敵対的生成ネットワークを用いた 対話中の人間の振る舞いのモデル化

西村 優佑* 中村 泰* 石黒 浩*

Human Behavior Modeling During Dialogue by Using Generative Adversarial Networks

Yusuke Nishimura*, Yutaka Nakamura* and Hiroshi Ishiguro*

In this research, human behavior during dialogue is modeled, with the goal of generating human-like motion for humanoid robots. Most of the previous studies on the human motion modeling aimed to model the single motion of human [8] and produce complicated behaviors by combining the models of the single motions [11]. In these studies, each model was associated with a label such as waving a hand, bowing and so on. However, since the human-like behavior in a daily life is diverse and ambiguous, defining a clear label for each motion of such behavior is difficult. To treat with this problem, in this paper, we collected the human motion data during a dialogue and propose a modeling method using Generative Adversarial Networks (GAN) [15]. The result of the human-robot experiment based on the subjective evaluation of the participants suggested that the human-like behavior of the robot could be generated by the proposed method.

Key Words: Human Robot Interaction, Human Motion Modeling, Generative Adversarial Networks, Human Behavior During Dialogue

1. はじめに

近年、日常生活をサポートするロボットの開発が盛んに行われている [1] [2]. このようなロボットが人間に受け入れられ、人間と共生するためには、音声対話システムなどの言語コミュニケーションだけでなく、ジェスチャーなどの非言語コミュニケーションの改善も必要である [3] [4]. しかし、多様な状況に直面する日常的な場面で、人間とのコミュニケーションを行うロボットに対して、適切な動作を生成するのは容易ではない。

人間は人間と円滑にコミュニケーションを行うことができる。そこで、人間の振る舞いをモデル化し、ロボットへ実装することを考える。特にアンドロイドなどの人間に酷似したロボットの場合、人間らしい動きが親和性に影響を与えることが示唆されている [5]. しかしながら、人間の動作は対話相手や周囲の状況に応じて変わるため、多様な動作を生成可能なモデル化が必要になる。さらに、人間は同じ状況に直面しても同じ動作を行うとは限らないため、一つの状況に対して正解動作が一意に決まるのではなく、一つの状況に対して異なる複数の正解動作が存在する。よって、対話中の動作は正解のうちの一つであること、同じ動作の繰り返しによる飽きを防ぐことが重要になると考えられる。

現在まで、モーションプリミティブ [6] [7] や隠れマルコフモデル (HMM) [8] [9] などの数理モデルを用いた人間の動作のモデル化が盛んに行われてきた。しかし、これらの手法では与え

られた一つの典型的な動作をモデル化することを基本としている。そのため、様々な状況で多様な動作を扱うためには、動作の分類を行う上位階層のモデル化を行うなどの工夫が必要である。例えば、HMM を用いた動作モデルでは、前もってラベルの与えられた動作をモデル化した HMM を複数用意し、どのラベルの動作を生成するかをルール化することで動作生成が行える [10] [11]. しかし、日常で行う多様な動作に対するラベル付けを行うことを考えると、必要なラベル数が本質的に決められない。また、限定されたラベルを付与する場合でも、人手による作業が必要であり、コストが高い。よって、ラベル付けを行うことなく、ライフログ [12] のように単純に蓄積したデータから動作を生成できる手法が求められる。

生成モデルは観測データの生成する確率分布を用いてモデル化する手法であり、乱数を入力として多様な疑似的サンプルを生成する枠組みとして注目を集めている [13]. それらの中でも、Variational AutoEncoder [14] や Generative Adversarial Networks (GAN) [15] などのニューラルネットワークを用いた生成モデルは高次元で複雑な構造を持つ画像や文章の生成にも成果を上げている [16] [17]. 本研究では、コミュニケーションを行うロボットの動作 (ジェスチャー) の生成を目的として、対話中の動作に対する生成モデルの定式化を行う。

