Bowing-Net: Motion Generation for String Instruments Based on Bowing Information

Asuka Hirata
Waseda University
Japan
asuka112358@suou.waseda.jp

Keitaro Tanaka
Waseda University
Japan
phys.keitaro1227@ruri.waseda.jp

Ryo Shimamura
Waseda University
Japan
s-ryo@akane.waseda.jp

Shigeo Morishima
Waseda Research Institute for Science and Engineering
Japan
shigeo@waseda.jp

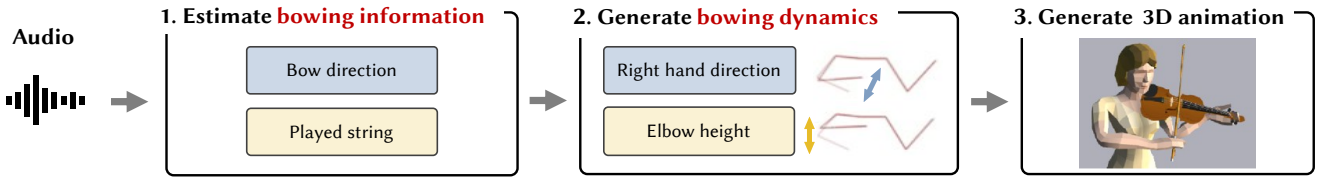

Figure 1: Overview of our proposed method. To reflect the bowing dynamics of the right hand to the final output animation, our model first estimates the bowing information from the given audio.

ABSTRACT

This paper presents a deep learning based method that generates body motion for string instrument performance from raw audio. In contrast to prior methods which aim to predict joint position from audio, we first estimate information that dictates the bowing dynamics, such as the bow direction and the played string. The final body motion is then determined from this information following a conversion rule. By adopting the bowing information as the target domain, not only is learning the mapping more feasible, but also the produced results have bowing dynamics that are consistent with the given audio. We confirmed that our results are superior to existing methods through extensive experiments.

CCS CONCEPTS

• **Computing methodologies** → **Motion processing**; • **Applied computing** → **Media arts**; **Sound and music computing**.

KEYWORDS

Motion generation, music information retrieval, neural networks.

ACM Reference Format:
Asuka Hirata, Keitaro Tanaka, Ryo Shimamura, and Shigeo Morishima. 2021. Bowing-Net: Motion Generation for String Instruments Based on Bowing Information. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '21 Posters), August 09-13, 2021.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3450618.3469170

SIGGRAPH '21 Posters, August 09-13, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8371-4/21/08.
https://doi.org/10.1145/3450618.3469170

1 INTRODUCTION

Creating natural animation for musical instrument performance is a costly process that requires laborious steps such as motion capture and manual tuning. Therefore, researches on body motion generation for musical instrument performance from an input of either musical note or raw audio have been conducted in the past. However, the problem with these existing methods is that they failed to create adequate results for string instruments as the mapping between playing procedures and produced sounds are far more ambiguous compared to instruments such as the piano.

In this paper, we propose Bowing-Net, a deep learning based method capable of producing plausible body motion for string instrument performance from the raw audio input. Raw audio is used as input instead of musical notes as in [Kugimoto et al. 2009] because audio characteristics such as timbre or tone, which affect the bowing dynamics, are not represented in the notes. Moreover, in contrast to methods such as [Shlizerman et al. 2018] [Kao and Su 2020], which directly estimate joint position from audio, we first estimate bow direction, *i.e.,* whether the bow moves up or down relative to the instrument, and the played string. Then the final body motion is determined so that they agree with the initially estimated bowing information. This is based on our insight that 1) bowing dynamics are visually essential elements and thus the bowing information should be the target domain, and 2) bowing information has less complicated correspondence with the audio than joint position and thus should be easier to predict. We show the effectiveness of our method through evaluation on the motion generated from violin performance audio.

2 METHOD

The overview of our method is shown in Figure 1. First, the bowing information, which consists of the bow direction and the played string, is estimated from the input audio. These two components are

Table 1: Quantitative evaluation.

	Bow direction acc.	F measure	String acc.
A2BD	0.560	0.461	-
TGM2B	0.493	0.422	-
Ours	**0.704**	**0.566**	0.801

estimated because they dictate the bowing dynamics and thus the movements of the right arm. Accordingly, body motion in 2D space is determined from the bowing information following a conversion rule. In the final step, the estimated 2D body motion is transferred to a 3D avatar.

