2P-05

弓動作を反映した弦楽器演奏モーションの自動生成

平田 明日香 † 田中 啓太郎 † 島村 僚 † 森島 繁生 ‡

† 早稲田大学 ‡ 早稲田大学理工学術院総合研究所

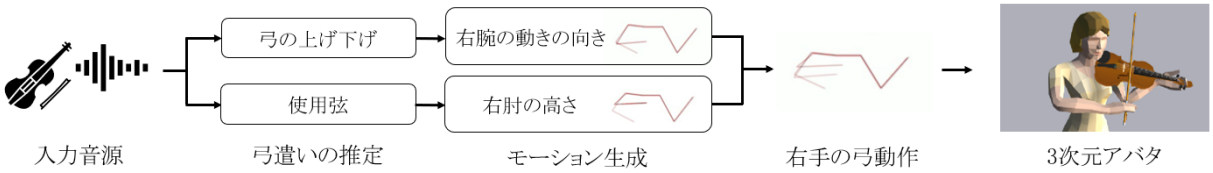

図 1: 提案手法概要図

1. はじめに

本稿では，弦楽器の演奏モーションの自動生成手法を扱う．演奏モーションの作成は演奏映像コンテンツ制作などの現場で広く求められているが，一般に，その作成にはモーションキャプチャやモーションの調整など多大なコストと労力を要する．そのため，演奏モーション生成の自動化が求められている．演奏モーションの自動生成に関する研究は広くなされており，その手法は 2 種類に大別される．一つは楽譜情報を入力とする手法，もう一つは音響情報を入力とする手法である．

楽譜情報を入力として演奏モーションを自動生成する手法では，ピアノ・ドラムの MIDI データから運指や演奏手順を自動で決定し，演奏モーションを生成している [1] [2]．鍵盤楽器や打楽器では楽器上のどの部分を弾いているかが楽譜情報から一意に定まる．しかし，弦楽器では異なる弦で同じ音を弾くことが可能であり，譜面上の同一音を弾く方法が複数考えられるため，楽譜情報のみから演奏手順を自動で決定することは困難である．

一方で近年，音響情報を入力とした深層学習による弦楽器演奏モーションの自動生成手法が提案されている [3] [4]．しかし，これらの手法では音響特徴量から直接モーションを生成しており，また既存の姿勢推定手法 [5] [6] を用いて推定されたポーズを正解としている．そのため，生成されるモーションの精度は正解ポーズの推定精度にも依存し，出力結果は音源に合わせて右手が充分に動いていない不自然なものとなる．演奏映像コンテンツの質を向上させるためには，モーションが演奏音源に対して音楽的に正しく自然なものである必要がある．特に弦楽器演奏において，右手の弓動作は弦楽器演奏の音色とモーションに強く影響する．したがって，自然な演奏モーションを生成するためには右手の弓動作をモーションに反映させる必要がある．

本研究では，弓遣いを構成する弓の上げ下げと使用弦を音響特徴量から推定し，その結果からルールベースで演奏モーション生成を行うことで，弓動作を反映したより自然なモーションを生成する．ヴァイオリン演奏音源から生成したモーションに対する評価実験によって本手法の有効性を確認する．

Automatic Generation of String Instrument Performance Motion Based on Bowing Mechanics, *Asuka Hirata†, Keitaro Tanaka†, Ryo Shimamura†, and Shigeo Morishima‡, (†Waseda University, ‡Waseda Research Institute for Science and Engineering)*

2. 提案手法

図 1 に提案手法の概要図を示す．まず，入力音源に対して弓遣いの推定を行い，その推定結果を基にルールベースでモーションの生成を行う．本手法により，右手の弓動作を反映した弦楽器演奏モーションを生成し，3 次元アバタに転写する．

2.1 弓遣いの推定

弓遣いの推定部分では，弓動作に特に強く影響する弓の上げ下げ，使用弦を推定した．入力音源の短時間フーリエ変換（Short Time Fourier Transform; STFT）によって得た周波数スペクトログラム $\boldsymbol{X}^{stft} = \{\boldsymbol{x}_1^{stft}, \ldots, \boldsymbol{x}_T^{stft}\} \in \mathbb{R}^{T\times F}$ を入力とし，各時刻における弓の上げ下げ $\hat{\boldsymbol{L}}^{ud} = \{\hat{l}_1^{ud}, \ldots, \hat{l}_T^{ud}\} \in \{0,1\}^T$（0 は下げ弓，1 は上げ弓）を得た．ここで，$T$ は時間フレーム数，F は STFT における周波数ビン数を表す．時系列を考慮するため，特徴抽出には長短期記憶（Long Short Term Memory; LSTM）を用いた．損失関数には binary cross entropy $\mathcal{L}_{bce}$ を用いた．

$$\mathcal{L}_{bce} = -\frac{1}{T}\sum_{t=1}^{T}(l_t^{ud}\cdot\log p_t^{ud} + (1-l_t^{ud})\cdot\log(1-p_t^{ud})) \quad (1)$$

