Supplementary Material – SyncViolinist: Music-Oriented Violin Motion Generation Based on Bowing and Fingering

Hiroki Nishizawa[1*] Keitaro Tanaka[1*] Asuka Hirata[1*] Shugo Yamaguchi[1]
Qi Feng[2†] Masatoshi Hamanaka[3] Shigeo Morishima[2]
[1]Waseda University [2]Waseda Research Institute for Science and Engineering [3]RIKEN

1. Dataset Post-processing and Synchronization

To meet the needs of our model training task, we post-processed the recorded audio, motion, and bowing/fingering information.

First, to extract audio features from the recorded signals, we used librosa, a Python library for music signal processing [3]. We applied a short-time Fourier transform (STFT) with a sliding window of length 2048 samples and a hop size of 1/30 seconds to obtain a 128-dimensional Mel-scaled spectrogram $X \in \mathbb{R}^{T\times 128}$.

Next, we used Shogun[1], a motion editing software, to process the motion data and obtain the rotation information in Euler angles for 61 joints and the 3D positions of the root (24 for the body and 38 for the fingers). Next, we downsampled the motion data to 30 fps and calculated the 3D joint position $Y \in \mathbb{R}^{T\times 62\times 3}$ from the rotation representation using forward kinematics, where T represents the number of time frames. Additionally, we retargeted the motion data to a common skeleton to eliminate the effect of variations in the violinists' body sizes and shapes while keeping their hand and fingertip positions intact.

Finally, to obtain the fingering and bowing information that is synchronized with the motion data, we used both the recorded MIDI signals and the MusicXML scores that were annotated by the violinists. It is worth noting that the MIDI signals were recorded simultaneously with the audio signals; thus, by annotating each label of the bowing/fingering information for each MIDI note, synchronization of audio features and bowing/fingering information can also be guaranteed at the same time. Since the bow direction and finger number labels were annotated in MusicXML by the violinists themselves, we were able to correspond the bow direction and finger number labels for each MIDI note actually played by synchronizing the MusicXML with the MIDI signals using a MIDI-to-score alignment method [4]. However, since this alignment is sometimes inaccurate, we manually corrected the information. For string labels, we obtained them by identifying which of the four channels had the largest output at each time frame since the MIDI violin used in our dataset acquisition process had outputs from different channels for each of the four strings. As for position labels, we decided them from the recorded MIDI note number (*i.e.*, pitch), played string, and the finger number obtained above.

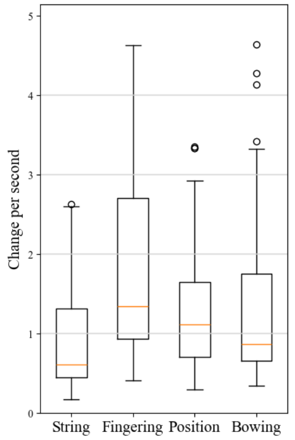

Figure 1. Statistics of motions of each component.

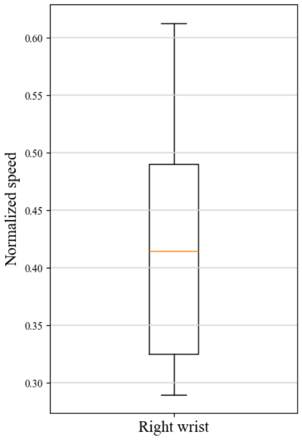

Figure 2. Speed statistics of wrist motions.

After annotating each label in bowing/fingering information to the MIDI notes, we converted them to ensure that they were consistent with the time frame of audio features.

2. Dataset Analysis

To ensure the diversity of the dataset, we conducted a comprehensive analysis covering both motion patterns and the music pieces included. Figure 1 presents statistics on the key motion components of the dataset, while Fig. 2 illustrates the variation in the right wrist's speed. Additionally, Figs. 3, 4, and 5 demonstrate the diversity of the musical pieces included in the proposed dataset.

*The first three authors contributed equally.
†Corresponding author.
[1] https://www.vicon.com/software/shogun/

3. Joint Details

Diagrams of the captured joints are illustrated in Fig. 6. Furthermore, additional lists (Tables 1 and 2) give the details of the joints available in the proposed dataset. To benchmark the dataset, we re-implemented previous models with minimal modifications (i.e. only reshaping the input and output dimensions, see Table 3) to account for the different numbers of key points.

