〈一般研究課題〉　人形浄瑠璃所作と拡散モデルを
用いたインタラクション動作生成

助 成 研 究 者　中京大学　　董　然

人形浄瑠璃所作と拡散モデルを用いたインタラクション動作生成

董　然
(中京大学)

Interaction Motion Generation using Puppet Theater Movements and Diffusion Models

Ran Dong
(Chukyo University)

Abstract：

Interactions between humans and virtual characters or humanoid robots often suffer from the "uncanny valley" phenomenon, wherein near-human movements nonetheless appear subtly unnatural. To address this issue, we draw inspiration from Bunraku— Japan's traditional puppet theater— wherein three puppeteers imbue expressionless puppets with emotional depth through nonverbal cues: the preparatory "Hodo," the initiating "Zu," and the rhythmic structure "Jo-Ha-Kyu." This study investigates the nonverbal coordination mechanisms employed in Bunraku, focusing on how the puppeteers achieve highly synchronized actions through "Hodo" and "Zu." We first captured full-body puppet motions using a combination of optical and magnetic motion capture systems. The recorded data were subsequently analyzed in the frequency domain via the Hilbert–Huang Transform (HHT) to extract the "Hodo" and "Zu" signals. Following this, we employed a Human Motion Diffusion Model by masking the left-hand channels and conditioning the model on the remaining joints, enabling the generation of plausible left-hand trajectories that align with the overall movement context. Finally, we demonstrated the transferability of the extracted motion data to a humanoid robot by mapping the captured motion onto the robot's kinematic structure, with further refinement achieved through HHT-based optimization to preserve the temporal and expressive characteristics of the original performance.

1. はじめに

近年、生成AIをはじめとする人工知能技術の進展に伴い、CGキャラクターや人型ロボットを介した人間との自然なインタラクションが求められている。しかしながら、ロボットやアバターの外見や動作が人間に近づくにつれ、微細な違和感が「不気味の谷」現象を引き起こし、ユーザーに拒絶感や不快感を与える問題が指摘されている。こうした中、ユネスコ無形文化遺産にも登録されている日本の伝統芸能「人形浄瑠璃(文楽)」においては、表情を持たない木製人形でありながら、義太夫の語りや三味線の音楽、三人の人形遣いの協調動作によって、豊かな感情表現が実現されている点に注目が集まっている。特に、主遣いが発する非言語的な合図「ホド」と「ズ」、および「序破急」に代表されるリズムの緩急に基づく即興的な動作制御メカニズムは、感情的かつ自然なインタラクションの鍵を握ると考えられる[1, 2]。

一方で、これまでの研究は文楽の動作やリズムの文化的・芸術的分析に留まり、工学的観点から「ホド」や「ズ」の非言語的合図の機能解明や、「序破急」メカニズムに基づく動作生成への応用については十分に探究されてこなかった。AIによるモーション生成の分野においても、既存の変分オートエンコーダーや拡散モデルによる潜在空間の正規分布へのマッピングは、生成される動作の即興性や多様性の限界という課題を抱えており、人間と自然に交流できる動作をデータ駆動的に生成するための新たな手法が求められている。

このような背景を踏まえ、本研究では文楽における「ホド」と「ズ」という非言語的合図と「序破急」のリズム構造に着目し、これらのメカニズムを拡散モデルに学習させることにより、CGキャラクターやロボットが人間と自然で感情豊かなインタラクションを可能にする動作を自律的に生成できるAIモデルの構築を目指した。文楽人形の所作を加速度センサーおよびモーションキャプチャにより収集し、義太夫や三味線の音声データと対応付けたマルチモーダルデータセットを構築した上で、ヒルベルト・ファン変換や多変量経験的モード分解を用いて「ホド」と「ズ」の非線形周波数特性を抽出し、解析を行う。その結果を踏まえて、Motion Diffusion Model (MDM) [3]に協調動作の自動生成ができるような学習フレームワークの開発を試みた。

