての過去の研究結果を元に定義されたルールに従い行われた．また Cassell ら [3] は後に，道案内の指示のテキストから Northwestern University Multimodal Autonomous Conversational Kiosk（NUMACK）の ECA 用の iconic ジェスチャを生成するシステムを実装した．このシステムにおいて，ジェスチャはテキストに付与された言葉の抽象的な意味表現に関する情報（大きさ，特徴，幅や高さなど）に基づき事前に定義されたルールに従い生成される．

一方で，機械学習などを用い統計的に発話の特徴から適したジェスチャを生成するアプローチも存在する．Chiu ら [4] は Deep Conditional Nerual Field（DCNF）モデルを提唱し，それを使用し発話の大きさやピッチなどの韻律的特徴とテキストを元に，既存のジェスチャに関する研究を踏まえ定義した 14 種類のジェスチャのうち最も適するジェスチャを時系列の適切な時点に割り振る予測タスクを行った．実験で高い予測精度を得ることができているが，事前に定義したジェスチャの種類以外を出力することができないという課題がある．他の手法として音声特徴から直接ジェスチャのモーションデータを生成する手法があるが，これを試みた研究はあまり多く見られない．Chiu ら [5] は音声の大きさ・ピッチなどの韻律的特徴と実際のジェスチャのモーションデータを使用し，hierarchical factored conditional restricted Boltzman machine を改良したモデルで機械学習を行いジェスチャの生成器を作成した．韻律的特徴は発話の意味的内容とは関連が薄いため，生成器は beat ジェスチャと呼ばれる，主に強調の意味合いで使われ発話の抑揚やリズムに関連のあるジェスチャを生成するように設計されている．よって，iconic，deictic，metaphoric などの発話の意味的内容に関連の深いジェスチャは対象とされていない．本研究では，この制約に囚われず，発話の意味的内容に関連のあるジェスチャも生成できることが期待できるジェスチャの生成法の検討を行った．

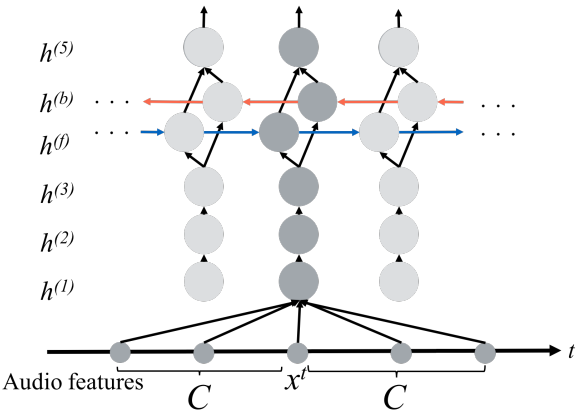

図 1: 学習ネットワーク概要図，C=25

3　データセット

学習に用いる入出力データとして，竹内らの作成した WAV 形式の音声データとそれに伴うジェスチャの BVH 形式のモーションキャプチャデータのデータセット [6] を使用した．データの記録は，事前に話すテーマと内容を決めた上，メモや資料を用いず話者が一方的に聴者に対して話すスピーチ形式で行われている．音声はマイク付きのヘッドセット，発話に伴うジェスチャは SPICE 社のモーショントラッキングシステム motive を使用して記録された．記録データは 1 センテンス毎に分割され音声とモーションデータがペアとなっている．BVH 形式のモーションデータは，関節のヒエラルキー等のスケルトンの定義と，各時間フレームにおける各関節の xyz 軸まわりのオイラー回転角の時系列データを含んでいる．この時系列データを学習の正解データとする．合計 609 センテンス（122.64 分）の音声およびジェスチャのモーションデータを学習・検証に用いた．

4　ネットワーク構成

Bi-Directional LSTM ユニットを使用したネットワークを TensorFlow[7] をバックエンドとし Keras[8] で実装した．音声認識の領域において MFCC による音声特徴を入力とし高い精度を実現した DeepSpeech[9] に用いられたニューラルネットワークの構造を大いに参考にしたものとなっている．ネットワークの構造の概要を図 1 に示す．

ネットワークは 5 層の隠れ層からなり，出力層以外の各層の後に batch normalization[10] と dropout（10%）[11] を適用する．最初の 3 層は再帰的でない全結合層となっている．1 層目 $h^{(l)}$ は 0.01 秒毎のタイムステップ t における音声データの MFCC 特徴ベクトルと前

後 25 ステップの音声特徴ベクトルを結合したベクトル $x^{(t)}$ を入力として受け取る．2，3 層目 $h^{(2,3)}$ は各タイムステップ毎の独立したデータを入力として受け取る．$h^{(1\sim3)}$ は式 1 のように計算される．

$$h_t^{(l)} = g(W^{(l)}h_t^{(l-1)} + b^{(l)}) \tag{1}$$

ここで，$g(z)$ は Rectified-Linear Unit（ReLu）の活性化関数，$W^{(l)}$ と $b^{(l)}$ はそれぞれ $h^{(l)}$ の重みの行列とバイアスパラメータである．

