2020 年臺灣國際科學展覽會
優勝作品專輯

作品編號　**190019**

參展科別　電腦科學與資訊工程

作品名稱　**Body Movement Generation for Expressive Violin Performance Applying Neural Networks**

得獎獎項　大會獎：三等獎

出國正選代表

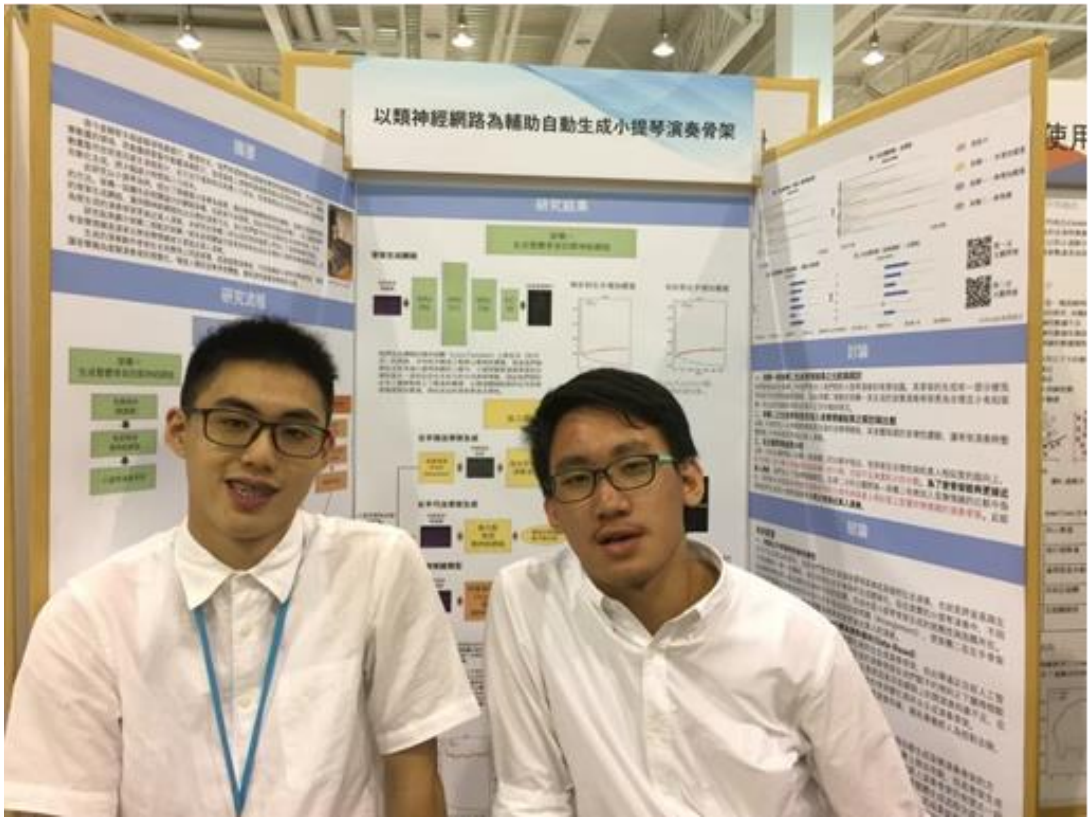
就讀學校　臺北市立建國高級中學

指導教師　蘇黎、王鼎中

作者姓名　劉峻瑋、林泓毅

關鍵詞　**Expressive Body Movement、Violin、Neural Networks**

作者簡介



我們是目前就讀建國中學三年級的林泓毅與劉峻瑋，很榮幸能夠參與此次的國際科展。這篇如果以後能夠刊出來，那代表我們應該是得獎了，因此我們想藉此特別感謝在科展路上曾經幫助過我的人。感謝中研院資科所音樂與文化科技實驗室的蘇黎教授一年半來的指導，讓兩個高中生擁有了一般研究生才能看見的視野與經驗，感謝資訊專題王鼎中老師的領導，感謝家人的支持與所有幫我們填問卷的人。

摘要

基於音樂輸入的動作骨架生成是一個正在興起的研究主題，然而在弦樂樂器的演奏骨架生成上，由於動作與音樂資訊間並非是一對一的對應關係，且在時間序列上非常注重前後關係，此問題仍非常具有挑戰性。在研究中，我們設計新的架構，將小提琴演奏者的演奏各部分拆解並分別生成。針對前人研究及此研究的研究結果，我們分別進行了客觀測試及主觀問卷的評估，兩方面皆顯示我們的研究結果較前研究進步。就我們所知，此篇研究是第一個嘗試在小提琴演奏動作上加入音樂情緒的研究。