4. Detailed Subjective Evaluation Results

Figure 7 presents the detailed results of the subjective evaluation. The left side shows the naturalness evaluation compared to other state-of-the-art methods, while the right side shows the naturalness evaluation compared to other ablative conditions. Tracks 1 and 2, 3 to 5, and 6 to 8 correspond to the test data of players 1, 2, and 3, respectively.

References

[1] Jiali Chen, Changjie Fan, Zhimeng Zhang, Gongzheng Li, Zeng Zhao, Zhigang Deng, and Yu Ding. A music-driven deep generative adversarial model for guzheng playing animation. *IEEE Transactions on Visualization and Computer Graphics*, 29(2):1400–1414, 2023. 6

[2] Hsuan-Kai Kao and Li Su. Temporally guided music-to-body-movement generation. In *Proceedings of the ACM International Conference on Multimedia*, pages 147–155, 2020. 6

[3] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25, 2015. 1

[4] Eita Nakamura, Kazuyoshi Yoshii, and Haruhiro Katayose. Performance error detection and post-processing for fast and accurate symbolic music alignment. In *Proceedings of the Conference of the International Society for Music Information Retrieval*, pages 347–353, 2017. 1

[5] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7574–7583, 2018. 6

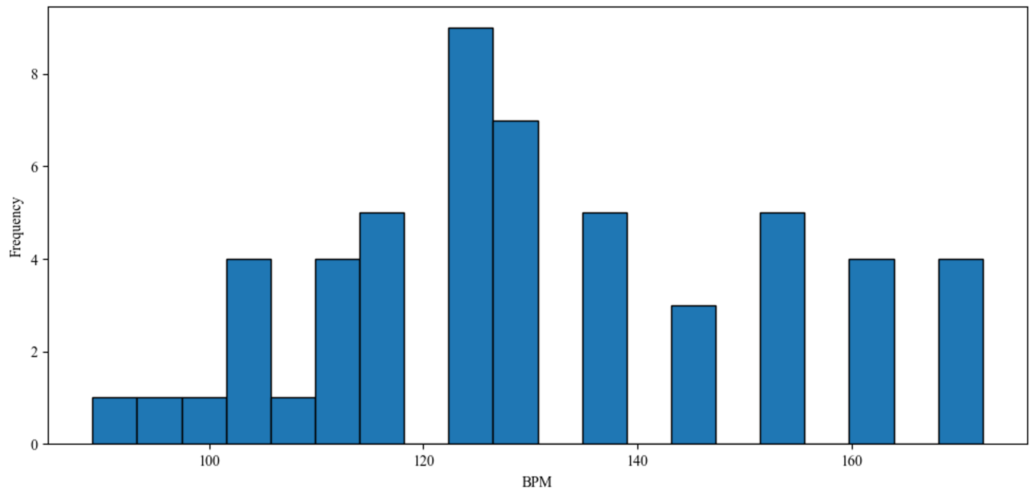

Figure 3. BPM distribution.

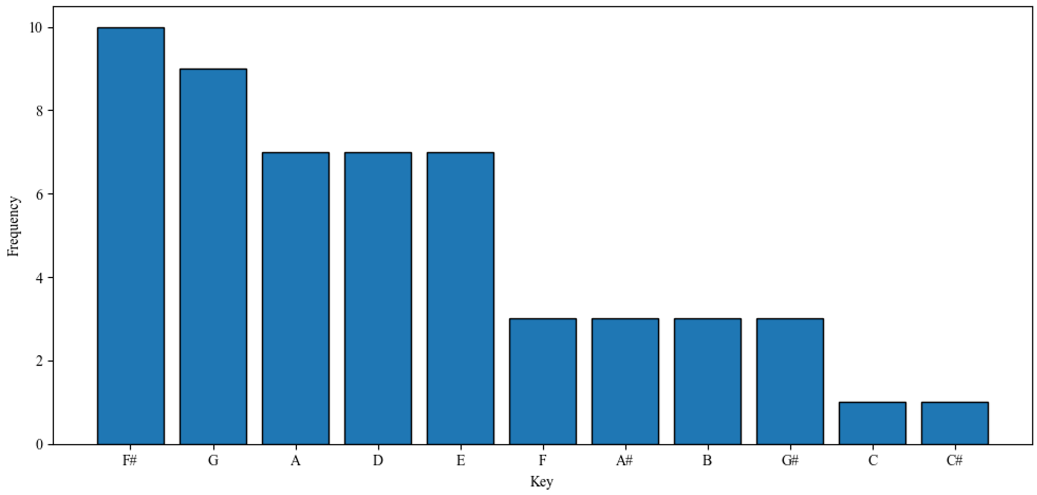

Figure 4. Key distribution.

