Temporally Guided Music-to-Body-Movement Generation
Hsuan-Kai Kao and Li Su
Institute of Information Science, Academia Sinica, Taiwan
{hsuankai,lisu}@iis.sinica.edu.tw

Figure 1: Our task is to create reasonable and natural playing movement with corresponding violin music.

ABSTRACT
This paper presents a neural network model to generate virtual violinist’s 3-D skeleton movements from music audio. Improved from the conventional recurrent neural network models for generating 2-D skeleton data in previous works, the proposed model incorporates an encoder-decoder architecture, as well as the self-attention mechanism to model the complicated dynamics in body movement sequences. To facilitate the optimization of self-attention model, beat tracking is applied to determine effective sizes and boundaries of the training examples. The decoder is accompanied with a refining network and a bowing attack inference mechanism to emphasize the right-hand behavior and bowing attack timing. Both objective and subjective evaluations reveal that the proposed model outperforms the state-of-the-art methods. To the best of our knowledge, this work represents the first attempt to generate 3-D violinists’ body movements considering key features in musical body movement.

CCS CONCEPTS
• Computing methodologies → Motion processing; • Applied computing → Media arts; Sound and music computing; • Human-centered computing → Sound-based input / output.

KEYWORDS
Neural networks, pose estimation, body movement generation, music information retrieval

1 INTRODUCTION
Music performance is typically presented in both audio and visual forms. Musician’s body movement acts as the pivot to connect audio and visual modalities, since musicians employ their body movement to produce the performed sound, and such movement also serves as the means to communicate their musical ideas toward the audience. As the result, the analysis, interpretation, and modeling of musicians’ body movement has been an essential research topic in the interdisciplinary fields for music training [7, 23], music recognition [10, 17], biomechanics, and music psychology [2, 6, 11, 27, 29]. Motion capture and pose estimation techniques [22] facilitated quantitative analysis of body motion by providing the data describing how each body joint moves with time. Beyond such research works based on analysis, an emerging focus is to develop a generative model that can automatically generate body movements from music. Such a technique can be applied to music performance animation, and human-computer interaction platforms, in which the virtual character’s body movement can be reconstructed from audio signal alone, without the physical presence of human musician. Several studies endeavor to generate body movement from audio and music signals, including generating pianist’s and violinist’s 2-D skeletons from music audio [16, 18, 26], generating hand gestures from conversational speech [8], and generating choreographic movements from music [13, 14].

In this paper, we focus on the generation of violinists’ body movement. Violinists’ body movement is highly complicated and intertwined with the performed sound. To investigate musical movement, previous research identified three main types of body movement in music performance. First, the instrumental movement leads to the generation of instrument sound; second, the expressive movement induces visual cues of emotion and musical expressiveness;
and third, the communicative movement interacts with other musicians and the audience [29]. Taking a violinist’s instrumental movement as an example, a bow stroke is a movement executed by the right hand to make the bow moving across the string. For a bowed note termed *arco*, there are two typical bowing modes: up-bow (the bow moving upward) and down-bow (the bow moving downward). The arrangement of bow strokes depends on how the musician segments a note sequence. In general, a group of notes marked with a *slur* on the score should be played in one bow stroke. Yet the music scores do not usually contain detailed bowing annotations for every note through the whole music piece, but only provide suggested bowing marks for several important instances, which renders musicians a lot of freedom to apply diverse bowing strategies according to their own musical interpretations. Given the flexibility of bowing in the performance practice, still, the bowing configuration should be arranged in a sensible manner to reflect the structure in music compositions. An unreasonable bowing attack (i.e., the time instance when the bowing direction changes) timing can be easily sensed by experienced violinists. Likewise, the left-hand fingering movement is also flexible to a certain extent: an identical note can be played with different strings at different fingering positions, depending on the pitches of successive notes. In addition to the instrumental movements (bowing and fingering motion), which are directly constrained by the written note sequence in the music scores, the expressive body movements also reflect the context-dependent and subject-dependent musical semantics, including the configuration of beat, downbeat, phrasing, valence, and arousal in music [2, 23]. In sum, the musical body movements have diverse functions and are attached to various types of music semantics, which leads to the high degree of freedom for movement patterns during the performance. The connection between the performed notes and body movements (including the right-hand bowing movements and left-hand fingering movements) is not one-to-one correspondence, but is highly mutual-dependent. Such characteristics not only make it difficult to model the correspondence between music and body movement, but also result in issues regarding the assessment of generative model: since there is no exact ground truth in body movement for a given music piece, it is not certain that if the audience’s perceptual quality can be represented by simplified training objective (e.g., the distance between the predicted joint position and a joint position selected from a known performance).