本研究の目的は、伝統芸能に内在する非言語的コミュニケーションの工学的理解を深めるとともに、即興性と感情豊かさを兼ね備えた動作生成AIの実現を通じて、CGキャラクターやロボットが人間との自然なインタラクションを可能にする基盤を提供することである。

2.「ホド」と「ズ」を用いた人形浄瑠璃の非言語的協調コミュニケーション

2.1 人形浄瑠璃における「ホド」と「ズ」

人形浄瑠璃(文楽)は、義太夫、三味線、人形遣いの「三業」が一体となり、感情豊かな物語を無表情の人形に宿す日本独自の伝統芸能である。この芸能において、一体の人形を三人の人形遣いが操作する「三人遣い」の技法は、世界的にも稀有であり、彼らがどのように動作を協調させているのかが注目される。三人のうち、頭と右手を操作する「主遣い」、左手を操作する「左遣い」、そして脚を操作する「足遣い」が、それぞれの役割を持ちながら一体の動作を作り上げるために、非言語的な合図を介して同期を取っている[4]。

この非言語的合図として最も重要なものが「ホド」と「ズ」である。「ホド」は、主遣いが人形を上下左右に微妙に揺らすことで左遣いや足遣いに「次に動作が来る」ことを知らせる信号であり、動

作の準備や方向性を示す役割を担う。一方、「ズ」は、主遣いが首や体幹に一瞬力を入れる、もしくはリズムに合わせて瞬間的に動作を加えることで、「今ここで動作を始める」という開始のタイミングや動作の強調を示す合図である。これらは、義太夫の語りや三味線のテンポに即した即興的な信号であり、左遣いや足遣いはこの「ホド」「ズ」に反応して動作を開始・調整することで、あたかも一人の人間が動いているかのような自然で連続的な動作を生み出している。

図1に示されるように、主遣い、左遣い、足遣いの三人が人形を操作している様子では、「ホド」と「ズ」の合図が不可欠な同期の鍵として機能している。特に、図1に示されるスペクトログラム(音声の時間周波数解析)からも分かるように、「Kyu」(急)は物語や動作のクライマックスであり、テンポの急激な変化とともに高い動作強度が要求される。このような場面では、主遣いは音楽のリズムの変化に合わせて「ホド」と「ズ」をより頻繁かつ強調して発し、他の遣い手がそれに即応することで、感情的クライマックスを共有・実現していることがわかる。このテンポ変化と非言語的信号の関係性は、文楽特有の非言語的協調コミュニケーションの本質であり、人とAI、ロボット間の自然な協調インタラクションを実現するための重要な手がかりとなる。

図1 序破急を用いた人協調コミュニケーション

2.2 非言語的協調コミュニケーションの社会応用の可能性

このような人形浄瑠璃における「ホド」と「ズ」に基づく協調メカニズムは、現代のAI・ロボティクス技術において、人とAIの協調、あるいはAI同士の協調制御に応用可能な貴重な知見を提供する。図2に示されるように、伝統芸能における人形遣い・義太夫・三味線・観客の構造は、日常生活におけるロボット、ユーザー、AIアシスタント、入力デバイスという構造に置き換えることができる[5]。