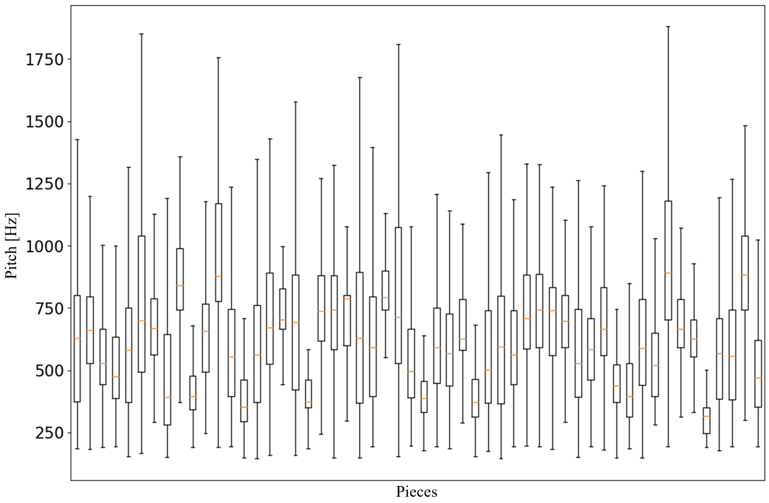

Figure 5. Pitch distribution.

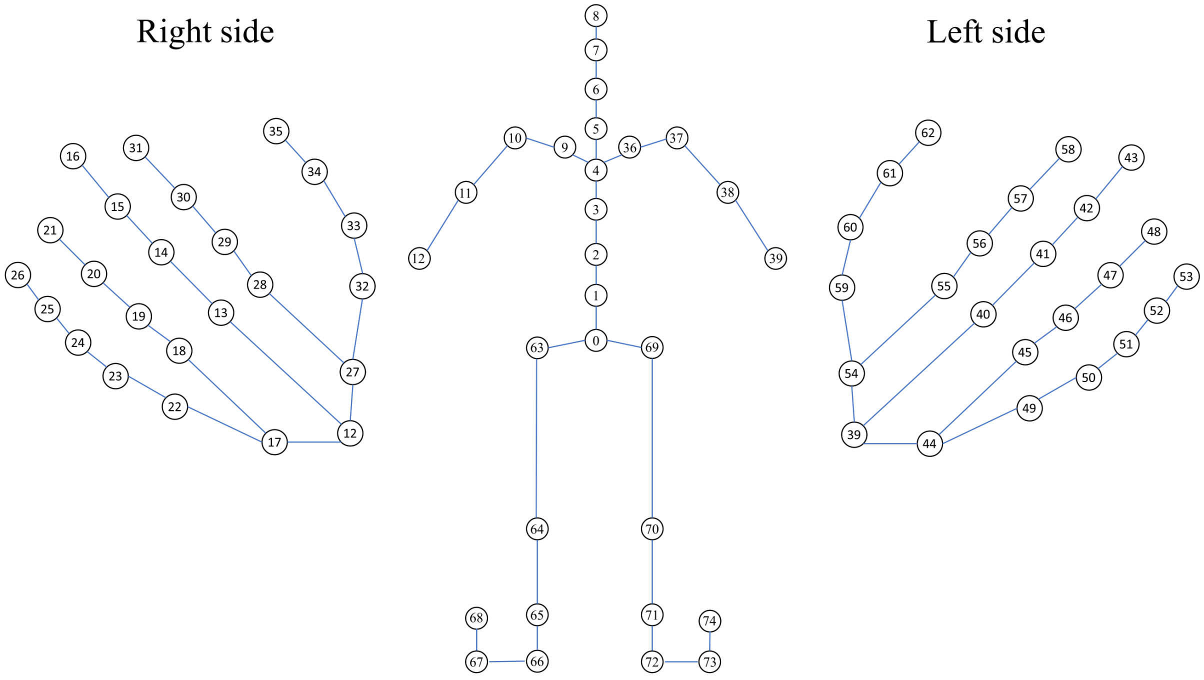

Figure 6. Diagrams of the captured joints from the proposed dataset.