2. 対話中の動作とその性質

対話中の人間は傾きやジェスチャーなど多様な動作を行っている [18]. しかしながら、個々の動作に明確な意味付けができるとは限らないため、「お辞儀」といった明確なラベルをすべての動作に与えることはできない。さらに、同じ「傾き」であっても動作の大きさや回数が違うように、同じ状況において

原稿受付 2018 年 11 月 1 日

*大阪大学

*Osaka University

■ 本論文は有用性 (要素分野) で評価されました。

Fig. 1 Human behavior during dialog

Fig. 2 Generation of interaction motion data

適切な動作が複数あり得るため、多様な動作を生成する枠組みが必要である。

これに対応する方法としては、状況や動作に応じたクラスを選択と与えられたクラスに対する典型的動作の生成を行うという方法が考えられる [8]。しかし、クラス数が不明で、その境界も曖昧である対話中に起こる動作に、そのまま適用することは困難である。

すなわち、模倣学習 [22] (Imitation learning) など与えられた軌道を再現することはまったく異なる軌道の動作が同じような意味を持つ場合がある対話中の動作生成法としては必ずしも適切ではない。本研究では、ランダムに生成した潜在変数から多様で多峰性のあるデータを生成できる GAN を用いたインタラクション動作のモデル化に取り組む。これによって、明示的にクラスを選択を行うことなく、多様な動作の生成が期待できる。

2.1 対話中の人間の動作の計測

本研究では、Fig. 1 のようにカメラを用いて、対話中の 2 人の人間の動きを計測し、この動作のモデル化に取り組む。記録した画像に対して OpenPose [19] を使用し、頭部や左右の肩などの 11 個の特徴点の二次元座標 (22 次元ベクトルで表現される人の姿勢の状態を示す変数である。以後スケルトンとよぶ。) を抽出する (Fig. 2)。

一つのスケルトンの時系列、すなわち、上述の 22 次元ベクトルをフレーム数結合した変数を以後“動作データ”とよぶ。また、対話中の 2 人の動作データを組み合わせたものを“インタラクション動作データ”とよぶ。インタラクション動作データはフレーム数 T の二つのスケルトン (44 次元) の時系列 ($\mathcal{R}^{T \times 44}$) で表される。以後、計測データから構築した、インタラクション動作データを $\mathbf{X}^{\mathcal{R}}$ と書く。本研究では、このインタラクション動作データのモデル化に取り組む。

一つのインタラクション動作データは Fig. 3 の A に示されるように、高次元空間の中の低次元の多様体中の一点に対応する。インタラクション中の人間は同じようなシチュエーションでも異なる動作をする場合がある。例えば「挨拶」を行う場合、Fig. 3 の B1 や B3 のように右手を挙げる場合と左手を挙げる

Fig. 3 Multimodality of interaction motion data

(a) Data collection environment

(b) An image taken by omnidirectional camera

Fig. 4 A scene of research experiment

場合がある。すなわち、Fig. 3 の A のように、インタラクション動作データの確率分布 (Interaction space) は多峰性を持つと考えられる。

このような複雑な分布から、適切なサンプルを抽出することは容易ではない。例えば、B1 や B3 のように適切なサンプルの単純な平均 B2 は、必ずしもインタラクション動作の分布する多様体に含まれるとは限らない。本研究では、このような確率分布をモデル化することで、複数の典型動作の生成が可能なモデルを獲得することが目的である。

2.2 インタラクション動作データに対する主成分分析

2 名の被験者が対話を行っている様子を撮影することで得たインタラクション動作データを主成分分析を用いて調べた。Fig. 4(a) と Fig. 4(b) にデータ収集の環境とカメラによる撮影画像の例を示す。Fig. 4(a) のように、被験者はテーブルを挟んで向かい合った状態で座り、実験者から与えられた話題について対話を行うように指示される。その様子をテーブルの中央に設置した全方位カメラから撮影した (Fig. 4(b))。なお、対話