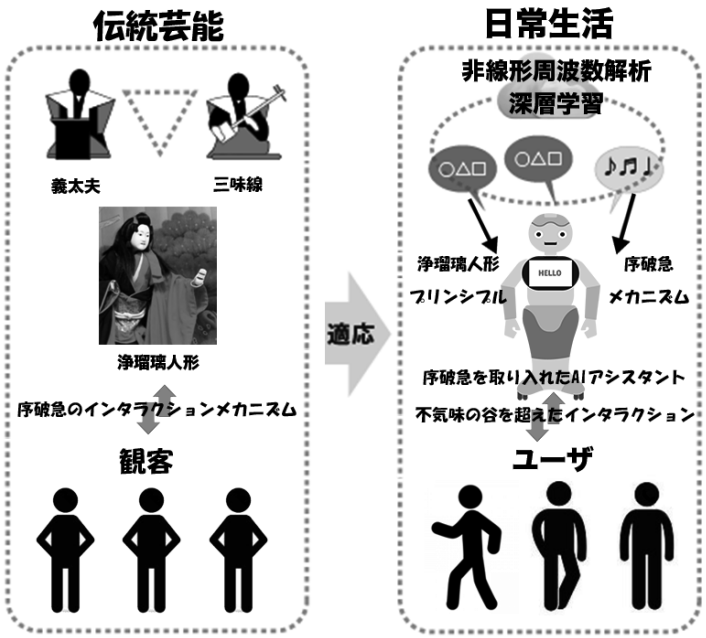

図2 人形浄瑠璃メカニズムの適用概念図

人形をロボット、義太夫と三味線をユーザーからの音声入力、そして人形遣いをAIのモーション制御モジュールに対応させることで、非言語的かつ即興的な指示に基づく動作生成が可能となる。これに

より、単にあらかじめプログラムされた動作を再生するのではなく、状況や環境、ユーザーの意図に応じた自然で感情豊かな応答をリアルタイムに実現できるロボット・CGキャラクターの実装が期待される。

さらに、不気味の谷問題の克服においても有効性が期待される。現在のロボットやCGアバターでは、外見のリアルさに比して動作のタイミングやニュアンスに乏しいことが、違和感の一因とされる。文楽のように「ホド」と「ズ」に基づく即興的・非言語的信号による協調動作を実現することで、動作の「間」や「リズム」の再現を可能とし、ユーザーが無意識に感じる違和感を軽減できる可能性がある。この非言語的協調メカニズムの導入は、教育、介護、観光、エンターテインメント分野でのインタラクティブなAIシステムの質的向上に寄与するものと考えられる。

2.3 拡散モデルを用いた協調動作のAI実装

本研究では、文楽における「ホド」と「ズ」に基づく非言語的協調メカニズムをAIに学習させるために、Motion Diffusion Model[3] を用いたモーション生成手法を提案する。Diffusion Modelは、データ分布から抽出されたサンプルに対して逐次的にノイズを付加し、学習時には逆拡散プロセスを介してそのノイズを除去することでデータの復元を学習する確率的生成モデルである。

このモデルにおいて、データ分布から得られる時系列モーションデータ $x_0^{1:N}$に対して、マルコフ過程としてノイズ付加プロセス ${x_t^{1:N}}_{t=1}^{T}$を次のように定義する：

$$q(x_t^{1:N}|x_{t-1}^{1:N}) = \mathcal{N}\left(\sqrt{\alpha_t}x_{t-1}^{1:N}, (1-\alpha_t)I\right) \tag{1}$$

ここで$\alpha_t \in (0,1)$は$\alpha_t = 1 - \beta_t$として時刻 t ごとに決まるパラメータである。逆拡散プロセスでは、ϵ_t を推定するのではなく、Motion Diffusion Model[3]に従い信号そのもの$\hat{x}_0$ を推定する形式を採用し、以下のシンプルな損失関数を用いて学習する：

$$\mathcal{L}_{\text{simple}} = E_{x_0 \sim q(x_0|c),\, t\sim[1,T]}\left[\|x_0 - G(x_t, t, c)\|_2^2\right] \tag{2}$$

ここで cは条件（本研究では「ホド」「ズ」等の合図を含む体幹・頭部・右腕のモーションデータ）、G は生成モデルを表す。本研究のアプローチでは、左腕の動作をマスク（欠損）し、c に基づき欠損部分（左腕の動作）をモデルが自動補完する問題設定とすることで、主遣いの非言語的合図に基づき左遣いの動作が決定される文楽特有の協調メカニズムをAIが学習・再現できるようにした。

このように、モデルには左腕以外の部位（体幹・頭部・右腕など）の時系列データを入力とし、出力として左腕の動作を生成させるタスクを定義する。左腕のマスクは$\mathcal{M}$として数式上次のように表される：$x_0^{\text{masked}} = \mathcal{M}(x_0)$。この$x_0^{\text{masked}}$ を入力として条件付き拡散モデルを学習することで、「ホド」と「ズ」に含まれる非言語的合図情報を利用し、左遣いの動作を推定可能とするフレームワークを構築した。