ここで，l_t^{ud}，p_t^{ud} はそれぞれ弓の上げ下げの正解ラベルと推定確率である．さらに，$\theta = 0.5$ として次のように $\hat{\boldsymbol{L}}^{ud}$ を得た．

$$\hat{l}_t^{ud} = \begin{cases} 1 & (p_t^{ud} \geq \theta) \\ 0 & (p_t^{ud} < \theta) \end{cases} \quad (2)$$

使用弦についても同様に，入力音源の STFT によって得られた周波数スペクトログラム $\boldsymbol{X}^{stft}$ を入力とし，LSTM を用いて各時刻における使用弦 $\hat{\boldsymbol{L}}^{string} = \{\hat{l}_1^{string}, \ldots, \hat{l}_T^{string}\} \in \{0,1,2,3\}^T$ を得た．$\hat{l}_t^{string}$ は 4 本の弦のうち各時刻 t においてどの弦を使用しているかを表している．たとえばヴァイオリンの場合，0，1，2，3 はそれぞれ E 線，A 線，D 線，G 線に対応する．損失関数には cross entropy $\mathcal{L}_{ce}$ を用いた．

$$\mathcal{L}_{ce} = -\frac{1}{T}\sum_{c=0}^{3}\sum_{t=1}^{T} l_{t,c}^{string}\cdot\log p_{t,c}^{string} \quad (3)$$

ここで，$l_{t,c}^{string}$，$p_{t,c}^{string}$ はそれぞれ使用弦 $c \in \{0,1,2,3\}$ についての正解ラベルの one-hot 表現と推定確率である．さらに，使用弦の推定確率から次のように推定ラベル $\hat{\boldsymbol{L}}^{string}$ を得た．

$$\hat{l}_t^{string} = c,\ p_{t,c}^{string} = \max_{c'\in\{0,1,2,3\}} p_{t,c'}^{string} \quad (4)$$

表 1: 弓遣いの定量評価結果

	上げ下げ正解率	切り返し F 値	弦正解率
A2BD	0.560	0.461	-
TGM2B	0.493	0.422	-
提案手法	**0.704**	**0.566**	0.801

2.2 2 次元モーション生成

弓遣いの推定結果を基に，ルールベースで 2 次元のモーション生成を行った．まず，使用弦の推定結果から右肘の高さを決定した．使用弦が奥の弦であれば肘の高さは高く，手前の弦であれば肘の高さが低くなるように決定した．次に，弓の上げ下げの推定結果から，上げ弓ならば右腕は上向きに．下げ弓ならば下向きに動くように右手首の座標を取得した．

2.3 アバタ転写

Unity 2020.1.8f1 を用いて 3 次元アバタへの転写を行った．上記で得られた 2 次元のモーションを 3 次元のアバタに転写するため，各関節点に対して z 座標を与えた．右手首，右肘の z 座標はヴァイオリンの上面にあり使用弦によって傾きの変化する平面上を通るように決定した．また，アバタの制御には，既存の演奏モーション生成手法 [3] [7] のアバタ転写部分に倣って逆運動学（Inverse Kinematics; IK）を用いた．IK には Lin ら [7] に倣って Unity Asset Store 内で提供されている Final IK 1.97 を採用した．また，ヴァイオリン本体のモデルは左手首，左肩に固定した．弓のモデルは右手首に固定し，ヴァイオリン本体に対して常に垂直な向きを向くように制御した．

3. 評価実験

本章では，提案手法により弓動作を反映したモーションの生成が可能であることを定量的・定性的に確認する．

3.1 実験条件

同一演奏者のヴァイオリン演奏音源 20 曲（58.2 分）を用いてデータセットを作成し，16 曲（49.0 分）をネットワークの学習，2 曲（4.2 分）を検証，2 曲（5.0 分）をテストに用いた．提案手法に用いるデータセットは，音源の STFT $\boldsymbol{X}^{stft}$ と弓の上げ下げ $\boldsymbol{L}^{ud} = \{l_1^{ud}, \ldots, l_T^{ud}\} \in \{0,1\}^T$，使用弦 $\boldsymbol{L}^{string} = \{l_1^{string}, \ldots, l_T^{string}\} \in \{0,1,2,3\}^T$ で構成される．STFT の窓幅，シフト幅はそれぞれ 256，64 とした．