ABSTRACT

Generating body movements based on given music audio recordings is an emerging research topic. This problem remains challenging particularly for string instruments, considering that the relationship between the musical note sequences and the body movement sequences in string instruments is not one-to-one correspondence and is highly contextual-dependent. In this paper, we take a divide-and-conquer approach to tackle the multifaceted characteristics of musical movement, and propose a framework for generating violinists' body movements. Both objective and subjective evaluation show that the proposed framework improves the stability as well as the perceptual quality of the generation outputs by using the task-specific models for bowing and expressive movement. To the best of our knowledge, this work represents the first attempt to generate violinists' body movements considering music expressivity.

1. INTRODUCTION

Musicians rely on their body movement to execute a music performance. The musical body movement is highly complex owing to 1) its diverse functions, 2) its context-dependent nature, 3) its person-dependent nature, and 4) the high-degree of freedom of movement during music performance. Musicians' body movement serves to produce the instrumental sound, to express their musical ideas, and to communicate with their co-performers [1]. Musicians intentionally select different movements to achieve the planed performance sound according to the musical compositional context [2]. For instance, different violinists may choose various bowing and fingering strategies depending on the musical interpretations they attempt to deliver. It has also been shown that individual musicians have their own distinguishable idiosyncratic movement features [3]. On the top of those, even when a particular musician playing the same musical composition, the performing movement may vary along with different expressive intentions [4]. All those dimensions contribute to the complexity and flexibility of musical movements.

Previous research has shown that different body movements from musicians generate diverse instrumental sound [5, 6]. And the correspondence between musicians' movement and the performed musical composition has also been demonstrated [7, 8]. An interesting yet challenging question is, in spite of its complexity, can the performing body movement being reversely generated from the given musical sound [9, 10]. The majority of existing studies tackle the movement-sound relation based on motion capture data collected in laboratory settings [11]. Such motion capture data have the virtue of being highly accurate and reliable, yet are with vast limitation in data collection and application. Furthermore, we suggest that the performance of existing end-to-end models to generate musicians' movement from

audio can be further improved by addressing the complexity and diverse functions of body movement [9, 10].

In this paper, the proposed model generates violinists' playing movement from the audio of violin performance. To tackle aforementioned issues, we take a divide-and-conquer approach to consider different functions of violinists' instrumental movement in their right hand (bowing model), left hand (position model), and expressive movement in upper body (expression model). Our data were derived from video sequences with pose estimation tools [12], which are more applicable to a nature scenario of music performance, and can potentially access to a crowd of data from existing video recordings. The model of generating body movement from music audio has the potential to be applied to computer graphics and animation, such as to synthesize virtual musicians directly from the recorded audio sequences.

2. Related Works

The task of modelling human body movement has been addressed by diverse approaches. Models to predict human movement (e.g. walking, running) has been built using frameworks including Gaussian processes [13], restricted Boltzmann machines [14], and hidden Markov models [15]. Deep learning frameworks such as Convolutional Neural Networks (CNN) [16] and Recurrent Neural Networks (RNN) [17] has proven to be efficient to generate movement sequences. Deep quaternion networks provide further improvement to implement the hyper-structure between components [18].

Regarding the correspondence between audio and body movement, the relationship between the speech and body movement has been explored [19]. Several attempts have been devoted to generate music-related movement. In [20], choreographic movements are automatically assigned to music according to the user's

preference. Pianists' and violinists' body movements were generated from given audio recordings using an end-to-end RNN model [9]. The model generating pianists' movement using the combination of CNN and RNN incorporated the information of bar and beat positions in music, and the model was proven to be capable of learning the movement characteristics of each pianist [10].

3. Proposed Method

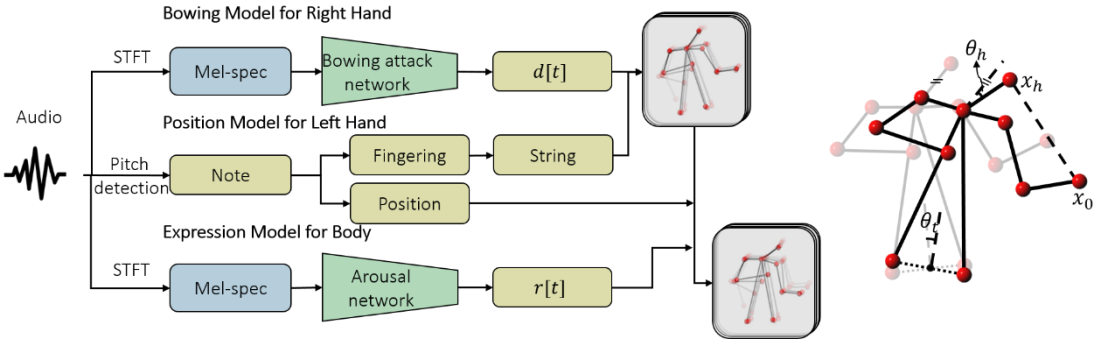


Fig. 1. Left: the body movement generation framework. Right: illustration of parameters (See Section 3.2 and 3.3)