4 層目 $h^{(4)}$ は Bi-Directional LSTM 層となっている．この層は，式 2 に示されるように $t = 0$ から順に再帰を行う前向きの LSTM ユニット $h^{(f)}$，および式 3 に示されるように t の終わりから逆順に後ろ向きに再帰を行う LSTM ユニット $h^{(b)}$ が含まれる．

$$h_t^{(f)} = \bar{g}(W^{(4)}h_t^{(3)} + W_r^{(f)}h_{t-1}^{(f)} + b^{(4)}), \tag{2}$$

$$h_t^{(b)} = \bar{g}(W^{(4)}h_t^{(3)} + W_r^{(b)}h_{t+1}^{(b)} + b^{(4)}) \tag{3}$$

$\bar{g}$ は LSTM ユニットの計算および ReLU 関数による活性化を表す．

5 層目は出力層となっており，前の層の前向き・後ろ向き再帰のユニットからの出力を入力として受け取り，重みとバイアスと合わせて演算した後 Linear 関数 $j(z)$ で活性化を行う．

$$h_t^{(5)} = j(W^{(5)}(h_t^{(f)} + h_t^{(b)}) + b_t^{(5)}) \tag{4}$$

最終的な出力は BVH 形式において定義される全身 51 関節の xyz オイラー回転角の予測値のベクトルとなる．出力された予測結果から Mean Squared Error を算出し，Adam optimizer[12] を使用し最適化を行う．

5　実験

節 4 で示したネットワークを使用し，各層 256 の隠れユニットを定義し 100epoch 分学習を行った．学習に用いたデータのうち，9 割をトレーニングに，残りの 1 割をバリデーションで用いるように割り振った．実装したネットワークの性能を評価するため，同等のデータセットおよびパラメータを使用し RNN を使用したベースラインネットワークで学習を行い，最終的な誤差の値を比較した．また，生成されたジェスチャの有用性を検証するため印象評価の実験を行った．

5.1　実験デザイン

original，predicted，および mismatched の 3 種類のジェスチャの印象に関する評価を検証する 1 要因 3 水準の実験を行った．ここで，original とはある発話音声に実際伴っていたジェスチャのモーションデータを用いたもの，mismatched は別の音声に伴っていたジェスチャのモーションデータを音声と再生時間が一致するよう末尾からデータを削除したものである．predicted は本研究の手法によって音声から生成されたジェスチャのモーションデータの fps を削減し，間を球面線形補間したものである．これは，生成されたジェスチャの動きの躍度を軽減するための処理となっている．

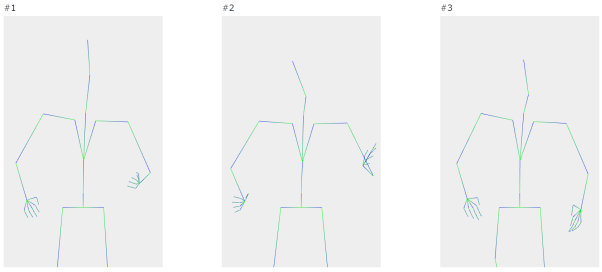

図 2: アンケート用 web ページでのジェスチャ表示例

5.2　評価方法および参加者

21~24 歳の大学生 20 名に，学習に用いていない 15 種類の音声データに対する original，predicted，mismatched の 3 種類のジェスチャそれぞれに持つ印象についてのアンケートに回答してもらった．これら 3 種類のジェスチャは，アンケート回答用の web ページにおいて横並びで表示され（図 2），対応する音声と同時に再生される．各種類のジェスチャが左，中央，右のどの位置で表示されるかは，それぞれ回数を均等に分配した上（どのジェスチャも合計左に 5 回，中央に 5 回，右に 5 回表示される）ランダムに設定される．その上，15 種類の音声の提示順序は参加者間でランダムに設定された．実験のデモを [13] にて参照できる．参加者はそれぞれのジェスチャに関して，以下に示す 3 つの設問に対し 7 段階で評価を行った．

- Q1: ジェスチャは自然だと思いますか？
- Q2: ジェスチャのタイミングは発話に対して適切だと思いますか？
- Q3: ジェスチャは実際の発話の内容に対して適切だと思いますか？

6　結果と考察

6.1　誤差比較の結果

表 1 に本研究で使用したネットワークとベースラインの RNN ネットワークの最終的な誤差を示す．音声

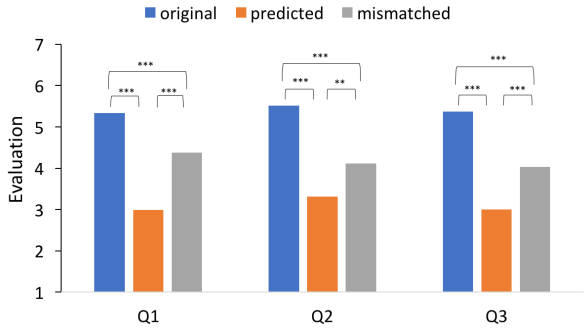

図 3: 各設問に対しての平均評価, *** p<.001, ** p<.01

表 1: 予測誤差の比較

Model	*Loss*
Bi-Directional LSTM	28.106
Simple RNN	217.477

からジェスチャを生成するタスクにおいて，本研究で使用した Bi-Directional LSTM を含む実装の方が RNN の実装より良い精度を得ることができることが分かる．

