The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

系列変換モデルを用いたプレゼンテーション動作の生成

Generation of Presentation Motion Using Sequence to Sequence Model

嶋津 章仁 *1 Akihito Shimazu
日永田 智絵 *1 Chie Hieida
長井 隆行 *1 Takayuki Nagai
中村 友昭 *1 Tomoaki Nakamura
武田 祐樹 *2 Yuki Takeda
原 豪紀 *2 Takenori Hara
中川 修 *2 Osamu Nakagawa
前田 強 *2 Tsuyoshi Maeda

*1電気通信大学 The University of Electro-Communication
*2大日本印刷 Dai Nippon Printing Co., Ltd.

In the presentation, gestures play a very important role as a means of improving the information transmission effect. Moreover, it was shown that the movement expressing the enthusiasm and intention of the presenter in an unconscious manner affects the success of the presentation and the impression on the audience. For these reasons, it is required for presentation robots to perform such movements; however, manual design of these movements is not an easy task. In this research, we propose a method to learn relationship between speech prosodic information and motion using a sequence to sequence model, and directly generate appropriate motions from prosodic information. This paper also proposes a method for generating motions representing meanings of specific words.

1. はじめに

近年，ロボットの社会進出が進んでおり，日常生活においてもロボットと人間が接する機会が増加している．そのようなロボットの例として，プレゼンテーションロボットが挙げられる．プレゼンテーションにおいて人間は様々な動きを行うが，人間が発話と共に行う動きがどのように生成されているかは，ロボットで再現できるほど十分に明らかにされていない．そのため，ロボットの動きは人手で設計するのが一般的である．人手による動きの設計は，特に自由度が高い場合には容易ではないため，プレゼンテーションの際の複雑かつ自然に見える動きを，どのように自動的に生成するかが課題となる．そこで，機械学習によって音声情報と動きを結びつけて学習させ，発話音声から自動的に動きを生成することを考える．

まず本研究では動きが，伝達内容に関する動き，文脈に関する動き，身体（個性）に依存した動き，の組み合わせによって生成されていると仮定する．つまり，これら 3 つの動きをそれぞれ実装し合成することによって，人間がコミュニケーションの際に行っている動きを再現できるのではないかと考える．本稿では，文脈をプレゼンテーションという状況に固定し，身体と伝達内容に関する動きの実現を目指す．

プレゼンテーションの目的は聴衆に情報をわかりやすく伝えることであり，プレゼンテーションロボットにも情報伝達をわかりやすく行うことが要求される．プレゼンテーションにおいて人間がジェスチャを行う理由の一つは，情報伝達を助けるためであると言われる [喜多 00]．Berger らは，ジェスチャを伴った音声と伴わない音声では，ジェスチャを伴った音声の方が聞き手の理解を約 2 倍正確にしたと報告している [Berger 71]．つまり，プレゼンテーションにおいて動きは情報伝達という意味で重要であり，前述の伝達内容に関する動きに相当する．

一方で Pentland は，ビジネスプレゼンテーションの成功が単なる言語的な情報の伝達ではなく，非言語的な情報によって意思や熱意が無意識に伝わることに関連していることを明らかにした [Pentland 08]．このことは，情報の伝達自体を目的としていない無意識に訴えかける動きの情報も重要であることを示唆している．これは，前述の身体に依存した無意識の動きに関係すると考える．例えば，発話の際には呼吸によって体が自然と動くため，そのリズムと同期して腕が動くといった物理的な現象があるはずであり，そうした発話との関係性に制約されつつも，意思や熱意が無意識に伝わるような動きがある．

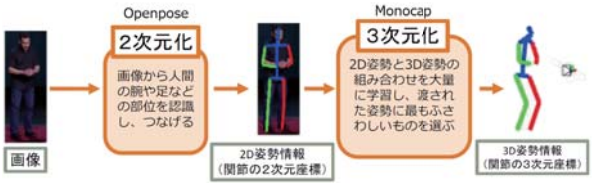

図 1: 3 次元姿勢抽出の流れ．

そこで本稿ではまず，意思や熱意を無意識のうちに表現している動き，つまりプレゼンテーションにおいて身体に依存した動きの実現に焦点を当てる．こうした動きはルール化することが難しいため，プレゼンテーション時の動きのデータを収集し，系列変換モデルによって韻律情報と動きの情報を結びつけることで自動生成する．そうすることで，プレゼンテーション時の平均的な動きが抽出され音声合成した発話の韻律に合わせて自動生成できる．この枠組みでは，学習データに多数の話者が含まれることによって個性が平均化され，より一般的な動きが生成される．逆に学習データを特定の個人に絞れば，個性を反映させることも可能である．さらに本稿では，情報伝達を助けるジェスチャ，つまりプレゼンテーションにおいて伝達内容を表現する動きについても検討する．