Fig. 5 Dimensionality of latent space of skeleton

ロボットの動作生成への応用を考慮して、ロボットの一人称視点に近い動画像を得るためにこのような配置とした。計3組の対話を24 [fps] で撮影し、合計約1時間分の動画を得た。

得られた動画から OpenPose [19] によりスケルトンを抽出し、インタラクション動作データを得た。全方位カメラから得られる画像のサイズは $736 \times 1,472$ [pixels] である。個人の画像に分けてスケルトン抽出を行うために、カメラから撮影した画像を 736×736 [pixels] のサイズの画像に分割し、各画像に対して OpenPose の処理を施した。この OpenPose によって得られる二次元座標（各軸の値: [0 736]）を最大値（736）で割ることで正規化（各軸の値: [0,1]）する。また、ノイズを除去するために、i) OpenPose の出力する推定信頼度が5%を下回る外れ値の除去し、ii) カットオフ周波数4 [Hz] のローパスフィルタを適用し平滑化を行った。その後、iii) 線形補間を用いて8 [fps] でデータをリサンプリングを行った。

Fig. 5 に姿勢データ（1時刻分のスケルトン）に対する主成分の累積寄与率を示す。横軸は主成分数、縦軸は累積寄与率を示す。実線は1人分のスケルトン（22次元）、一点鎖線は2人分のスケルトン（44次元）に対する累積寄与率である。累積寄与率が99%を超える最小の主成分数を圧縮次元とよぶと、1人の場合は圧縮次元が14であるのに対して、2人の場合は24であった。

動作データの潜在的な次元は元のスケルトンの持つ次元（ $2 \times 14 = 28$ 次元）よりも小さい空間（24次元）を持つ。この結果は、インタラクション動作データの持つ潜在的な次元は、対話中の2人の動きが独立ではなく、双方が他方の動作に影響を与えていることを示唆している。また、前後の時刻で大きくスケルトンが変化する（独立性が高い）とは考え難いため、比較的低次元の潜在変数を持つ生成モデルを用いることができると考えられる。

3. インタラクション動作の生成モデル

生成モデルとは、観測データが何らかの確率的な過程に従って生成されていると仮定し、観測データが生成される確率分布を推定する枠組みである。生成モデルを用いてインタラクション動作データをモデル化することで、複数の正解動作があり得る、すなわち、多峰性のある動作データの生成が可能になると期待できる。

本研究では、GANを用いたインタラクションデータのモデル化に取り組む。GANはDiscriminatorとGeneratorとよばれる二つのニューラルネットワークによって構成される生成モデル学習の枠組みである。Generatorは潜在変数としてランダム

に生成した入力からリアルサンプル $\mathbf{X}^R \in \mathcal{R}^{T \times 44}$ （2.1節で示した計測データから生成したインタラクション動作データ）と同次元のベクトルでされる“フェイク”サンプル $\mathbf{X}^F \in \mathcal{R}^{T \times 44}$ を生成するネットワークである。

Discriminatorは入力されたデータ（ベクトル）が実際に観測されたリアルサンプルであるか、Generatorの出力したフェイクサンプルであるかを判別するネットワークである。入力データを $\mathbf{x}$ と書くと、その出力は

$$y = D(\mathbf{x})$$

のように表され、この出力 y は入力 $\mathbf{x}$ のリアルサンプルらしさに関する指標である。DiscriminatorはGeneratorの生成したフェイクサンプルを正しく判別できるように、GeneratorはDiscriminatorが誤判別するように、敵対的に学習を行う。この学習の結果、Generatorはリアルサンプルとの判別が困難なフェイクサンプルを出力するように学習する。