In this paper, we propose a 3-D violinist’s body movement generation system, which incorporates musical semantics including the beat timing and bowing attack inference mechanisms. Following the track in [18], we model the trunk and the right hand segments separately, and further develop this approach into an end-to-end, multi-task learning framework. To incorporate the musical semantic information in model training, the beat tracking technique is applied to guide the processing of input data. Moreover, a state-of-the-art 3-D pose estimation technique is employed to capture the depth information of skeleton joints, which is critical in identifying bowing attacks. The pose estimation process provides reliable pseudolabels motion data to facilitate the training process. To investigate the non-one-to-one motion-music correspondence, we propose a new dataset, which contains music with multiple performance versions by different violinists for the same set of repertoire. The generative models are evaluated on multiple performance in order to reduced the bias. To the best or our knowledge, this work represents the first attempt to generate 3-D violinists’ body movements, as well as to consider information from multiple performance versions for the development of body movement generation system.

The rest of this paper is organized as follows. Section 2 presents a survey of recent research regarding body movement generation techniques. The proposed method is introduced hereafter, where Section 3 describes the data processing, and Section 4 describes the model implementation. Section 5 reports the experiment and results, followed by the conclusion in Section 6.

2 RELATED WORK

Music body movement analysis The role of music body movement has been discussed in many studies [6]. Music performer’s body movements are divided into three types: 1) the instrumental movement such as striking the keyboard on piano or pressing the strings on violin; 2) the expressive movement such as body swaying and head nodding; and 3) the communicative movement such as cuing movement suggesting tempo changes in music [29]. Studies showed that music performers could intentionally adopt different body movements to achieve the planed performance sound according to the musical context [6, 11, 19, 20]. For instance, different violinists may choose various bowing and fingering strategies depending on the musical interpretations they attempt to deliver.

Previous research has shown that body movements from different music performers generate diverse instrumental sounds [5, 20]. The correspondence between music performer’s movement and the musical composition being performed has also been discussed [9, 27]. Recently, a study employs body movement data with the recurrent neural network (RNN) model to predict dynamic levels, articulation styles, and phrasing cues instructed by the orchestral conductor [10]. Since detecting musical semantics from the body movement data is possible, an interesting yet challenging task is to generate body movement data from given musical sound [16, 26].

Generating audio from body movement Techniques have been developed to generate speech or music signals from body movement [4, 31]. [4] generated human speech audio from automatic lip reading on the face videos, whereas [31] generated co-speech movements including iconic and metaphoric gestures from speech audio. [3] applied Generative Adversarial Networks (GAN) to produce music performer’s images based on different types of timbre. [1] generated music from gathering the motion capture data. In the field of interactive multimedia, using gesture data to induce sound morphing or related generation task is also commonly used.

Body movement generation from audio Several attempts have been devoted to generate music-related movement. The commonly seen topics of body movement generation from audio include generating body movements from music, generating gestures from speech, and generating dance from music [26] used an RNN with long-short-term-memory (LSTM) units to encode audio features, and then employed a fully-connected (FC) layer to decode it into the body skeleton keypoints of either pianists or violinists. In [13], choreographic movements are automatically generated from music according to the user’s preference and the musical structural
context, such as the metrical and dynamic arrangement in music. Another recent work on pianists’ body skeleton generation [16] also considers musical information including bar and beat positions in music. The model combining CNN and RNN was proven to be capable of learning the body movement characteristics of each pianist.

3 DATA AND PRE-PROCESSING

In this section, we introduce the procedure to compile a new violin performance dataset for this study. The data pre-processing procedure is summarized in Figure 2.

3.1 Dataset

We propose a newly-collected dataset containing 140 violin solo videos with total length of around 11 hours. 14 selected violin solo pieces were performed by 10 violin-major students from music college. This dataset therefore contains diverse performed versions and individual musical interpretations based on the same set of repertoire, which is specifically designed for the exploration of non-one-to-one correspondence between music motion and audio. The selected repertoire contains 12 conventional Western classical pieces for violin solo ranging from Baroque to post-Romanticism, plus two non-Western folk songs.

We collected 10 different versions performing identical music pieces, which allows us to derive 10 sets of bowing and fingering arrangements, as well as pseudolabel (i.e., the skeleton motion data extracted from pose estimation method) for each music piece. The multi-version design of the dataset is incorporated with our data splitting strategy to explore diverse possible motion patterns corresponding to identical music piece. The skeleton and music data are available at the project link (see Section 6).

3.2 Audio feature extraction

We apply *librosa*, a Python library for music signal processing [21], to extract audio features. Each music track is sampled at 44.1 kHz, and the short-time Fourier transform (STFT) is performed with a sliding window (length = 4096 samples; hop size = 1/30 secs). Audio features are then extracted from STFT, including 13-D Mel-Frequency Cepstral Coefficients (MFCC), logarithm mean energy (a representation for sound volume), and their first-order temporal derivative, resulting in a feature dimension of 28.