これまでに，アニメーションのための人の動き生成手法が多く提案されている．例えば，文献 [Alemi 15] では，Restricted Boltzman Machine を用いた情動的な動作の生成を行っている．しかし多くの研究では，発話に伴う動作の生成を考えていない．竹内らは，発話に伴うジェスチャを Bi-Directional LSTM で生成することを提案している [竹内 17]．しかし，この研究で

連絡先: 嶋津 章仁，電気通信大学大学院情報理工学研究科，東京都調布市調布ケ丘 1-5-1，ashimazu@apple.ee.uec.ac.jp

図 2: 抽出された 3 次元姿勢.

はプレゼンテーションという文脈は考えられていない．また，学習データをモーションキャプチャで収集しており，十分な数のデータによる学習が実現されていないという問題がある．

2. 提案手法の概要

提案手法は，発話音声のエネルギーとピッチから成る韻律情報を入力とし，上半身合計 9 関節の 3 次元ベクトルの時系列を出力する．これは，再帰型のニューラルネットワークを応用した，系列変換（seq2seq）モデルを学習することで実現する．ここでの問題は，大量の学習データが必要になることである．学習に必要なデータは，音声データと発表者の 3 次元姿勢データであるが，これらのデータを大量に集めるために，web 上にあるプレゼンテーションの映像から，発表者の 3 次元姿勢を抽出する手法を提案する．本稿ではさらに，情報伝達を目的としたジェスチャを生成し，系列変換モデルによって生成した動作に合成する．

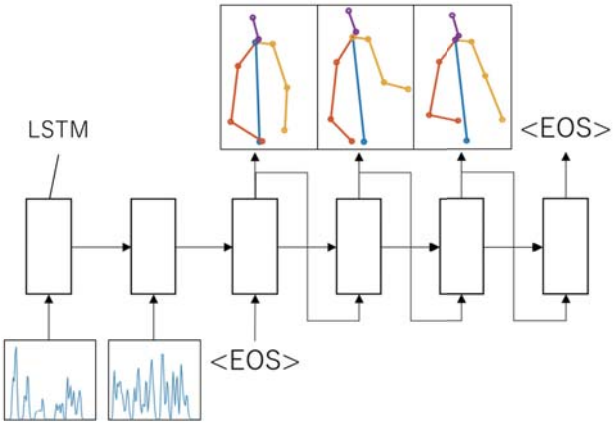

図 3: 学習モデルの概略.

3. 提案手法

3.1 学習用データセットの作成

ここでは，系列変換モデル学習のためのデータセット作成について述べる．

3.1.1 画像からの 3 次元姿勢推定

2 次元画像中の人の 3 次元姿勢推定は，OpenPose [Cao 17] による 2 次元姿勢抽出手法と，MonoCap [Zhou 16] による 3 次元姿勢推定手法を組み合わせることで実現する．これは，この二つの手法の組み合わせが精度と計算速度の点で優れているためである．図 1 に，推定の流れを示す．まず OpenPose を用いて，映像中の対象の 2 次元姿勢を取得する．その結果を MonoCap に渡し，学習した 2 次元姿勢と 3 次元姿勢の組み合わせから，似た姿勢を考慮することで 3 次元化を行う．また，プレゼンテーション動画では，スライドや観客のみが映っているフレームも多く存在する．本稿では，話者が多くの場合に中心で大きく映っているという事前知識を併用して，話者のトラッキングを行った．提案した手法で 3 次元姿勢推定を行った結果の例を，図 2 に示す．発表者の 3 次元姿勢推定の精度を評価するために，一つのプレゼン動画（約 15 分）に対して，フレーム毎に抽出の成否を人手で判定した．その結果，抽出率は 94.8% であり，プレゼンテーションの動きを学習するための精度としては十分であることが分かった．

3.1.2 韻律の抽出

音声のピッチは，森勢らによって提案された音声分析合成システム WOLRD [Morise 16] を用いて行う．また音声のエネルギーは，発表者の音声波形の各サンプルを二乗し，ピッチの計算周期で加算することで算出する．