Node number	Joint name	Parent node	Body parts	
			in generation	in evaluation
0	Hips	-	Others	-
1	Spine	0	Others	-
2	Spine1	1	Others	-
3	Spine2	2	Others	-
4	Spine3	3	Others	-
5	Neck	4	Others	-
6	Neck1	5	Others	-
7	Head	6	Others	-
8	End Site	7	Others	-
9	RightShoulder	4	Others	-
10	RightArm	9	Others	-
11	RightForeArm	10	Right hand	Right arm (RA)
12	RightHand	11	Right hand	Right arm (RA)
13	RightHandMiddle1	12	Right hand	-
14	RightHandMiddle2	13	Right hand	-
15	RightHandMiddle3	14	Right hand	-
16	End Site	15	Right hand	-
17	RightHandRing	12	Right hand	-
18	RightHandRing1	17	Right hand	-
19	RightHandRing2	18	Right hand	-
20	RightHandRing3	19	Right hand	-
21	End Site	20	Right hand	-
22	RightHandPinky	17	Right hand	-
23	RightHandPinky1	22	Right hand	-
24	RightHandPinky2	23	Right hand	-
25	RightHandPinky3	24	Right hand	-
26	End Site	25	Right hand	-
27	RightHandIndex	12	Right hand	-
28	RightHandIndex1	27	Right hand	-
29	RightHandIndex2	28	Right hand	-
30	RightHandIndex3	29	Right hand	-
31	End Site	30	Right hand	-
32	RightHandThumb1	27	Right hand	-
33	RightHandThumb2	32	Right hand	-
34	RightHandThumb3	33	Right hand	-
35	End Site	34	Right hand	-

Table 1. Detailed list of the captured joints from the proposed dataset (1/2).

Node number	Joint name	Parent node	Body parts	
			in generation	in evaluation
36	LeftShoulder	4	Others	-
37	LeftArm	36	Others	-
38	LeftForeArm	37	Left arm	Left arm (LA)
39	LeftHand	38	Left arm	Left arm (LA)
40	LeftHandMiddle1	39	Left hand	-
41	LeftHandMiddle2	40	Left hand	-
42	LeftHandMiddle3	41	Left hand	-
43	End Site	42	Left hand	Left fingers (LF)
44	LeftHandRing	39	Left hand	-
45	LeftHandRing1	44	Left hand	-
46	LeftHandRing2	45	Left hand	-
47	LeftHandRing3	46	Left hand	-
48	End Site	47	Left hand	Left fingers (LF)
49	LeftHandPinky	44	Left hand	-
50	LeftHandPinky1	49	Left hand	-
51	LeftHandPinky2	50	Left hand	-
52	LeftHandPinky3	51	Left hand	-
53	End Site	52	Left hand	Left fingers (LF)
54	LeftHandIndex	39	Left hand	-
55	LeftHandIndex1	54	Left hand	-
56	LeftHandIndex2	55	Left hand	-
57	LeftHandIndex3	56	Left hand	-
58	End Site	57	Left hand	Left fingers (LF)
59	LeftHandThumb1	54	Left hand	-
60	LeftHandThumb2	59	Left hand	-
61	LeftHandThumb3	60	Left hand	-
62	End Site	61	Left hand	Left fingers (LF)
63	RightUpLeg	0	Others	-
64	RightLeg	63	Others	-
65	RightFoot	64	Others	-
66	RightForeFoot	65	Others	-
67	RightToeBase	66	Others	-
68	End Site	67	Others	-
69	LeftUpLeg	0	Others	-
70	LeftLeg	69	Others	-
71	LeftFoot	70	Others	-
72	LeftForeFoot	71	Others	-
73	LeftToeBase	72	Others	-
74	End Site	73	Others	-

Table 2. Detailed list of the captured joints from the proposed dataset (2/2).