3. 実験結果

3.1 ヒルベルト・ファン変換を用いた「ホド」と「ズ」の抽出

本研究では、文楽における人形操作における非言語的協調コミュニケーションの中核をなす「ホド」と「ズ」の抽出に、ヒルベルト・ファン変換（Hilbert-Huang Transform, HHT）を用いた。文楽における人形の動きは非線形かつ非定常であり、従来の短時間フーリエ変換（STFT）や連続ウェーブ

レット変換(CWT)では、これらの微細な信号を適切に抽出できないという課題があった[6]。そのため、本研究では、図3が示すように、MEMD (Multivariate Empirical Mode Decomposition)により人形から採取した動作信号をIMF (Intrinsic Mode Function)に分解し、各IMFにヒルベルト変換を適用して瞬時周波数を算出するHHT手法を使用した。

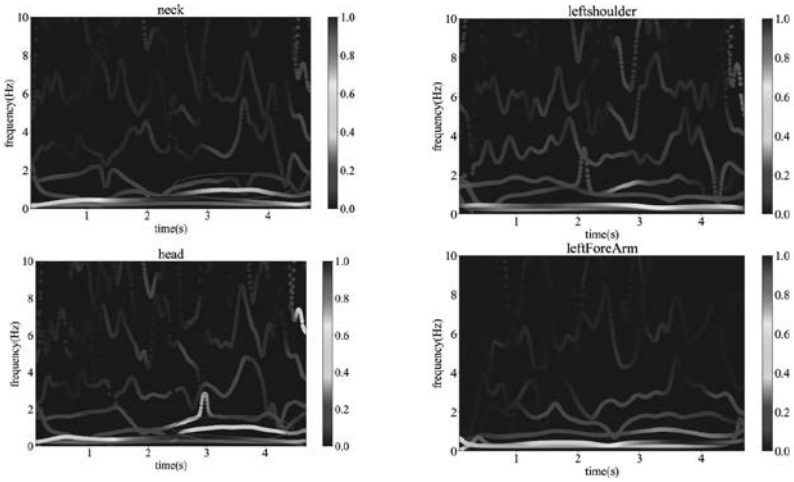

図3 「ホド」と「ズ」が行われた人形モーションデータ(代表的な型「手を合わせる」)に対してHHTによる関節ごとのスペクトル解析

実験では、文楽の代表的な型「手を合わせる」のシーンを対象に、人形の頭部、体幹、左右の腕にセンサーを配置し、モーションキャプチャデータを収集した。図4が示した解析の結果、主要な信号として「ホド」はIMF6、「ズ」はIMF8に対応することが明らかとなった。これにより、各IMFの瞬時周波数スペクトルおよび位相の推移から、主遣いが左遣いに対して「ホド」と「ズ」を送信するタイミングを特定できた。図4では、特に「ホド」から「ズ」への移行のタイミングが瞬時周波数において顕著に示されており、非言語的協調信号が一体的に動作内に埋め込まれている様子が確認できた。

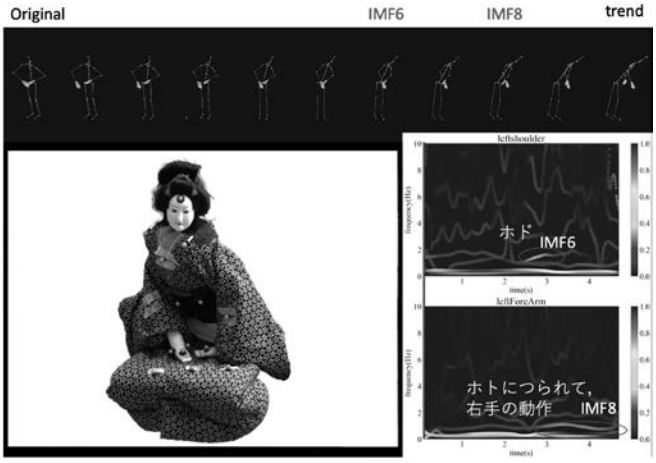

図4 「ホド」と「ズ」のHHTスペクトル解析の例(人形は着物バージョンに差し替え、実データとは異なる例示用画像)