3.3 Skeletal keypoints extraction

The state-of-the-art pose detection method [22] is adopted to extract the 3-D position of violinists’ 15 body joints, resulting in a 45-D body joint vector for each time frame. The 15 body joints are: head, nose, thorax, spine, right shoulder, left shoulder, right elbow, left elbow, right wrist, left wrist, hip, right hip, left hip, right knee, and left knee. The joints are extracted frame-wisely at the video’s frame rate (30 fps). All the joint data are normalized, such that the mean of all joints over all time instances is zero. The normalized joint data are then smoothed over each joint using a median filter (window size = 5 frames).

3.4 Data pre-processing

The extracted audio and skeletal data are synchronized with each other with the frame rate of 30 fps. To facilitate the training process, the input data are divided into segments according to the basic metrical unit in music. Beat position serves as the reference to slice data segments, considering the fact that the arrangement of bowing stroke is highly related to the metrical position. To obtain beat labels from audio recordings, we first derive beat positions in the MIDI file for each musical piece, and the dynamic time warping (DTW) algorithm is applied to align beat positions between the MIDI-synthesized audio and the recorded audio performed by human violinists. The beat positions are then used for the data segmentation. Each data segment starts from a beat position, and is with the length of 900, i.e., 30 seconds. According to the average tempo in the dataset, 30 seconds is slightly longer than 16 bars in music, which provides a sufficient context for our task. All the segmented data are normalized in feature dimension by z-score.

For the data splitting, a leave-one-piece-out (i.e., 14-fold cross-validation) scheme is performed by assigning 14 pieces to the testing set by turns. We take the recordings of one specific violinist for training and validation, and take the recordings of the remaining nine violinists for testing. For the training and validation data, we choose the recordings played by the violinist whose performance technique is the best among all according to expert’s opinion. Within the training and validation set, 80% of the sequences are for training and
20% of the sequences are for testing. This 14-fold cross-validation procedure results in 14 models. Each model is evaluated on the piece performed by the remaining nine violinists in the testing set. The results will then be discussed by comparing the nine different performance versions and their corresponding ground truths. This evaluation procedure can reflect the nature of violin performance, in which multiple possible motion patterns may correspond to a identical music piece in different musician’s recordings.

For the cross-dataset evaluation, we also evaluate our model using the URMP dataset [15], which has been used in previous studies for music-to-body-movement generation [16, 18]. The URMP dataset comprises 43 music performance videos with individual annotations for music-to-body-movement generation [16, 18]. The URMP dataset is constructed by an LSTM RNN, followed by a linear projection layer. The final output of the model is the superposition of the generated body skeleton and the right-hand skeleton to output the whole estimated full skeleton

\[ \mathbf{Y}_{\text{full}} = \mathbf{Y}_{\text{body}} + \mathbf{Y}_{\text{rh}} \]

where the scalar weights. Accordingly, each state perceives the global information, with the down-sampling convolutional layer in the encoder part of the U-net structure starts from an average pooling layer to downsample the full sequence, and is followed by two consecutive convolutional blocks, each of which consists of one convolutional layer, a batch normalization layer, and ReLU activation. Such computation repeats by \( N \) times until the bottleneck layer of U-net. In this paper, we set \( N = 4 \). The main function of the encoding process of the U-net is to extract high-level features from low-level ones; in our scenario, it functions as a procedure to learn structural features from the frame-level audio features. The self-attention layer between the encoding and decoding parts of the U-net will be introduced in the next section. Next, the encoder part of the U-net starts from an upsampling layer using linear interpolation, which is concatenated with the down-sampling convolutional layer in the encoder part through the skip-connection, and then followed by two convolutional blocks. This calculation also repeats by \( N \) times until the features are converted into another modal. Compared to the original architecture of U-net, we do not directly transform audio features to skeleton; rather, we first convert such output representation into another high-dimensional feature, which is levered for generation task with the remaining LSTM network. Moreover, we find out that the bowing attack accuracy can be improved by stacking multiple blocks in U-net with self-attention. The whole block is framed by dash line, as illustrated in Figure 3.

\[ \mathbf{y}^{(i)} \leftarrow \mathbf{y}^{(i)} + \mathbf{W}_r \mathbf{y}^{(i)} \]

### 4.2 Self-attention

Music typically has a long-term hierarchical structure. Similar patterns may appear repeatedly in a training sample. To decode the order of the body movement sequence is a critical issue. However, the U-net utilizes convolutional blocks in downstream to encode audio features to symbolic representation, it merely deals with the local structure in a limited kernel size. To solve the problem of long-term sequential inference, recently, the self-attention mechanism [28] has been widely applied in sequence-to-sequence tasks, such as machine translation, and text-to-speech synthesis. Different from the RNN-based models, in Transformer, representation is calculated by the weighted sum of each frame of the input sequence across different states, and the more relevant states are given more weights. Accordingly, each state perceives the global information, and this would be helpful for modeling long-term information such as notes and music structure. We therefore apply the self-attention mechanism at the bottleneck layer of U-net.