3.2 身体に依存した動きの学習

身体に依存した動きの学習は，図 3 に示すように，音声の韻律を入力として動きを生成する系列変換モデルによって実現する．系列変換モデルは，Sutskever らによって提案された再帰型ニューラルネットワークに基づく学習モデルである [Sutskever 14]．エンコーダ—デコーダモデルとも呼ばれ，入力の時系列をベクトルにエンコードし，そのベクトルから出力の時系列をデコードする．提案手法では，入力は音声のピッチと音声エネルギーの時系列であり，これを発話文章単位でエンコードする．デコーダ側では，発表者の上半身 9 関節の 3 次

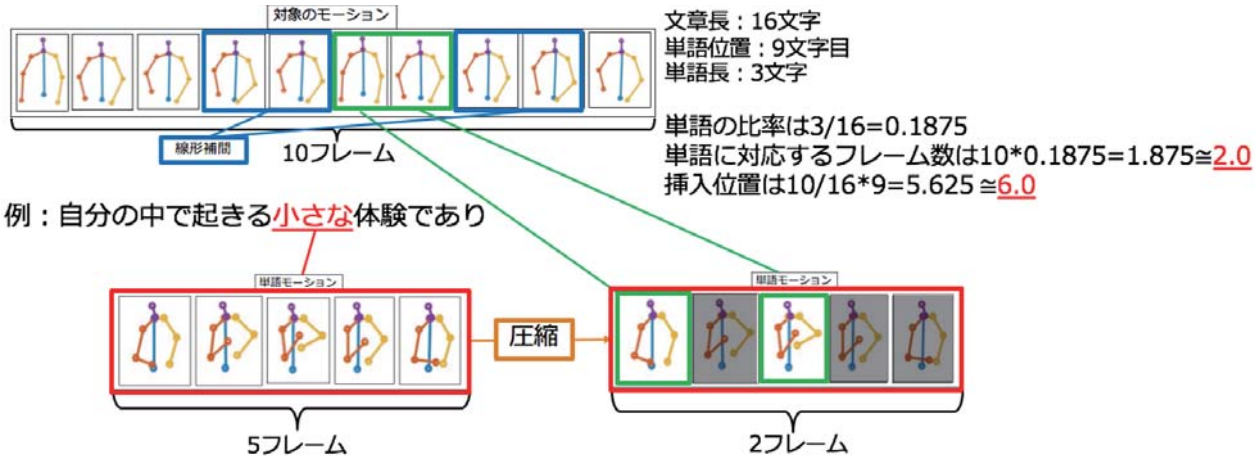

図 4: 単語的意味を表現する動きの合成手法.

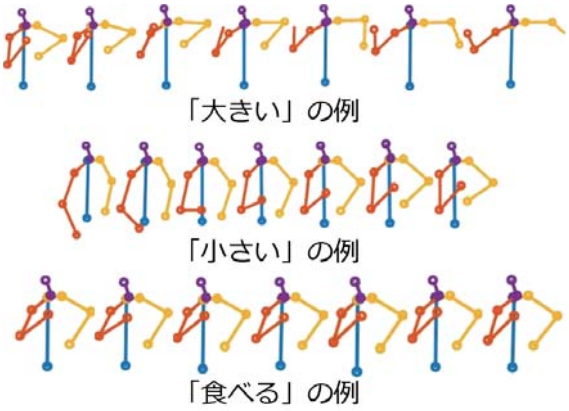

図 5: 作成した単語の意味に対応した動きの例.

元姿勢ベクトルをエンコーダ側からの情報を基に復元する．文章単位の系列変換モデルであるため，文章内での動きの滑らかさは韻律情報の滑らかさによってある程度保証されるが，文章間で不連続になる可能性がある．提案手法では，文章間の発話のないところ補間することで，文章間の姿勢を滑らかにつなぐこととする．

3.3 情報伝達を行う動きの合成

情報伝達を目的とした動きとしては様々なものが考えられるが，ここでは特定の単語の意味を表現する動きを考える．系列変換モデルによって生成された動きに対して，単語的意味を表現する動きを合成する手法の概略を図 4 に示す．文章中の対象単語の発話位置から動きを挿入する時間的位置を算出し，かつ挿入部分にモーションが収まるよう時間的に伸縮することで動きの合成を行う．なお，対象とする単語は手動で選択することとする．また，挿入する単語動作自体も前述の 3 次元姿勢抽出手法を用いて著者の動作から自動で生成したものを用いる．つまり選択した各単語に対して著者がそのイメージで動作を行い，その映像から 3 次元姿勢の抽出を行い単語動作ライブラリとして登録する．