本研究では安定した学習を実現する学習アルゴリズムとして WGAN-GP (Wasserstein GAN - Gradient Penalty) [21] を利用した。WGAN-GPにおいては、Discriminatorの出力は正誤の2値ではなく、リアルサンプルらしさ（評価値）を出力する。Discriminatorの学習においてはリアルサンプルに対する評価値の期待値とフェイクサンプルに対する評価値の期待値の差を最大化するように学習を行う。

Discriminatorの学習における目的関数は

$$L_D = L_{WGAN} + \alpha L_{GP} + \beta \{ \gamma L_1 + (1 - \gamma) L_2 \}$$

となる。 $L_{WGAN} = E[D(\mathbf{x} = \mathbf{X}^F)] - E[D(\mathbf{x} = \mathbf{X}^R)]$ は評価値の期待値の差である。また、 L_{GP} はDiscriminatorのランドスケープの滑らかさに関するペナルティ項である[21]。また、 L_1 と L_2 はそれぞれDiscriminatorの結合重みパラメータ $\mathbf{w}_a$ に対する l_1 および l_2 ノルムであり、結合重みに対する正則化項である。 α , β と γ はハイパーパラメータであり、10.0, 0.0001 と 0.7 に設定した。最適化手法はAdamを用いた。

Generatorは標準正規分布に従う潜在変数ベクトル $\mathbf{z} \in \mathcal{R}^m$ を入力として、フェイクサンプル $\mathbf{X}^F \in \mathcal{R}^{T \times 44}$ を出力する。GeneratorはこのフェイクサンプルをDiscriminatorへ入力し、その出力の期待値を最小化するように重みを更新する。Generatorの損失関数は

$$L_G = -E[D(\mathbf{x} = \mathbf{X}^F)]$$

のように計算される。最適化手法はAdamを用いた。

4. 生成された動作の評価

本研究では、インタラクション動作データとして、4秒間の動作を8 [fps] で計測したデータ（ $T=32$ ）を扱い、これを1個のリアルサンプルとする。すなわち、インタラクション動作データの各サンプルは1,408次元のベクトルとなる。本研究では、合計約1時間分の動画を収録し、特徴点を抽出することで、合計44,608個のリアルサンプルからなるデータセットを学習データとして用いた。

実験に用いたDiscriminatorとGeneratorのネットワーク構造をTable 1に示す。それぞれ5層のニューラルネットワークであり、エンコーダとデコーダのように対称な構造を持つように設計した。また、Fig. 5より、1フレームの情報に24次

Table 1 Network architecture

Discriminator	Generator
Input $\mathbf{x}$	Input $\mathbf{z} \sim \mathcal{N}^{(30)}(0, I)$
$\mathbf{h}_1 = \text{Full}(1408, 704, \text{LR})(\mathbf{x})$	$\mathbf{z}_1 = \text{Full}(30, 88, \text{LR})(\mathbf{z})$
$\mathbf{h}_2 = \text{Full}(704, 352, \text{LR})(\mathbf{h}_1)$	$\mathbf{z}_2 = \text{Full}(88, 176, \text{LR})(\mathbf{z}_1)$
$\mathbf{h}_3 = \text{Full}(352, 176, \text{LR})(\mathbf{h}_2)$	$\mathbf{z}_3 = \text{Full}(176, 352, \text{LR})(\mathbf{z}_2)$
$\mathbf{h}_4 = \text{Full}(176, 88, \text{LR})(\mathbf{h}_3)$	$\mathbf{z}_4 = \text{Full}(352, 704, \text{LR})(\mathbf{z}_3)$
$\mathbf{y} = \text{Full}(88, 1, \text{None})(\mathbf{h}_4)$	$\mathbf{X}^F = \text{Full}(704, 1408, \tanh)(\mathbf{z}_4)$
Full(input size, output size, activation) is full connection layer.	
“LR” : LeakyReLU	
“None” : Identity function	
“tanh” : Hyperbolic tangent