定量評価の尺度には弓の上げ下げの正解率，弓の上げ下げの切り返しの F 値，及び使用弦の正解率を用い，提案モデルにより正しい弓遣いが推定可能かを評価した．弓の切り返し $\boldsymbol{L}^{attack} = \{l_1^{attack}, \ldots, l_T^{attack}\} \in \{0,1\}^T$ は以下のように定めた．

$$l_t^{attack} = |l_t^{ud} - l_{t-1}^{ud}| \tag{5}$$

F 値については，許容誤差 δ を設定し，正解ラベルが $l_t^{attack} = 1$ のとき，$\tau \in [t-\delta, t+\delta]$ の範囲で $l_\tau^{attack} = 1$ として推定結果 $\hat{\boldsymbol{L}}^{attack}$ との F 値をとった．ただし，$\delta = 3$ (1.25 秒) とした．定性評価では，生成されたモーションが音源に対して弓動作を充分に反映した動きをしているかを確認した．

ベースラインとして入力音源から直接演奏者の姿勢を推定する既存手法 Audio to Body Dynamics [3], Temporally Guided Music-to-Body-Movement Generation [4] を採用した．ただし, [4] については公開されている学習済みのモデルを使用した．ベースライン手法の弓の上げ下げの推定結果は各関節の座標の時間変化から取得し，定量評価は弓の上げ下げに関して行った．

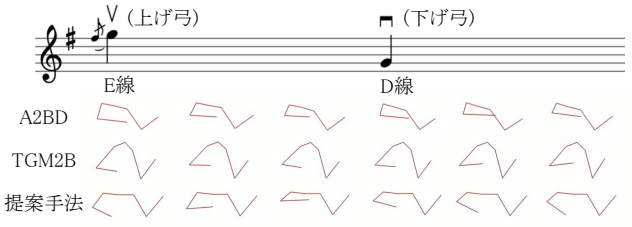

図 2: モーションの生成結果例

3.2 定量評価

表 1 にベースライン手法及び提案手法における弓の上げ下げの正解率及び F 値を示す．提案手法による結果は，正解率・F 値ともにベースライン手法を上回っており，弓の上げ下げ・切り返しにおける提案手法の有効性が確認できた．また，提案手法における使用弦の正解率は 0.801 であった．

3.3 定性評価

図 2 にテストデータの一部の楽譜と，提案手法・ベースライン手法における両腕のスケルトン時系列の生成例を示す．楽譜には弓の上げ下げ及び使用弦を表記した．提案手法により生成されたスケルトン時系列では楽譜通りの弓の上げ下げ及び弓の切り返しを反映できているのに対し，ベースライン手法では右手首が充分に動いておらず，上げ弓・下げ弓を反映できていない．また，提案手法による生成例では，譜面上での使用弦の変化も肘の高さの変化に反映されている．このように，提案手法により弓動作を反映したモーションが生成可能であることがモーションの生成結果からも確認できた．

4. おわりに

本稿では，弓動作を反映した演奏モーションの自動生成手法を提案した．評価実験の結果，提案手法における弓の上げ下げの正解率，弓の上げ下げの切り返しの F 値がベースライン手法を上回り，生成されたモーションの可視化結果とあわせて本手法の有効性が確認できた．本研究の最終目的は，人間にとって自然な演奏モーションを生成することであるため，今後はアバタに対する主観評価を含めた実験を行う．加えて，現在は単一演奏者の演奏データのみを使用しているため，複数演奏者のデータに拡張し，本手法の有効性を確認する．

謝辞 本研究は，JST ACCEL (JPMJAC1602) および JSPS 科研費 (JP19H04137) の補助を受けた．

参考文献

[1] N. Kugimoto *et al.*: “CG Animation for Piano Performance,” *ACM SIGGRAPH*, 3–7, 2009.
[2] 粟井修司ら: “ドラム演奏支援のための動作生成,” 映像情報メディア学会誌, vol. 67, No. 6, pp. J155-J162, 2013.
[3] E. Shlizerman *et al.*: “Audio to Body Dynamics,” *IEEE Conference on Computer Vision and Pattern Recognition*, 7574–7583, 2018.
[4] H. Kao *et al.*: “Temporally Guided Music-to-Body-Movemet Generation ,” *Proceedings of the 28th ACM International Conference on Multimedia*, 147–155, 2020.
[5] Z. Cao *et al.*: “Realtime multi-person 2d pose estimation using part affinity fields,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7291–7299, 2017.
[6] D. Pavllo *et al.*: “3D human pose estimation in video with temporal convolutions and semi-supervised training,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7753–7762, 2019.
[7] Y. Lin *et al.*: “A Human-Computer Duet System for Music Performance,” *Proceedings of the 28th ACM International Conference on Multimedia*, 772–780, 2020.