作成した動きの一例を，図 5 に示す．図より，単語の動作を表しているような動作が取得されていることが分かる．

4. 実験

提案手法の有効性を評価するための実験について述べる．

4.1 実験設定

本稿ではデータとして，TED [TED 17] の 119 の動画を用いた．この全動画に対して，前述の 3 次元姿勢抽出と韻律の抽出を行い，それぞれ文章ごとに分割することで音声（韻律）情報と 3 次元姿勢情報の 10485 対の時系列データを得た．これは時間にして，7 時間 23 分 27 秒分のデータである．これらのデータを，学習データと評価データに分割して実験を行った．

4.2 生成モーションの評価手法

生成されたモーションをどのように客観評価するかは，重要な問題である．主観評価も必要であるが，ここではその前段として客観的な指標によって生成モーションを評価することを考える．プレゼンテーション時のモーションの自然さという視点で考えると，動きと音声の同期性が重要である．これは，音声とモーションが同一の話者で同時に生成されていることを考えれば当然であり，その同期性を観測者が「動きの自然さ」として捉えると考えることは，不自然ではない．そこで本稿では，「各関節の速さと音声エネルギーの変化量」の相関係数を指標とする．まずこの指標が使えることを示すために，前述のデータに対して相関係数を計算した．また音声とモーションをランダムに入れ替えたデータに対しても同様に計算し，それらの差異を検定したところ，統計的に優位に入れ替えのないデータの相関係数が大きかった．そこで，入れ替えのないデータの相関係数の分布と入れ替えたデータの相関係数の分布をそれぞれ正規分布で表し，その二つの分布の対数尤度比という形で生成したモーションの自然性を評価することとした．

4.3 結果

学習データを用いて系列変換モデルを学習し，そのモデルを用いて，残りの 10 個の評価用データの音声からモーションを生成した．結果の例を図 6 に示す．生成したモーションに対して，対数尤度比スコアを算出した結果を図 7 に示す．スコアが 0 より大きい場合に，音声と動きの同期性という点で自然なモーションであると言える．生成したモーションによるスコアの平均が 0.85，実際に人が動作しているデータを用いた場合のスコアの平均は 1.10 であり，提案手法による生成モー

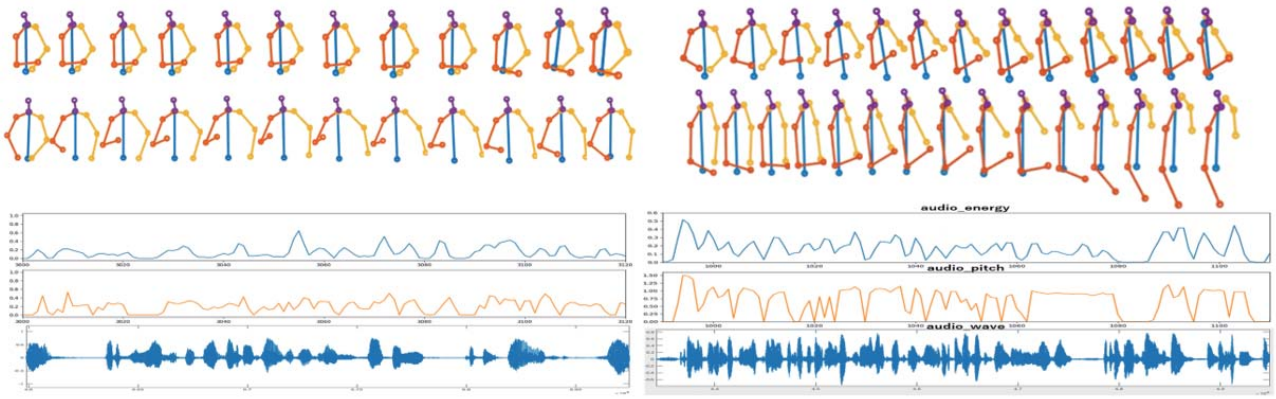

図 6: 生成モーションの一例; 上段) オリジナルの動き，中段) 系列変換モデルによって生成された動き，下段) 音声の情報.