Fig. 1 illustrates the proposed body movement generation framework for expressive violin performance. The framework consists of three models to deal with movements in different body parts, namely the *bowing model* for right hand, the *position model* for left hand, and the *expression model* for upper body. Due to the limitation of the training data and pose estimation accuracy, we consider only 2-D pose generation of the upper body, while depth information, lower body movement and fingering are not discussed. The framework takes audio as input, and outputs a 20-D sequence containing the 2-D location of 10 body joints: head, neck, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left hip, and right hip. The three models are discussed as follows.

3.1. Bowing model

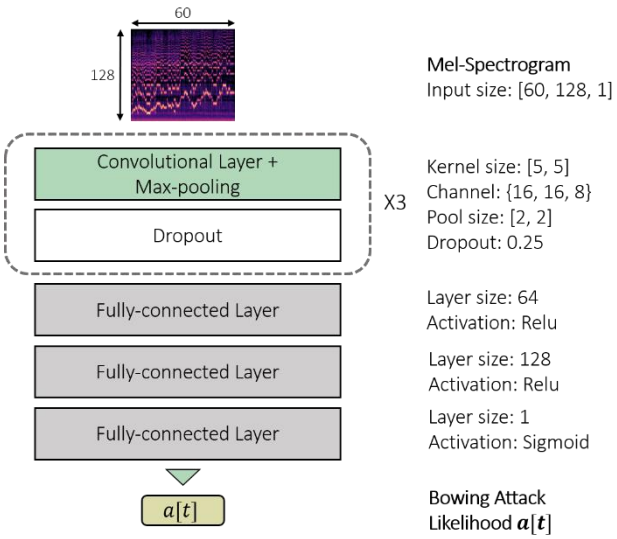


Fig. 2. The proposed network for bowing attack prediction.

In violin performance, there are usually different ways of bowing for violinists to perform the same music piece. Our goal is thus to generate a reasonable bowing arrangement according to a given music piece. To facilitate the discussion, we assume that the performance always starts from up-bow, as the same approach also applies for the music started with down-bow. In this way, the bowing movement generation problem is simplified into two subtasks: bowing attack prediction, and bowing speed rendering.

Fig. 2 shows the bowing attack prediction network. It is a CNN trained to predict *bowing attack*, the time instance when the bowing direction changes (i.e., from up-bow to down-bow. or from down-bow to up-bow). A 2-second segment of mel-spetrogram centered at time t is fed into the network, and it outputs the likelihood of bowing attack $a[t]$ at t, $a[t] \in \{0, 1\}$, where $a[t]$ is labeled as 1 if the bow is changing its direction at time t and labeled as 0 if not. The output layer is a sigmoid

function and an output threshold at 0.5 is used. The kth predicted bowing attack of a piece is denoted as a_k. The bowing direction $d[t] \in \{-1, 1\}$ is rendered from $a[t]$ by switching the sign of $d[t]$ when $a[t] = 1$, such that $d[t] = 1$ represents up-bow and $d[t] = -1$ represents down-bow.

Then we discuss body movement generation at the interval $t \in [a_k, a_{k+1}]$ between the kth and the $(k+1)$th predicted bowing attack for every k. Denote $\delta = a_{k+1} - a_k$. We first select the upper-body movements having a whole up-bow on each of the four strings from the training data as the up-bow *templates* $\mathrm{x}_s[q] \in \mathbb{R}^{20}$, where the subscript s represent the four strings; $s \in \{\mathrm{G}, \mathrm{D}, \mathrm{A}, \mathrm{E}\}$. The down-bow templates are made by reversing the up-bow templates. We normalize the length of these templates to $\delta_0 = 28$ frames (0.924s), i.e. $q \in [0, \delta_0]$. The body movement $\hat{\mathrm{x}}[t]$ corresponding to music is generated by *stretching* or *truncating* the templates:

$$\hat{\mathrm{x}}[t] \coloneqq \begin{cases} d[t]\mathrm{x}_s[(t - a_k)\delta_0/\delta], & \delta > \varphi; \\ d[t]\mathrm{x}_s[(t - a_k)], & \delta < \varphi; \end{cases}$$

where $\varphi_v = 15$ frames and s will be determined by the position model. More specifically, if the interval δ is longer than 15 frames, we assume it as full-bow and stretch the template; while if the interval is shorter then 15 frames, we truncate it.