この結果は、従来の定常信号解析では見落とされていた微細な協調信号の解明につながり、文楽の非言語的コミュニケーションの動態的理解に貢献するものである。

3.2 拡散モデルを用いた協調動作生成

本研究ではさらに、文楽人形の協調動作をAIにより自動生成する可能性を探るため、Motion Diffusion Modelを応用した。文楽人形のモーションデータにおいて、左手の動作をmaskし、他の部位(頭部、右手、体幹等)の動作情報のみを入力として、左手の動作を補完するタスクを設定した。

このタスクにおいては、MDMの定義式：$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$ に基づき、入力された部分動作情報から欠損部分を条件付き拡散モデルにより推論するアプローチを採用した。masking操作は、条件付き生成モデルの手法に基づき、特定関節の情報を欠損させたデータに対し、生成モデルがその部分を再構成するよう訓練を行った。実験結果では、Ground Truthの左手動作と比較し、生成された3つの試行データのいずれもが「手を合わせる」に至る協調動作を完全には再現できな

かった。図5に示すように、生成された左手動作は右手との接触点に十分近づくことができず、「ホド」と「ズ」の非言語的信号による誘導情報が、単なる動作データだけでは十分にモデル化できないことが示唆された。

この結果は、先行研究[7]が指摘したように、AIによる非言語的協調動作生成には、単なる関節座標や角度データに加え、時間的・リズム的文脈情報(Jo-Ha-Kyuのテンポ変化等)が不可欠であることを示しており、今後のモデル改善の指針となる知見を提供した。

3.3 人形データとHHTを用いたロボットへの最適化

最後に、文楽人形の協調動作データを用いて、ロボットにおける動作模倣の最適化を試みた。ここでは、MEMDをベースとするHHTを活用したロボット動作最適化手法を適用し、ロボットのモーター特性に適応した動作再構成を行った[8]。このアプローチにより、単にモーションキャプチャデータをロボットに実装した場合に生じるノイズや過負荷動作を抑制しつつ、人形特有の非線形な動作リズムを保持したまま、ロボットでの模倣を実現できた。図6では、MEMDによる分解結果およびヒルベルトスペクトル解析に基づくロボット動作の最適化過程を示している。

本研究の結果は、文楽人形の協調動作を模倣するロボット設計において、非線形性と周波数特性に着目した新たな実装手法として有効であり、将来的なヒューマノイドロボットの芸術表現やエンターテインメント応用への展開が期待される。

