3D skeleton motion generation of double bass from musical score

Takeru Shirai and Shinji Sako

Nagoya Institute of Technology
clf19021@ict.nitech.ac.jp

Abstract. In this study, we propose a method for generating 3D skeleton motions of a double bass player from musical score information using a 2-layer LSTM network. Since there is no suitable dataset for this study, we have created a new motion dataset with actual double bass performance. The contribution of this paper is to show the effect of combining bowing and fingering information in the generation of performance motion, and to examine the effective model structure in performance generation. Both objective and subjective evaluations showed that the accuracy of generating performance motion for double bass can be improved using two types of additional information (bowing, fingering information) and improved by constructing a model that takes into account bowing and fingering.

Keywords: LSTM network, Performance motion generation, 3D model, Double bass

1 Introduction

Double bass plays an important role as the foundation in various forms of ensemble music such as orchestral music, chamber music, wind music, and jazz. In addition, the double bass plays a solo role while accompanied by the piano or orchestra. In the case of the bowed stringed instrument to which the contrabass belongs, there is so much visual information that the timing of the sound can be shared among the players by the motion of the right arm, and the pitch can be estimated by the shifting and fingering of the left hand.

In an actual ensemble performance, visual information is an important element for conveying performance timing and specific musical expressions to other players and for facilitating ensemble performance [1]. In particular, visual information is considered to be highly important in situations where many people are playing together in an ensemble, such as in an orchestra or wind band.

In spite of the fact that visual information is one of the most important elements in playing music as described above, among the major study fields of music information processing, the studies on automatic performance generation (i.e. performance rendering) mainly focus on performance sounds, and only a few studies focus on visual information of performances.

Therefore, in this study, we aim to generate performance motion for the double bass. There are two major technical issues: the generation of natural playing motion and the

naturalness of the 3D model appearance and rendering accuracy. In this study, we first focus on the former, which is the more essential issue.

There have been some studies on automatic performance generation focusing on visual information, but they have targeted piano [2] and violin [3, 4], and generated performance motion using actual performance sounds or MIDI as input data. In the case of a bowed stringed instrument, it is considered that it is difficult to generate the performance motion such as bowing and fingering from the pitch information because motion is not uniquely determined from the pitch information. In order to solve this problem, we propose a method to generate performance motion using the musical score information that includes not only pitch information, but also bowing and fingering information that greatly affects performance motion. Some methods have been proposed for automatically estimating fingering from musical score [5], and a method combining these studies would be promising, but in this study, we suppose these additional information are added manually.

Since there is no suitable dataset for this study, it is necessary to construct a dataset of musical scores and 3D motions. In previous studies, joint points extracted by using body tracking technology of video data were used as motion information [2–4]. This approach is also superior in that it does not interfere with the playing motion. However, in this study, which targets a large instrument such as a double bass, it is considered difficult to obtain accurate performance motion using this technique because part of the performer is hidden by the instrument. Therefore, we collect motion data using the inertial motion capturing device.

In this study, we adopted LSTM (Long Short Term Memory) network as a model for the conversion between musical score data and motion data. In particular, we verify the effect of using additional information (bowing and fingering information) as input data by comparing the accuracy of the generated motion only from pitch information and with additional information. Furthermore, we design a series model that learns the right arm motion from the bowing information and the left arm motion from the fingering information independently, and verify the effect of changing the structure of the model.

2 Related works

Li et al. [2] proposed a method to generate a pianist's 2D motion from MIDI sound sources of a piano performance. They used a Convolutional Neural Network (CNN) to extract the stream of the piano performance and the features of the beat structure, and used these as input data to the 2-layer LSTM network, and used the 2D performance motion from a fixed position as output data. In the subjective experiment, no significant difference was found between the human motion and the generated motion in 75% of the songs, indicating that the system does not generate extremely unnatural motion.

Liu et al. [3] proposed a method for generating violinist's performance motion from actual performance sounds. In this method, a model for predicting the bowing of the right arm and a model for predicting the expressive motions of the whole body were constructed from the Mel-spectrogram[1] obtained by performing STFT (Short Time

[1] Spectrogram in the Mel scale, a perceptual measure of pitch in human hearing.

Fourier Transform) on the input sound source. And a model for predicting the position, fingering, and strings of the left arm from the data obtained by pitch detection is constructed independently. In addition, a model that predicts the position of the left arm, fingering, and strings based on the data obtained from pitch detection was constructed independently, thereby realizing the generation of violinist's full-body performance motions.