Fig. 6 Training curves

元程度必要であることから、それに数次元を増加し、経験的に Generator の入力（潜在変数）の次元は 30 とした。

既存の研究として、LSTM (Long Short Term Memory) や GRU (Gated Recurrent Unit) などの RNN (Recurrent Neural Network) を用いて時系列データを扱い、人間の動作予測 [23] [24] や文章生成 [29] などを行う研究がなされている。RNN の再帰的な構造を用いることで、学習パラメータの数の削減が期待できる。しかし、RNN は IIR (Infinite Impulse Response) システムであるため、出力が初期状態によって変化する。対話自体は長いコンテキストを持つと考えられるが、本研究では、動作自体は直近の振る舞いにしか影響を受けないものと考え、動作の時間を 4 秒間とした。ロボットの一つの動作プリミティブの長さは 2 秒程度のものが多く [8] [22]、4 秒間はそのバイグラム程度の長さになっていると考えられる。よって、本研究では、全結合層による単純なフィードフォワードネットワークモデルを利用した。近年、このような時系列モデルを扱うフィードフォワードネットワークとして CNN (Convolutional Neural Network) を自己回帰モデルとして扱う研究が盛んであり [25] [26]、その応用により高精度化が期待できる。

4.1 生成モデルの学習曲線

Fig. 6 に Discriminator と Generator の学習曲線を示す。横軸はエポック数、縦軸は各ネットワークの損失関数の値である。すべてのリアルサンプルを 1 度だけ使用して学習することを 1 エポック (Epoch) とよぶ。バッチサイズは 64 とした。リアルサンプル数が 44,608 であったため、1 エポックにつき 697 回重みが更新される。本報告では、40,000 エポックで学習を停止した。

4.2 生成された動作の評価

本報告では、動作生成の手法として、線形確率的主成分分析 (PPCA) を用いた方法との比較を行った。PPCA では、最尤推定法を用いて主成分を推定した。潜在変数の次元数は累積寄与率 99% を基準として決め、315 次元の潜在変数を持つモデルとなった。

(a) PPCA

(b) GAN

Fig. 7 Generated motions

Fig. 7 (a) と **Fig. 7 (b)** にそれぞれ PPCA と GAN の各条件によって生成した動作の例を示す。なお、学習後の Generator が生成したインタラクション動作データは微細な高周波の振動を含んでいたため、GAN 条件の動作に対してはカットオフ周波数 2 [Hz] のローパスフィルタを適用している。Fig. 7 (a) のように、PPCA 条件では 2 名が常時かつ同時に動くような動作が多く生成される。それに対して、Fig. 7 (b) のように、一方が動作をする (Active) ときはもう一方は動作をしなくなる (Inactive) といった、GAN からは会話のターンテイキングのように見える動作も生成された。

また、本研究では、計測したデータ (Real), GAN, PPCA

Table 2 Questionnaire results

	Real	GAN	PPCA
相手の話を聞いている	98	122	37
オープンアティチュード	59	55	161
頭部を触る	39	21	6
不自然な動作	14	20	5
腕を組む	12	4	11
ガッツポーズ	11	10	11
その他	0	0	0
総回答数	233	232	231
エントロピー	1.33	1.12	0.93

Fig. 8 Comparison of percentage of motion labels

の各条件から生成したインタラクション動作データに対して、動作に対するラベル付けの多様さと容易さをアンケートにより評価した。なお、各条件からは2人分の動作データが生成されるが、負担を軽減するために被験者は1人分の動作だけを視聴して、評価した。

主観評価は以下の手順で行った。

- (1) 各条件から5個（計15個）のインタラクション動作データを生成する。
- (2) 生成された15個のインタラクション動作データ、それぞれを1人分の動作データに分割し、合計30個の動作データを作る。
- (3) 上記の動作データからランダムに一つの動作データを選択し、そのスケルトン動画を被験者に提示する。
- (4) 被験者は提示された動画に対するラベル付けとその確信度を入力する。
- (5) 30個すべての動画を評価し終わるまで、(3)と(4)の作業を繰り返す。