**Scaled Dot-Product Attention** Given input sequence \( X \in \mathbb{R}^{L \times d} \), we first project \( X \) into three matrices, namely query \( Q := X W_Q, \) key \( K := X W_K \) and value \( V := X W_V \), where \( W_Q, W_K, W_V \in \mathbb{R}^{d \times d} \). The scaled dot-product attention computes outputs for a sequence vector \( X \) as:

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{Q K^T}{\sqrt{d}} \right) V, \quad (1)
\]

where the scalar \( \frac{1}{\sqrt{d}} \) is used to avoid overflowed value leading to very small gradient.

**Multi-Head Attention** Multi-head attention allows the model to jointly attend to the information from different representation subspaces at different positions. The scale-dot product is computed

\[
\text{Attention}(Q, K, V) = \sum_{i=1}^{H} \text{Attention}^i(Q, K, V),
\]

where \( H \) is the number of heads.

**Transformer** The overall architecture of the Transformer is shown in Figure 3. It is different from RNNs in that the order of the body movement sequence is a critical issue. However, the U-net utilizes convolutional blocks in downstream to encode audio features to symbolic representation, it merely deals with the local structure in a limited kernel size. To solve the problem of long-term sequential inference, recently, the self-attention mechanism [28] has been widely applied in sequence-to-sequence tasks, such as machine translation, and text-to-speech synthesis. Different from the RNN-based models, in Transformer, representation is calculated by the weighted sum of each frame of the input sequence across different states, and the more relevant states are given more weights. Accordingly, each state perceives the global information, and this would be helpful for modeling long-term information such as notes and music structure. We therefore apply the self-attention mechanism at the bottleneck layer of U-net.
times in parallel with different head, and the th head can be expressed as follows:

\[
\text{Head}_h(Q_h, K_h, V_h) = \text{softmax} \left( \frac{Q_hK_h^T}{\sqrt{d_h}} \right) V_h. \tag{2}
\]

For each head, queries, keys, and values are projected into a subspace with dimension \(d_h\), where \(d_h = d/h\), \(W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d \times d_h}\) and \(Q_h, K_h, V_h \in \mathbb{R}^{L \times d_h}\). The output of each head are concatenated and linearly projected, and skip connection is applied with input \(X:\)

\[
\text{MultiHead} = \text{Concat}(\text{Head}_1, ..., \text{Head}_h) W^O, \tag{3}
\]

\[
\text{MidLayer} = \text{MultiHead} + X, \tag{4}
\]

where \(W^O \in \mathbb{R}^{(k \times d_h) \times d}\).

**Relative Position Representations** While there is no any positional information applied in scaled dot product, the same input at different time steps would contribute to the same attending weights. To solve the problem, we apply the relative position encoding [25] in the scaled dot-product self-attention. Two learnable embeddings \(R^K\) and \(R^V\) represent the distance between two positions in sequence vector \(X\), where \(R^K, R^V \in \mathbb{R}^{L \times d}\), and they are shared across all attention heads. We then modify Equation 1 as follows:

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QR^K^T + Q(R^K)^T}{\sqrt{d}} \right) (V + R^V). \tag{5}
\]

By adding the term \(Q(R^K)^T\) in numerator, the original matrix multiplication in Equation 1 would be injected the relative position information. The similar way is also applied to the value term, \(V + R^V\).

**Position-wise Feed Forward Network** Another sub-layer in the self-attention block is position-wise a feed-forward network. It consists in two linear transformation layers with a ReLU activation between them, which is applied to each position separately and identically. The dimensionality of input and output is \(d\), and the inner layer has the dimensionality of \(d_{ff}\). The outputs of this sub-layer are computed as:

\[
\text{FFN}(X) = \text{max}(0, XW_1 + b_1)W_2 + b_2, \tag{6}
\]

where the weights \(W_1 \in \mathbb{R}^{d \times d_{ff}}, W_2 \in \mathbb{R}^{d_{ff} \times d}\) and biases \(b_1 \in \mathbb{R}^{d_{ff}}, b_2 \in \mathbb{R}^{d}\).