6.2 生成ジェスチャの印象評価の結果

アンケートによる印象評価の結果を図 3 に示す．ジェスチャの種類別の評価の平均に対して one-way ANOVA を行ったところ，ジェスチャの種類に主効果が検出された（Q1 に対して $F(2, 19) = 31.440$, $p < .001$, Q2 に対して $F(2, 19) = 26.545$, $p < .001$, Q3 に対して $F(2, 19) = 31.294$, $p < .001$）．また，Bonferroni の方法を下位検定として用いた t 検定による一対比較を行った結果，図 3 に示されるように，全ての設問に関して全てのジェスチャ種類間の平均評価に対して有意差が検出された．つまり，本研究の手法で生成されたジェスチャはどの設問に関しても他 2 条件のジェスチャに比べて有意に低く評価される結果となった．

このような結果となった原因はいくつか考えられる．一つは，生成されたジェスチャが他 2 種類のジェスチャに比べ頻繁に動いていたことが挙げられる．ジェスチャが必要と思われた箇所でジェスチャをしないことよりも，必要ないと思われた箇所でジェスチャをする方が低く評価されることが予想されるため，せわしなく動いていたことによりジェスチャの自然さとタイミングに対する評価が下がったと考えられる．また，ジェスチャをより人間らしくするために行った球面線形補間が結果としてあまりよいとは言えない印象を与えたことも挙げられる．人間の自然な動きには速度の緩急が存在するが，球面線形補間することにより動きが等速になる区間が多く見られるようになったため低評価に繋がったと考えられる．

7 結論

本論文では，Bi-Directional LSTM を使用したニューラルネットワークを用い学習を行い，発話音声の MFCC による音声特徴から適するジェスチャのモーションデータを直接生成する手法の検討を行った．ネットワークは，過去および未来のデータの連続性を考慮に入れ，発話音声と伴うジェスチャのモーションキャプチャデータがペアになったデータセットより音声特徴と BVH 形式におけるスケルトン定義に基づく 51 関節の xyz 軸周りのオイラー回転角との関係を学習できることが期待される．最終的な誤差の比較の結果から，本研究で用いたネットワークによる予測の精度は通常の RNN を用いた実装を大きく上回ることが示された．生成されたジェスチャは人間らしい動作の体裁は保てたが，実験より元々ある音声データに伴ったジェスチャおよび別の音声データに伴ったジェスチャに比べ印象が優位に低く評価される結果となった．

8 今後の展望

印象評価実験の結果を受け，データセットおよび学習ネットワークにいくつか改良を行った．一つ目に，単純にデータの不足により学習がうまくすすめられていなかったことが予想できたため，データを合計 1182 センテンス（392 分）に増量した．合わせて，epoch と誤差の推移を考慮し，学習時の epoch の数を 500 に新しく設定した．加えて，生成されたジェスチャが頻繁な動作をしていたことに関し，予測時のフレームレートが高すぎることが影響していると考えられたため，学習におけるタイムステップを $t = 0.01$ から $t = 0.05$ に変更した．また，発話とジェスチャの関連性を学習する際，可能な限り広い範囲を考慮できることが望ましいため，入力時に結合する前後のコンテキストを前後 30step に設定した．最後に，足にあたる関節 8 種類を予測から除外し，合計で 43 関節の回転角を予測するようにした．これらの変更に加え，音声特徴とジェスチャのモーションデータの関係性を学習するタスクにより適した形となるようにニューラルネットワークの構造やパラメータの調整を進める必要があると考える．改良後の結果に応じ，再度予測の精度や生成ジェスチャの印象評価を行う．

参考文献

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[2] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. Beat: the behavior expression animation toolkit. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 477–486. ACM, 2001.

[3] Justine Cassell, Stefan Kopp, Paul Tepper, Kim Ferriman, and Kristina Striegnitz. Trading spaces: How humans and humanoids use speech and gesture to give directions. *Conversational informatics*, pp. 133–160, 2007.

[4] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. Predicting co-verbal gestures: a deep and temporal modeling approach. In *International Conference on Intelligent Virtual Agents*, pp. 152–166. Springer, 2015.

[5] Chung-Cheng Chiu and Stacy Marsella. How to train your avatar: A data driven approach to gesture generation. In *Intelligent Virtual Agents*, pp. 127–140. Springer, 2011.

[6] Kenta Takeuchi, Souichirou Kubota, Keisuke Suzuki, Dai Hasegawa, and Hiroshi Sakuta. Creating a gesture-speech dataset for speech-based automatic gesture generation. In *HCI International 2017 – Posters' Extended Abstracts: 19th International Conference, HCI International 2017, Vancouver, BC, Canada, July 9–14, 2017, Proceedings, Part I*, pp. 198–202, Cham, 2017. Springer International Publishing.

[7] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[8] François Chollet, et al. Keras, https://github.com/fchollet/keras, 2015.

[9] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

[11] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929–1958, 2014.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[13] https://youtu.be/mas4ikgtobu.