上記の手順を1セッションとして、被験者は計4セッション分の評価を行う。最初の1セッションは練習として、残りの3セッションを集計対象とした。

アンケートの集計結果をTable 2に、その棒グラフをFig. 8に示す。被験者は計8名であり、各条件の回答数は計240個である。一つの動画に対して複数のラベルを付けるなどの例外が生じた場合はデータを棄却した。その結果、Real条件からは7個、GAN条件からは8個、PPCA条件からは9個のデータを棄却した。

Fig. 8より、PPCA条件はオープンアティチュードの割合が高く、GAN条件はRealの比率に近い結果となった。また、Table 2の最下行に各条件でラベルの頻度に関するエントロピーを示す。PPCA条件よりもRealとGANの条件の方がエントロピーが高い。これは、GANを用いた提案手法によってより

(a) Horizontal

(b) Vertical

Fig. 9 Distribution of wrist position

多様な動作が生成されていることを示唆している。

また、各条件のインタラクション動作にターンテイクングが含まれていることを確かめるために、2名それぞれの動作を被験者へ提示し、「話し手」、「聞き手」、「分からない」の三つのアノテーションを行ってもらった。アノテーションはペアとしてインタラクション動作となる動画の同時刻の振る舞いについて行ったため、同時刻の2人の動作に対して、それぞれ「話し手」、「聞き手」、「分からない」のいずれかにラベルが付与される。2名の動作に対して、通常、対話中に多くは含まれないと考えられる「話し手、話し手」というラベルの組み合わせが含まれている割合を調べると、Real条件、GAN条件とPPCA条件はそれぞれ22%、11%、71%となった。また、「話し手、聞き手」というラベルの組み合わせが含まれている割合を調べると、Real条件、GAN条件とPPCA条件はそれぞれ53%、46%、24%となった。結果として、PPCA条件は同時に「2名が話し手」と認識されるような動作を最も多く含んでおり、一方で、計測データ（Real）とGAN条件は「一方が話し手で、もう一方が聞き手である」と認識されるような状況が半分程含まれている。このことから、GANを用いた提案手法によって実際の対話中の動作（Real）に含まれる、ターンテイクングを考慮した対話中の動作を生成していることが示唆される。

また、リアルサンプルと各手法によって生成されるフェイクサンプルの位置と速度の分布を比較した。分布の例として、手先の位置の分布をFig. 9に、手先の速度の分布をFig. 10に示す。それぞれ、低頻度のデータの可視化のために二重線の上下で縮尺を変えて図示している。各グラフの横軸は元の画像サイズ（736 × 736 [pixels]）における位置と速度の値、縦軸はそれぞれの値の頻度を示す。各特徴点の位置分布は胴体の座標を原点として計算した。速度分布は時刻 t フレームめの特徴点座標

Fig. 10 Distribution of wrist velocity

Fig. 11 KL distance between real sample and fake sample

と $t-1$ フレームめの座標との差に 1 秒間のフレーム数 (8 [fps]) を掛けた数値の分布である。各分布の計算時に特徴点の対称性 (右手と左手や個人) を考慮して、データを統合した。

Fig. 11 にリアルサンプルの分布を基準とした GAN および PPCA, それぞれの手法を用いて生成されたフェイクサンプルの分布の間の Kullback-Leibler (KL) 距離を比較した結果を

Fig. 12 Distribution of speed of all feature points

示す。Fig. 11 (a) と Fig. 11 (b) より、位置と速度分布の観点から、PPCA 条件よりも GAN 条件の方が訓練データにより近い分布となっていることが分かる。Fig. 8 のように GAN 条件から生成された動作ラベルがリアルサンプルの比率に近い結果となったのは、生成動作の位置と速度の分布が近いことが要因となっていると考えられる。