Additionally, we also place an extra position-wise feed forward network after the last layer of U-net. While the outputs of U-net is contextualized, the position-wise feed forward network make it more similar to the skeletal representation.

### 4.3 Generation

For the generated body sequence \(\hat{Y}^{(body)}\), we directly feed audio features into the LSTM RNN network, followed by a dropout and a linear projection layer, as shown in the lower branch of Figure 3. This branch of model directly generates the sequence of 39-D body skeleton \(\hat{Y}^{(body)}\). For the right-hand sequence generation \(\hat{Y}^{(rh)}\), the output of position-wise feed forward network would be fed into two components. The first one is identical to the body sequence generation network, and the second component is a network to refine the right-hand motion. While directly producing full right-hand from one branch may limits the variation of wrist joint, we take the output of position-wise feed-forward network into another branch to generate the 3-D coordinate for the right-hand wrist joint. Therefore, the right-hand output is a 6-D right-hand sequence, whose last three dimensions (represented as wrist joint) are added by the output of refine network. Finally, we concatenate the outputs of body sequence and right-hand sequence:

\[
\hat{Y}^{(full)} = \text{Concat}(\hat{Y}^{(body)}, \hat{Y}^{(rh)}). \tag{7}
\]
The model is optimized by minimizing the $L_1$ distance between the generated skeleton $\hat{Y}(\text{full})$ and the ground truth skeleton $Y(\text{full})$: 

$$L_{\text{full}} := \|Y(\text{full}) - \hat{Y}(\text{full})\|.$$ 

4.4 Implementation details

In our experiments, we use 4 convolutional blocks ($N = 4$) in the downstream and upstream subnetworks of U-net individually, and all the dimensions of convolutional layers in U-net are set to 512. In the bottleneck layer of U-net, we adopt 1 attention block. The number of head in the attention block is 4, and $d$ is set to 512. The inner dimension of feed forward network $d_{ff}$ is set to 2048. The dimension of the LSTM unit is 512, and the dropout rate for all dropout layers is 0.1. Besides, we further stack two full components ($M = 2$) composed of U-net and self-attention as our final network architecture.

The model is optimized by Adam with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$, and adaptive learning rate is adopted over the training progress:

$$lr = k \cdot d^{-0.5} \cdot \min(n^{-0.5}, n \cdot \text{warmup}^{-1.5}),$$

where $n$ is the step number, $k$ is a tunable scalar. The learning rate is increased linearly for the first warmup training steps and is decreased thereafter proportionally to the inverse square root of the step number. We set $\text{warmup} = 500, k = 1$ for training model with 100 epochs, and batch size is set to 32. Furthermore, we use the early-stopping scheme to choose optimal model when the validation loss stops decreasing for 5 epochs.

5 EXPERIMENT

5.1 Baselines

We compare our method with two baseline models, which share $\epsilon$ as the discriminator for comparison.

Audio to body dynamics Both our work and [26] aim to generate body movement in music performance. [26] predicts reasonable playing movement based on piano and violin audio. Their model consists of 1-layer LSTM layer with time delay and 1-layer fully connected layer with dropout. We follow their setup and use MFCC and the ground truth skeleton as the pseudo-labels, whereas we extracts 15 body joints in 3-D space.

Speech to gesture Another work in [8] aims to predict the speaker’s gesture based on the input speech audio signal. Compared to our task, their predicted gesture motion are short segments ranging from 4-12 seconds, while our music pieces generally range from one to ten minutes. Convolutional U-net architecture is applied to their work, and a motion discriminator is introduced to eliminate single motion output. Although applying discriminator may increase the distance between the generated motion and ground truth (i.e., $L_1$ loss), the model is capable of producing more realistic motion. In this paper, we only take their model without discriminator as the baseline for comparison.

5.2 Evaluation metrics

There is no standard way to measure the performance of a body movement generation system so far. To provide a comprehensive comparison among different methods, we propose a rich set of quantitative metrics to measure the overall distance between the skeletons and also the accuracy of bowing attack inference.

$L_1$ and PCK While $L_1$ distance is the objective function in the training stage, we also used it to evaluate the difference between generated motion and ground truth. Note that we report the results by averaging over 45-D joints and across all frames. Considering that the motion in right-hand wrist is much larger compared to other body joints, we calculate another $L_1$ hand loss for the 3-D wrist joint alone. The Percentage of Correct Keypoints (PCK) was applied to evaluate the generated gesture in speech in a prior work [8], and we adapt PCK to 3-D coordinate in this paper. In the computation of PCK, a predicted keypoint is defined as correct if it falls within $\alpha \times \max(h, w, d)$ pixels of the ground truth keypoint, where $h, w$ and $d$ are the height, width and depth of the person bounding box, and we average the results using $a = 0.1$ and $a = 0.2$ as the final PCK score.