3 Proposed method

Our model is based on a previous study by Li et al. [2]. The difference between our model and the previous study is that the output data is not 2-dimensional but 3-dimensional, and the input is not derived from MIDI but from score information, which is a sequence of symbols. We need to consider a model that addresses these differences.

We construct a 2-layer LSTM network, and use MAE (Mean Absolute Error) as the loss function and Adam [6] as the optimize function. The output vectors of the LSTM network are fed to all the coupling layers to obtain the positional and rotational information of each joint point in each frame in 6 dimensions.

We also attempt to apply the framework constructed by Liu et al. [3]which consists of three models: a bowing model for the right arm, a position model for the left arm, and a representation model for the upper body. In this study, since we are trying to generate performance motion using manually additional information (bowing, fingering) rather than performance sound data, we can treat these information as more accurate and reliable than that obtained by estimation.

Extract the sequence of pitch, bowing, and position from the musical score information as shown in the Fig. 1 into a format that can be input to the LSTM network, each with the same period. As a result, the pitch sequence is a 30-dimensional sequence consisting of $\{E0, F0, \ldots, A3\}$, the bowing sequence is a 2-dimensional sequence consisting of $\{down\text{-}bow, up\text{-}bow\}$, and the position sequence representing fingering information is a 12-dimensional sequence consisting of $\{0, 1, \ldots, 11\}$.

The three sequences extracted from the above score information are used as input data, and the sequence representing body motions are used as output data to construct a model. The goal is to verify the significance of each data and to design a model that is suitable for learning. By comparing the accuracy of the generated performance motion by the designed models, we can verify whether the bowing and position information used as additional information are significant in improving the accuracy of the generated performance motion.

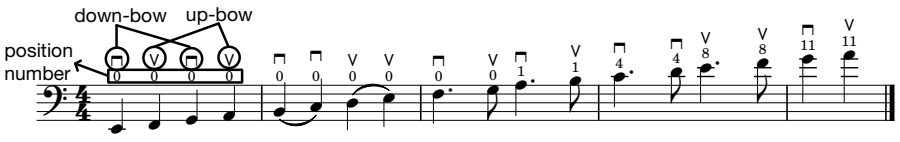


Fig. 1: Sample of musical score

	input	output
Model1	pitch(30-d)	upper body(90-d)
Model2	bowing+position+pitch(44-d)	upper body(90-d)
Model3	bowing(2-d), position(12-d), pitch(30-d)	right arm(24-d), left arm(24-d), other(42-d)

Table 1: Structure of the three models

The structure of the three models we designed is summarized in the Table 1.

4 Experiment

4.1 Dataset

We use ten pieces from the collection of exercises "Franz Simandl / 30 Etudes for the Double Bass" from No.1 to No.10 for the training data, and three pieces from No.11 to No.13 for the test data. The total time to play these 13 pieces at the tempo specified in the score is about 30 minutes.

Motion data We use an inertial motion capture PERCEPTION NEURON made by NOITOM to construct a dataset of the performance motion of one male double bass player. The bvh file is a motion capture data file format proposed by Biovision, and consists of two parts: a hierarchy part describing the tree structure of each joint point, and a motion part describing the motion data. In this study, the hierarchy part was defined as the 15 joint points of the upper body with the hip as the parent node. And since the motion part describes the position information and rotation information of each of the 15 joint points, it is represented by a 90-dimensional sequence. In this dataset, the coordinates of the parent node are set to the origin.

Since the experiment was intended to be performed at the tempo dictated by the musical score, we recorded the music performance played to a metronome. The frame rate was set to 30 fps in accordance with previous research [4]. Since the accelerometers at each joint point may deviate from their default positions due to motion during performance, calibration (correction of deviations in sensor position information) was performed after each etude was recorded.

Musical score data The musical score data was authorized for the target etudes in MusicXML format [7] using the score authoring software MuseScore. Since this study does not target the generation of expressive motion, we exclude tempo changes, volume marks, and detailed articulation instructions such as tenuto and staccato from the authoring.

In the original score, there is no bowing and fingering information for all notes, so we added symbols as bowing information and position numbers as fingering information, as shown in the Fig. 1. The position number in this case is not the actual position where the string is pressed, but the position of the index finger when pressing the string,

and is defined as position number($= \{0, 1, \cdots, 10, 11\}$), starting with the lowest pitch position.

Extract a 30-dimensional pitch sequence, a 2-dimensional bowing sequence, and a 12-dimensional position sequence for every etude in order to get the pitch, bowing, and position information from the xml data into a format that can be input to LSTM network. In this process, each information was extracted at 30 fps to match the frame rate of the motion data.