また、特徴点の速さの平均値の分布を Fig. 12 に示す。横軸は速さの平均値、縦軸は頻度を示す。 k 番目の特徴点の水平方向と垂直方向の速度をそれぞれ H_k と V_k とすると、特徴点の総数は 11 点であるため、速さの平均値 $\bar{s}$ は

$$\bar{s} = \frac{1}{11} \sum_{k=1}^{11} \sqrt{H_k^2 + V_k^2}$$

のように計算される。Fig. 12 より、Real 条件はほかの条件よりも速さが小さい成分が多く含まれ、なおかつ裾が広い分布となっている。PPCA 条件はほかの条件よりも速さが大きい成分が多くなっており、GAN 条件は PPCA 条件と同じような速さの区間で頻度の値を取っているが、比較的速さが小さい成分が多く含まれている。これらのことから、PPCA 条件では、ゆっくりながら常に動き続ける動作が生成されやすいのに対して、GAN 条件は静止と動作を切り替えているような動作が生成されやすいことが分かる。

5. ま と め

本報告では、インタラクション動作のモデル化を生成モデルを用いて定式化した。GAN を用いた提案手法によって線形モデルである比較手法よりも多様な動作が生成されていることを確かめた。また、提案手法によって実際の人間の動きに近い動作が生成されていることが示唆された。

近年、小型のイメージセンサを利用できる三次元姿勢推定のシステム (RGB-D センサ [27], OpenPose 3D [28]) が盛んに開発されている。これらのシステムには、セットアップ等が簡便で多量のデータの取得が容易であること、マーカを取り付けるなどのシステムの制約が少ないこと、人型のロボット等の機器の搭載に物理的な制約がある場合にも利用しやすいなどのメリットが考えられる。CNN の利用などのネットワークの構成の最適化とともに、これらのシステムを用いた三次元動作モデルの構成は今後の課題である。

動作データ生成モデルとオンラインで取得した過去の動作データを条件として利用することで、ヒューマノイドの動作生成を行うことができる [20]。このためには、現在までの動作をネットワークに入力するもの [20] や潜在変数の効果を調査し、制御する方法 [16] の利用が考えられる。また、ヒューマノイドの動