Bowing attack accuracy The bowing attack indicates the time slot when the direction of the hand is changing. We first take both right-hand wrist joint sequences having length $L$ as $\hat{y}^{(r,w)}$ and $y^{(r,w)}$. We consider that the tolerance should be more strict and only calculated once, which means that if all real bowing attacks are set the ground truth as 1 in the range $[i - \delta, i + \delta]$ for a bowing attack located at time $i$, which suggests that the predicted bowing attack $\hat{A}(i)$ is a true positive (i.e., correct) prediction, if real bowing attack is located on the range $[i - \delta, i + \delta]$. Otherwise, it would be a false prediction. Notice that all the ground truth bowing attacks are only calculated once, which means that if all real bowing attacks near $\hat{A}(i)$ have been calculated before, then $\hat{A}(i)$ is a false prediction. Another previous work [18] also introduced bowing attack accuracy as an evaluation metric, and set the tolerance value as $\delta = 10$ (i.e., 0.333s). We consider that the tolerance should be more strict and set $\delta = 3$ (i.e., 0.1s) in this paper. The F1-scores for bowing attack labels on axes $x, y, z$ (width, height and depth) are calculated, and represented as Bowx, Bowy and Bowz, respectively, whereas the average of these bowing attack accuracy is shown as bow avg.

Cosine similarity In this paper, our goal is not to generate identical playing movement as the ground truth, and the cosine similarity can be used to evaluate the similarity between generated skeleton and ground truth. The Cosine similarity is defined as:

$$\text{Cosine similarity} = \frac{\sum(Y_{\text{full}} \cdot \hat{Y}_{\text{full}})}{\|Y_{\text{full}}\| \cdot \|\hat{Y}_{\text{full}}\|}.$$
We compare our proposed method with two baselines [26] [8] for Table 2: The comparison between baselines and our proposed model in different evaluation metrics.

| Method                  | L1 avg | L1 hand avg | PCK  | Bowx | Bowy | Bowz | Bow avg | Cosine Similarity |
|-------------------------|--------|-------------|------|------|------|------|---------|------------------|
| A2B [26]                | 0.0391 | 0.0925      | 0.7240| 0.4169| 0.4462| 0.4062| 0.4231  | 0.6865           |
| S2G [8]                 | 0.0365 | 0.0910      | **0.7590** | 0.3800| 0.4330| 0.3729| 0.3953  | 0.6740           |
| Unet1                   | 0.0382 | 0.0870      | 0.7354| 0.4350| 0.4773| 0.4130| 0.4417  | 0.6850           |
| Unet1 + FFN + Refine Network | 0.0377 | **0.0840**  | 0.7416| 0.4427| 0.4870| **0.4157** | 0.4485  | 0.6934           |
| Unet2 + FFN + Refine Network | 0.0379 | 0.0860  | 0.7394| **0.4476** | **0.5165** | 0.4080 | **0.4574** | **0.6968** |

Table 2: The comparison between baselines and our proposed model in the URMP dataset.

|        | Bowx | Bowy | Bowz | Bow avg |
|--------|------|------|------|---------|
| A2B [26] | 0.4554 | 0.4775 | 0.4448 | 0.4893 |
| S2G [8]  | 0.3985 | 0.4660 | 0.4030 | 0.4225 |
| Our     | 0.4827 | 0.6286 | 0.4160 | **0.5090** |

similarity is therefore a suitable measurement to evaluate the general trend of bowing movement. We compute cosine similarity for 3-D right-hand wrist joint between generated results and the ground truth, and then take the average over three coordinates across all frames.

It should be noted that the above evaluation metrics cannot measure the absolute performance of body movement generation, since there is no standard and unique ground truth. Instead, the above evaluation metrics are to measure the consistency between the generated results and a version of human performance.

5.3 Quantitative results

We compare our proposed method with two baselines [26] [8] for the average performance over 14-fold test (as shown in Table 1). Three variants of the proposed methods are presented: First, Unet1 represents the model with one single block (i.e. M = 1, see Figure 3) composed by U-net with self-attention. Second, Unet1 + FFN + Refine Network adds a position-wise feed forward network and a refine network after Unet1. The last one, Unet2 + FFN + Refine Network, adopts two U-net blocks (M = 2) instead. The values reported in Table 1 can be understood based on a reference measurement: the mean of right arm length in our dataset is 0.13. For example, an L1 average value at 0.0391 mean that the average L1 distance between ground truth and prediction is around 0.0391/0.13≈30% of the length of right arm. In addition, it should be noted that L1 avg are generally smaller than L1 hand avg, which is owing to the fact that joints on trunk mostly move with small quantity, whereas right-hand wrist exhibit obvious bowing motion covering wide moving range.