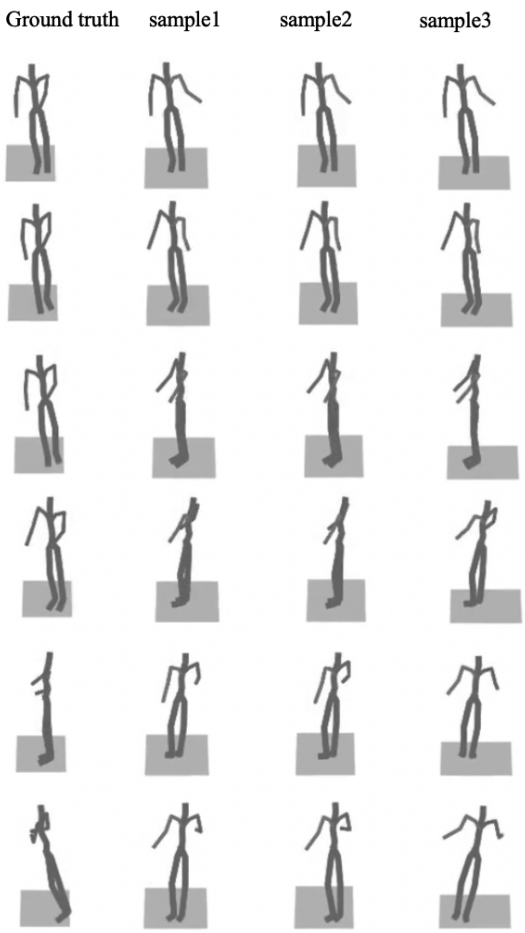

図5 MDMによる協調動作の自動生成

図6 HHTを用いた人形モーションのロボット実装(人形は着物バージョンに差し替え、実データとは異なる例示用画像)

4. まとめと今後の課題

本研究では、日本の伝統芸能である人形浄瑠璃(文楽)において、三人の人形遣いが非言語的合図「ホド」と「ズ」を用いて協調動作を実現しているメカニズムに着目し、この非言語的協調コミュニケーションの工学的解明およびAIによる動作生成への応用に取り組んだ。ヒルベルト・ファン変換を用いてモーションデータを周波数領域で解析し、「ホド」と「ズ」の信号抽出を行った。さらに、

Motion Diffusion Modelを用いて左遣いの動作を補完する協調動作生成の試みを行い、最後に抽出した動作データをロボット模倣動作の最適化に応用した。

今後の課題としては、拡散モデルによる協調動作生成の精度向上が挙げられる。本研究で用いたMotion Diffusion Modelにおいては、左手の動作をmaskし、他の部位の情報から補完するタスク設定を行ったが、生成された動作においては、右手との「手を合わせる」動作の接触タイミングや位置関係を正確に再現することができなかった。これは、拡散モデルの逆拡散過程において解の一意性が保証されないこと、すなわち逆拡散におけるランジュバン方程式(Langevin equation)の数値解法が唯一の解に収束しないことが一因であると考えられる。

この問題を解決するためには、逆拡散過程において追加情報を条件付ける必要がある。HHT解析によって抽出された「ホド」と「ズ」の周波数成分や位相情報をembedding情報として拡散モデルに入力することで、逆拡散における推定のガイドとし、解の一意性を高めることが有効であると考えられる。今後は、このembedding情報を統合した新たな条件付き拡散モデルの構築と評価を行い、非言語的協調動作のより自然かつ正確な生成を実現することが課題である。

参考文献

[1] 渋谷友紀, 森田ゆい, 福田玄明, 等. 文楽人形遣いにおける呼吸と動作の非同期的関係: 日本の古典芸能における「息づかい」の特殊性[J]. 認知科学, 2012, 19(3): 337-364.

[2] 丹波明.「序破急」という美学: 現代によみがえる日本音楽の思考型[J]. 2004.

[3] Tevet, Guy, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. "Human Motion Diffusion Model." *In The Eleventh International Conference on Learning Representations*.

[4] 渋谷友紀, 櫻哲郎, 佐々木正人, 等. 文楽における浄瑠璃と人形遣いの呼吸の同期[J]. 認知科学, 2017, 24(4): 518-539.

[5] Dong R, Chen Y, Cai D, et al. Robot motion design using bunraku emotional expressions focusing on Jo-Ha-Kyū in sounds and movements[J]. *Advanced Robotics*, 2020, 34(5): 299-312.

[6] 董然, 蔡東生. ヒルベルト-ファン変換を用いたダンスモーション解析[J]. 電子情報通信学会論文誌 D, 2019, 102(12): 843-853.

[7] Dong R, Cai D, Hayano S, et al. Investigating the Effect of Jo-Ha-Kyū on Music Tempos and Kinematics across Cultures: Animation Design for 3D Cha racters Using Japanese Bunraku Theater[J]. *Leonardo*, 2022, 55(5): 468-474.

[8] Dong R, Chang Q, Er M J, et al. Motion Capture-Based Robotic Imitation: A Keyframeless Implementation Method Using Multivariate Empirical Mode Decomposition[J]. *IEEE/ASME Transactions on Mechatronics*, 2024.