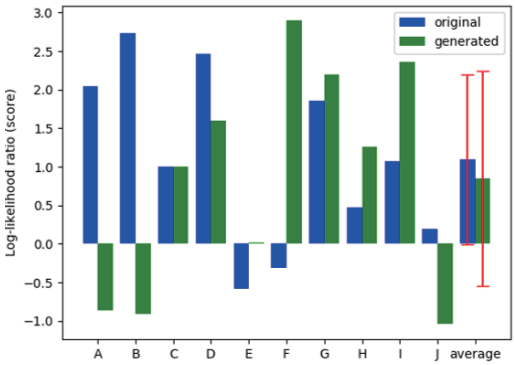

図 7: 対数尤度比スコアによる評価.

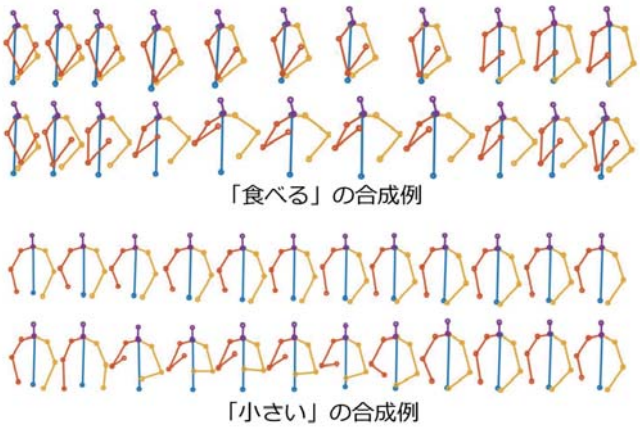

図 8: 単語の意味を表す動きの合成例; 上段) 合成前の動き，下段) 合成後の動き.

ションは，同期性という視点では人のプレゼンテーションと同等に自然であることが分かる.

さらに，系列変換モデルによって生成した動きに対して，単語の意味に応じた動きを合成した結果の例を図 8 に示す.「食べる」といった動詞や,「小さい」といった形容詞が意味的に強調された動作が生成されていることが分かる．合成した動きに対する評価は，今後の課題である.

5. まとめ

本稿では，プレゼンテーションロボット向けモーション生成手法を提案した．提案手法は系列変換モデルに基づいており，学習に必要な大量のデータをプレゼンテーション動画から自動生成した．実験によって，生成されたモーションが，音声と動きの同期性という基準で自然なモーションであることを示した．また単語の意味に相当する動きを生成し，系列変換モデルによって生成された動きと合成する手法を提案した．これにより実際に意味の強調された動きが生成できることを確認したが，今後，生成された動きを主観評価する必要がある．生成された動きのロボットへの実装も今後の課題である.

参考文献

[喜多 00] 喜多壮太朗, “特集, ジェスチャーの認知科学：ひとはなぜジェスチャーをするのか,” 認知科学 7(1), 9-21, 2000, 2000

[Berger 71] K.W.Berger *et al.*, “Extra-facial gestures in relation to speechreading,” Journal of Communication Disorders pp.302-308 1971, 1971

[Pentland 08] A.Pentland, Honest Signals, The MIT press, 2008

[Alemi 15] O.Alemi, W.Li, P.Pasquier, “Affect-expressive movement generation with factored conditional Restricted Boltzmann Machines,” ACII 2015,2015

[竹内 17] 竹内健太, 長谷川大, 白川真一, 金子直史, 佐久田博司, 鷲見和彦, “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, ” HAI2017,2017

[Cao 17] Z.Cao *et al.*, “Realtime Multi-Person 2D Pose Estimation using Affinity Fields,” CVPR2017, 2017

[Zhou 16] X.Zhou *et al.*, “MonoCap: Monocular Human Motion Capture using a CNN Coupled with a Geometric Prior,” CVPR2016, 2016

[Morise 16] M.Morise *et al.*, “WORLD: a vocoder-based high-quality speech synthesis system for real-time applications,” IEICE transactions on information and systems, 2016

[Sutskever 14] I.Sutskever *et al.*, “Sequence to Sequence Learning with Neural Networks,” NIPS2014, 2014

[TED 17] TED Conferences LLC, “TED (Technology Entertainment Design),” https://www.ted.com/, Retrieved 2017