作生成システムを実現するために、対面者の動作データや音声などの環境情報や対話のコンテキストを条件とした動作を生成するモデルにも取り組む必要がある。

参 考 文 献

- [1] H. Iwata and S. Sugano: "Design of human symbiotic robot twenty-one," ICRA, pp.580-586, 2009.
- [2] T. Mukai, et al.: "Tactile-based motion adjustment for the nursing-care assistant robot riba," ICRA, pp.5435-5441, 2011.
- [3] A. Mehrabian: Silent messages: Implicit communications of emotions and attitudes. Belmont, CA: Wadsworth, 1981.
- [4] R.L. Birdwhistell: Kinesics and context: Essays on body motion communication. University of Pennsylvania press, 2010.
- [5] 森政弘, 他: "不気味の谷", Energy, vol.7, no.4, pp.33-35, 1970.
- [6] Ijspeert, et al.: "Movement imitation with nonlinear dynamical systems in humanoid robots," IEEE International Conference on Robotics and Automation, vol.2, pp.1398-1403, 2002.
- [7] 中西淳, 他: "運動学習プリミティブを用いたロボットの見まね学習", 日本ロボット学会誌, vol.22, no.2, pp.165-170, 2004.
- [8] 稲邑哲也, 他: "ミメシス理論に基づく見まね学習とシンボル創発の統合モデル", 日本ロボット学会誌, vol.22, no.2, pp.256-263, 2004.
- [9] 高野渉, 他: "身体運動と言語を結ぶロボットの統計的情報処理", 日本ロボット学会誌, vol.30, no.7, pp.674-683, 2012.
- [10] 稲邑哲也, 他: "幾何学的シンボル操作による多様な動作パターンの認識・生成を実現する原始シンボル空間の構成法", 日本ロボット学会誌, vol.27, no.5, pp.564-574, 2009.
- [11] T. Wataru and N. Yoshihiko: "Real-time unsupervised segmentation of human whole-body motion and its application to humanoid robot acquisition of motion symbols," Robotics and Autonomous Systems, vol.75, pp.260-272, 2016.
- [12] L. Bao, et al.: "Activity recognition from user-annotated acceleration data," International Conference on Pervasive Computing, pp.1-17, 2004.
- [13] I.J. Goodfellow, et al.: Deep Learning. MIT press Cambridge, 2016.
- [14] D.P. Kingma and M. Welling: "Auto-Encoding Variational Bayes," ICLR 2014, 2014.
- [15] I.J. Goodfellow, et al.: "Generative Adversarial Networks," NIPS 2014, 2014.
- [16] A. Radford, et al.: "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," ICLR 2016, 2016.
- [17] Z. Hu, et al.: "Toward Controlled Generation of Text," PMLR 2017, 2017.
- [18] T. Nishida: Conversational Informatics: An Engineering Approach (Wiley Series in Agent Technology). Wiley Online Library, 2007.
- [19] S-En Wei, et al.: Convolutional pose machines. CVPR, 2016.
- [20] 西村優佑, 中村泰, 石黒浩: "敵対的生成ネットワークを用いたインタラクション中の人間の振る舞いのモデル化", 研究報告数理モデル化と問題解決 (MPS), vol.2018-MPS-117, no.6, pp.1-6, 2018.
- [21] I. Gulrajani, et al.: "improved training of wasserstein Gans," NIPS 2017, 2017.
- [22] M. Takamitsu, et al.: "Learning Stylistic Dynamic Movement Primitives from multiple demonstrations," IROS 2010, 2010.
- [23] K. Fragkiadaki, et al.: "Recurrent network models for human dynamics," ICCV 2015, 2015.
- [24] J. Martinez, et al.: "On human motion prediction using recurrent neural networks," CVPR 2017, 2017.
- [25] C. Donahue, et al.: "Adversarial Audio Synthesis," ICLR 2019, 2019.
- [26] A. Oord, et al.: "Wavenet: A generative model for raw audio," arXiv:1609.03499, 2016.
- [27] L. Keselman, et al.: "Intel realsense stereoscopic depth cameras," CVPR 2017, 2017.
- [28] H. Joo, et al.: "Total capture: A 3d deformation model for tracking faces, hands, and bodies," CVPR 2018, 2018.
- [29] Y. Lantao, et al.: "Seqgan: Sequence generative adversarial nets with policy gradient," AAAI 2017, 2017.

西村優佑 (Yusuke Nishimura)

2018 年大阪大学基礎工学部システム化学科卒業。2015 年同大学大学院基礎工学研究科システム創成専攻入学, 現在在学中。対話ロボットの動作生成に関する研究に従事。(日本ロボット学会学生会員)

石黒 浩 (Hiroshi Ishiguro)

1991 年大阪大学大学院基礎工学研究科博士課程修了。工学博士。その後、京都大学情報学研究科助教授、大阪大学工学研究科教授等を経て、2009 年より大阪大学基礎工学研究科教授。ATR 石黒浩特別研究所客員所長 (ATR フェロー)。2017 年から大阪大学名誉教授。専門は、ロボット学、アンドロイドサイエンス、センサネットワーク等。2011 年大阪文化賞受賞, 2015 年文部科学大臣表彰受賞。(日本ロボット学会正会員)

中村 泰 (Yutaka Nakamura)

2004 年奈良先端科学技術大学院大学博士課程修了。工学博士。2006 年大阪大学大学院工学研究科助手, 2010 年同大学基礎工学研究科助教, 2013 年大阪大学未来戦略機構特任准教授を経て、2018 年より大阪大学基礎工学研究科特任准教授。知能ロボット、機械学習に興味を持つ。