It can be observed from the table that our model outperforms A2B both in L1 and PCK, which indicates that our model applying U-net based network with self-attention mechanism can improve the performance for learning ground truth movement. Although S2G has competent performance for L1 avg and PCK, our model boosts bowing attack accuracy more than 4% compared to S2G. Also, after adding the position-wise feed forward network and the refined network, we get better performance in L1 hand, bowing attack x, y, z and cosine similarity. This proves that the two components play a critical role for learning hand movement. Further, stacking two U-net blocks can increase bowing attack accuracy about 1%. Overall, stacking two blocks of U-net and adding two components can achieve the best results in most metrics. This best model outperforms the baseline A2B model significantly in a two-tailed t-test (p = 8.21 × 10^-8, d.f. = 250).

5.4 Cross-dataset evaluation

To explore if the methodology and designed process can adapt to other scenarios (e.g., different numbers of recorded joints, different positions such as standing or sitting, etc.), a cross-dataset evaluation is performed on the URMP dataset. The same process mentioned in Section 3 is applied to extract audio features and to estimate violinists’ skeleton motion. However, the URMP dataset only contains 13 body joints, whereas 15 joints are extracted from our recorded videos. Considering the different skeleton layouts between two dataset, only the averaged bowing attack accuracy, and the accuracy on three directions are computed as illustrated in Table 2. Our method (i.e. Unet2 + FFN + Refine Network) in Table 2 represents the best model demonstrated in the quantitative results. It can be observed that our proposed method outperforms two baselines for bowing attack accuracy, and it is demonstrated that our model is well-adapted to different scenarios and datasets.

5.5 Robustness test

To test the robustness of our model to tempo variation, we compare the average bowing attack F1-scores on the same music pieces with different tempi. It is expected that the performance of a robust model should be invariant with diverse tempi. It should be noted that only the longest piece in the dataset is tested in this experiment, and all results shown in Table 3 are the Bow avg values only. As shown in Table 3, our proposed model achieves better results compared to two baselines in five settings of tempo, which verifies the robustness of the proposed method with different performance tempi. The bowing attack accuracy is more likely to improve with faster tempo, since the prediction has a better chance to fall between the range \([i - \delta, i + \delta]\).
5.6 Subjective evaluation
Since A2B shows better performance than S2G, we take only the ground truth, A2B, and our model for comparison in the subjective evaluation. The material for evaluation is 14 performances played by one randomly selected violinist (length of each performance = 94 seconds). The ground truth, the generated movements by A2B and our model are presented in a random order for per music piece. And the participants are asked to rank ‘the similarity level compared to human being’s playing’ and the ‘rationality of the movement arrangement’ among the three versions. Within 36 participants in this evaluation, 38.9% of participants have played violin, and 41.7% have gotten music education or worked on music-related job. The results for all participants are shown in Figure 4, whereas the results specifically for the participants who have played violin are shown in Figure 5.

For the rationality, our results and the ground truth are much more reasonable than A2B, and the difference is more evident in Figure 5 compared to Figure 4. For the extent of being like human, the results are quite similar to the results of rationality in Figure 5, whereas no obvious trend is observed in Figure 4. This result may be owing to the limitation that only the violinist’s skeleton is included in the evaluation. In the future work, we consider to incorporate the violin bow as a part of our architecture to generate more vivid animations.

5.7 Qualitative results
For a more comprehensive demonstration of our result, we illustrate one example of the generated skeletons of the proposed method, the baseline method, and ground truth, as shown in Figure 6. In this example, we choose one bar from one of the music pieces in the testing data and show the corresponding skeletons. Figure 6 clearly shows that the movements of ground truth skeletons are consistent to the down-bow and up-bow marks in the score. It can be observed that the skeletons generated by the proposed model also exhibit consistent bowing direction in the right hand, while the skeletons generated by A2B do not show any changes of the bowing direction within this music segment.

6 CONCLUSION
In this paper, we have demonstrated a novel method for music-to-body movement generation in 3-D space. Different from previous studies, which merely apply conventional recurrent neural networks on this task, our model incorporates the U-net with self-attention mechanism to enrich the expressivity of skeleton motion. Also, we design a refinement network specifically for the right wrist to generate more reasonable bowing movements. Overall, our proposed model achieves promising results compared to baselines in quantitative evaluation, and the generated body movement sequences are perceived as reasonable arrangement in subjective evaluation, especially for participants with music expertise. Codes, data, and related materials are available at the project link.1

ACKNOWLEDGMENTS
This work is supported by the Automatic Music Concert Animation (AMCA) project funded by the Institute of Information Science, Academia Sinica, Taiwan. The authors would also like to thank Yu-Fen Huang for editing this paper and discussions.