4.2 Objective evaluation

For objective evaluation, we compare the generated data with correct data (motion data collected under the same conditions as when the data set was constructed), using the following two criteria.

1. Average of the difference of coordinates at each joint point in each frame

2. Average of the ratio of the change between adjacent frame at each joint point

In the criterion 1, accuracy is verified by the difference of the amount of motion of all joint points, so the smaller the value, the higher the accuracy. The criterion 2 takes into account the problem that the only evaluation based on the criterion 1 is not sufficient because of the not so small difference in the amount of motion among joints. The criterion 2 verifies the accuracy by the ratio, so the closer the value is to 1, the higher the accuracy.

The results for the criterion 1 are shown in Fig. 2(a), and the results for the criterion 2 are shown in Fig. 2(b). From these two figures, it can be seen that the order of accuracy is Model1 $<$ Model2 $<$ Model3.

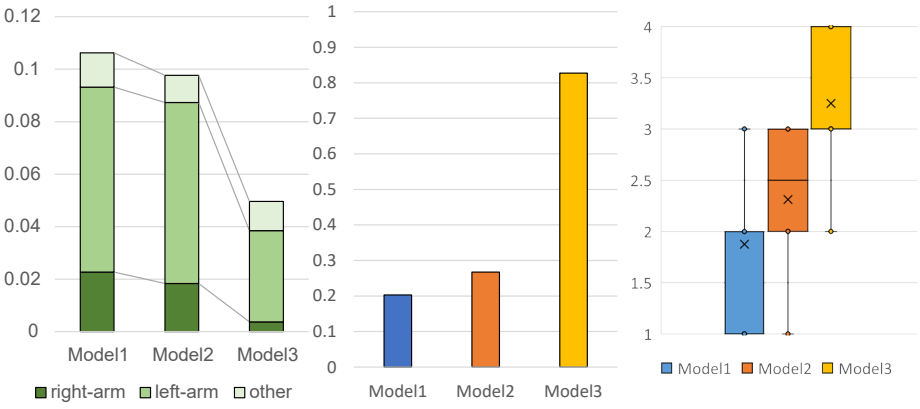
4.3 Subjective evaluation

The subjective evaluation is based on the naturalness of the performance motion. In order to make this evaluation, it is necessary to have a person who can concretely imagine the performance motions of the player from the score information, so the subjects of the evaluation experiment were limited to double bass players. After checking the score, the subjects watched a movie of the generated motion data played on Blender. In this evaluation experiment, the order of playback was randomized. A total of 16 double bass players, 8 males and 8 females in their early 20s, evaluated the naturalness in four levels: “1: unnatural”, “2: somewhat unnatural”, “3: somewhat natural” and “4: natural”.

From Fig. 2(c),which show the results of subjective evaluation using a box-and-whisker diagram, it can be seen that the order of accuracy is Model1 $<$ Model2 $<$ Model3. This is consistent with the result of the objective evaluation.

5 Conclusion

In this study, we proposed a method for generating performance motions of a double bass, for which it is difficult to predict performance motions from audio signals, by using musical score information (pitch, bowing, and fingering) as input data. As a result



(a) Objective evaluation(1) **(b)** Objective evaluation(2) **(c)** Subjective evaluation

Fig. 2: Results of evaluation

of the experiments, it was demonstrated that there was a positive effect of providing additional information (bowing and fingering), and that a higher effect could be obtained by learning the right arm and the left arm independently from the bowing and fingering information. As a future task, the generation of expressive performance motion is considered. In addition, the generation of realistic performance motions using 3D human models will be useful for performance training for beginners.

References

1. Satoshi Kawase. Communication between ensemble performers: Coodination cues. *Japanese psychological review*, Vol. 57, No. 4, pp. 495–510, 2014. (in Japanese).
2. Bochen Li, Akira Maezawa, and Zhiyao Duan. Skeleton Plays Piano: Online Generation of Pianist Body Movements from MIDI Performance. In *International Society for Music Information Retrieval*, pp. 218–224, 2018.
3. Jun-Wei Liu, Hung-Yi Lin, Yu-Fen Huang, Hsuan-Kai Kao, and Li Su. Body Movement Generation for Expressive Violin Performance Applying Neural Networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3787–3791. IEEE, 2020.
4. 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*, pp. 7574–7583, 2018.
5. Wakana Nagata, Shinji Sako, and Tadashi Kitamura. Violin fingering estimation according to skill level based on hidden markov model. In *International Computer Music Conference*, 2014.
6. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
7. Michael Good. MusicXML: An internet-friendly format for sheet music. In *Xml conference and expo*, pp. 03–04. Citeseer, 2001.