Method	before	after
Shlizerman et al. [5]	98	255
Kao and Su [2]	45	225
Chen et al. [1]	188	248
Ours	-	225

Table 3. Reshaping of dimensions for existing methods.

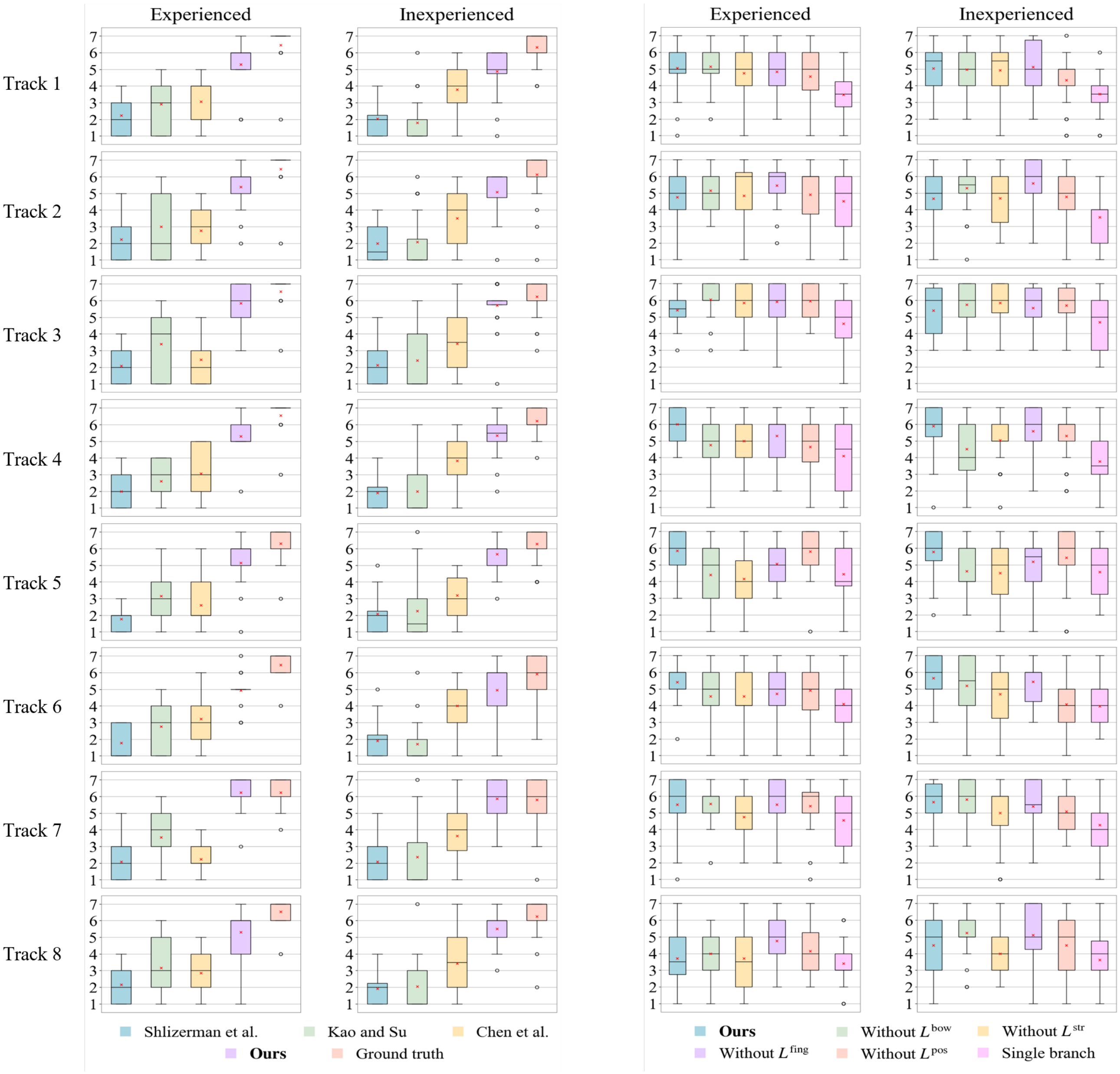

Figure 7. Detailed results of the subjective evaluation against state-of-the-art methods (left) and ablative conditions (right).