1https://github.com/hsuankai/Temporally-Guided-Music-to-Body-Movement-Generation
REFERENCES

[1] Tamara Berg, Dehaleena Chattopadhyay, Margaret Schedel, and Timothy Vallier. 2012. Interactive music: Human motion initiated music generation using skeletal tracking by kinect. In Proc. Conf. Soc. Electro-Acoustic Music United States.

[2] Birgitta Burger, Sari Saarikallio, Geoff Luck, Marc R. Thompson, and Petri Toivainen. 2013. Relationships Between Perceived Emotions in Music and Music-induced Movement. Music Perception: An Interdisciplinary Journal 30, 5 (2013), 517–533.

[3] Lei Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. 2017. Deep cross-modal audio-visual generation. In Proc. Thematic Workshops of ACM MM. 349–357.

[4] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip reading sentences in the wild. In IEEE Conference on Computer Vision and Pattern Recognition. 3444–3453.

[5] Sofia Dahl, Frédéric Bevilacqua, and Roberto Bresin. 2010. Gestures in performance. In musical Gestures: Routledge. 46–88.

[6] Jane W Davidson. 2012. Bodily movement and facial actions in expressive musical performance by solo and duo instrumentalists: Two distinctive case studies. Psychology of Music 40, 5 (2012), 595–633.

[7] Anne Farber and Lisa Parker. 1987. Discovering music through Dalcroze eurhythmics. Music Educators Journal 74, 3 (1987), 43–45.

[8] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. In IEEE Conference on Computer Vision and Pattern Recognition. 3497–3506.

[9] Egal Haga. 2008. Correspondences between music and body movement. Ph.D. Dissertation. Faculty of Humanities, University of Oslo Unipub.

[10] Yu-Fen Huang, Tsung-Ping Chen, Nikki Moran, Simon Coleman, and Li Su. 2019. Identifying Expressive Semantics in Orchestral Conducting Kinematics. In International Society of Music Information Retrieval Conference. 115–122.

[11] Yu-Fen Huang, Simon Coleman, Eric Barnhill, Raymond MacDonald, and Nikki Moran. 2017. How do conductors’ movements communicate compositional features and interpretational intentions? Psychomusicology: Music, Mind, and Brain 27, 3 (2017), 148.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In IEEE Conference on Computer Vision and Pattern Recognition. 1125–1134.

[13] Ryo Kikutsumi, Kosetsu Tsukuda, Satoru Fukayama, Naoya Iwamoto, Masataka Goto, and Shigeo Morishima. 2016. A choreographic authoring system for character dance animation reflecting a user’s preference. In ACM SIGGRAPH.

[14] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to Music. In Advances in Neural Information Processing Systems.

[15] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Chenliang Xu. 2018. Skeleton Plays Piano: Online Generation of Pianist Body Movements from MIDID Performance. In International Society of Music Information Retrieval Conference. 218–224.

[16] Bochen Li, Chenliang Xu, and Zhiyao Duan. 2017. Audiovisual source association for string ensembles through multi-modal vibrato analysis. Proc. Sound and Music Computing (2017).

[17] Jun-Wei Liu, Hung-Yi Lin, Yu-Fen Huang, Hsuan-Kai Kao, and Li Su. 2020. Body Movement Generation for Expressive Violin Performance Applying Neural Networks. In IEEE International Conference on Acoustics, Speech and Signal Processing. 3497–3501.

[18] Jennifer MacRitchie and Massimo Zacari. 2012. The intentions of piano touch. In Proc. Conf. Soc. Electro-Acoustic Music United States.

[19] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In Advances in neural information processing systems.

[20] Marcelo M. Wanderley, Bradley W. Vines, Neil Middleton, Cory McKay, and Wesley Hatch. 2005. The musical significance of clarinetists’ ancillary gestures: An exploration of the field. Journal of New Music Research 34, 1 (2005), 97–113.

[21] Yu-Te Wu, Berlin Chen, and Li Su. 2018. Automatic music transcription leveraging generalized cepstral features and deep learning. In IEEE International Conference on Acoustics, Speech and Signal Processing. 401–405.

[22] Jitendra Malik. 2019. Learning Individual Styles of Conversational Gesture. In Proc. Conf. Soc. Electro-Acoustic Music United States.

[23] Youngwoo Yoon, Woo-Ri Ko, Minso Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In International Conference on Robotics and Automation. 4303–4309.

[24] J. Hlavac. 2005. The musical significance of clarinetists’ ancillary gestures: An exploration of the field. Journal of New Music Research 34, 1 (2005), 97–113.

[25] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D relative position representations. In Advances in Neural Information Processing Systems.