For the estimation of the bowing information, the short-time Fourier transform (STFT) of the given audio is first calculated. Then, the bow direction $\hat{L}^{\mathrm{ud}} = \hat{l}^{\mathrm{ud}}_{1:T} \in \{0,1\}^T$ (0:down, 1:up) is estimated, where T is the number of total time frames. To take time series into account, we utilize a long short-term memory (LSTM) network, whose output p^{ud}_t is the probability that the bow direction is up at frame t ($0 \leq p^{\mathrm{ud}}_t \leq 1$). The bow direction $\hat{l}^{\mathrm{ud}}_t$ is obtained by applying a threshold $\theta = 0.5$ to p^{ud}_t as follows:

$$\hat{l}^{\mathrm{ud}}_t = \begin{cases} 0 & (p^{\mathrm{ud}}_t < \theta) \\ 1 & (p^{\mathrm{ud}}_t \geq \theta). \end{cases} \tag{1}$$

Similarly, the played string $\hat{L}^{\mathrm{str}} = \hat{l}^{\mathrm{str}}_{1:T} \in \{0,1,2,3\}^T$ is obtained using another LSTM network, where $\hat{l}^{\mathrm{str}}_t$ designates which of the four is being played at frame t. For example, in the case of the violin, 0, 1, 2, and 3 correspond to E, A, D, and G string, respectively.

Based on the estimated bowing information above, we determine the bowing dynamics in 2D space following a conversion rule. Specifically, we determine the positions of the right elbow and the right hand, which decide the bowing dynamics. First, we determine the height of the right elbow from the estimation results of the played string, as skilled players fix the position of their right elbow while they keep playing the same string. In the case of the violin, the elbow height is lower (higher) when a player uses strings for the higher (lower) tone. Based on this nature, we manually determine the elbow position for each of the four strings. In string transition, we smoothly change the elbow positions using interpolation.

Next, since the bow movement always follows a unique trajectory for each string, the movement of the right hand is determined from the elbow position and the estimation result of the bow direction. We make the right hand move at a constant speed except for long tones where the speed is made slower. Finally, we transfer the 2D bowing dynamics to a 3D avatar by calculating the z-coordinates of each joint.

3 EVALUATION

Both quantitative and qualitative evaluations were conducted on our generated results. We built a dataset containing 20 pieces (58.2 minutes) of audio by a single violin player. The train:validation:test split ratio is 16:2:2. The dataset in the proposed method is composed of audio and ground truth bowing information (*i.e.,* the per-frame bow direction and played string). To evaluate whether correct bowing can be estimated by the proposed model, we measured the accuracy for bow direction and played string, and also the F measure for the bow attack. Note that the bow attack l^{att}_t indicates the point of transition between bow directions, which is defined as

$$l^{\mathrm{att}}_t = |l^{\mathrm{ud}}_t - l^{\mathrm{ud}}_{t-1}|. \tag{2}$$

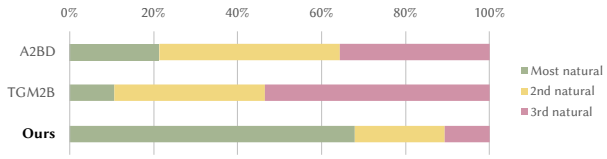

Figure 2: Subjective evaluation for naturalness.

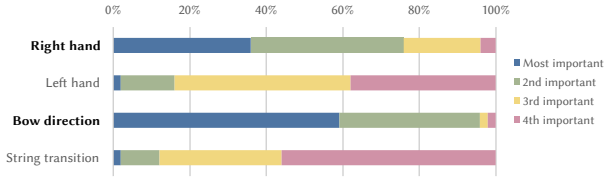

Figure 3: Subjective evaluation for important elements.

We compared our method with two baselines [Shlizerman et al. 2018] [Kao and Su 2020], which generate body movements directly from given audio. The bow directions of the baseline methods were obtained by calculating the displacement of the joint positions between consecutive frames. Table 1 shows the quantitative result. These results show that our method is better at reproducing the bowing dynamics than other methods. Also, the accuracy for the played string by our method is 0.801.

In addition, we conducted a subjective evaluation. We asked 28 participants to answer two questions after watching 10 videos of comparison between results generated by the two baseline methods and ours. First, for each video, we asked them to order the naturalness of the videos. Next, we asked them to order the importance of some visual elements when judging the naturalness of the body motion in the previous question. Figure 2 and Figure 3 show the results. Our results are shown to be more natural than other results. Also, it is shown that the naturalness of our results is most likely due to its quality of the bowing dynamics of the right hand and especially the bow direction. For generated results, please refer to our supplementary video.

4 CONCLUSION

We proposed a novel method of motion generation for string instrument performance from raw audio. We confirmed that our generated motion appears more natural than the baselines through extensive experiments. For future work, we plan to expand our dataset to multiple performers and further generalize our method to string instruments other than the violin.

ACKNOWLEDGMENTS

This work was supported by JST Mirai Program No. JPMJMI19B2, and JSPS KAKENHI Nos. 19H01129 and 19H04137.

REFERENCES

Hsuan-Kai Kao and Li Su. 2020. Temporally Guided Music-to-Body-Movement Generation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 147–155.

Nozomi Kugimoto, Rui Miyazono, Kosuke Omori, Takeshi Fujimura, Shinichi Furuya, Haruhiro Katayose, Hiroyoshi Miwa, and Noriko Nagata. 2009. CG animation for piano performance. In *SIGGRAPH'09: Posters*.

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