3.2. Position model

The position of left hand on the fingerboard $\mathrm{x}_{lh}[t] \in \mathbb{R}^2$ depends on the contextual pitch sequence. We first assume that at $t = 0$ the left hand position started with the lowest position where the pitch is allowed to be played. Then, the pitch sequence estimated from the audio using the YAAPT algorithm [21] is applied to derive the left hand position of all notes. This is done with a greedy strategy: the left

hand position of the $(n+1)$th note is as near as possible to the position of the nth note. By assuming that the left hand always lies on the line between the lowest position $x_0[t]$ (i.e., 0th position on the fingerboard) and the head position $x_h[t]$, as illustrated in Fig. 1, the left hand at the n_pth position is synthesized as $\hat{x}_{lh}[t] = x_0[t] + n_p d(x_h[t] - x_0[t])$, where d is the ratio between the fingerboard position and the distance between head and the 0th position. After knowing n_p and pitch, the string number s can be uniquely determined and this information is then used to generate right hand bowing. Note that detailed left hand fingering generation is not implemented in this paper and is left as future work.

3.3. Expression model

Generating violinists' body movement with musical expression is a rather unexplored problem, probably because violin performance data containing annotations of music expression is rare. As a preliminary step toward this direction, we focus only on the *arousal* aspect of music expression, since [22] indicates that the music arousal level highly correlates to the head acceleration and torso tilt, and there are music datasets with arousal labels [23] for training an arousal prediction model.

The expression model is therefore an arousal-predicting network and a parametric model relating arousal to head and torso tilt. The arousal predicting network is a convolutional recurrent neural network (CRNN) based on [24]. It contains three convolutional layers of CNN with a receptive filter size of 3x3 and a fully-connected layer, followed by a bidirectional GRU and a maxout layer. MSE is used as the loss function. Based on the data annotation, the outputs arousal value is between -1 and 1. For better discussion, the final output arousal value is scaled to the range between 0 and 1.

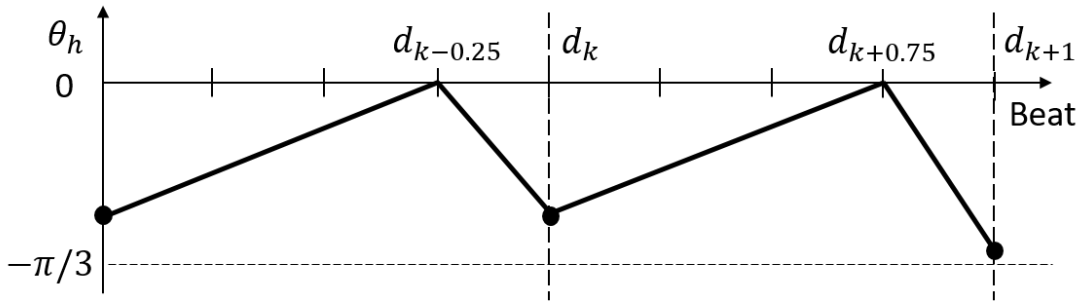


Fig. 3. Conceptual illustration of head movement angle (black solid bold line) for 4-beat music. The lowest angle (black dot) at every downbeat is determined by the average arousal value every the beat interval before the downbeat.

The predicted arousal $r[t] \in \{0, 1\}$ is then used to control the torso tilt angle $\theta_t[t]$ and the head tilt angle $\theta_h[t]$ (See Fig. 1). Taking the middle of the two hip joints as reference, the $\theta_t[t]$ measures the counterclockwise rotation of the neck joint from the vertical axis. Inspired from [22], high arousal implies large torso tilt, so we let $\theta_t[t]$ be proportional to $r[t]$: $\theta_t[t] \coloneqq -\beta_t r[t]$. We set $\beta_t = 2/\pi$, meaning that the maximal tilt is -90 degree. In comparison to torso, the behavior of head is more rhythmic, implying the need to incorporate beat information into the generation process. We define $\theta_t[t]$ as the extra tilt angle of head compared to the original template, as shown in Fig. 1. We adopt the `madmom` library [25] to estimate the beat and downbeat positions and design the rule of movement as: the head moves forward and backward periodically, and it moves forward to the lowest position (i.e. minimal θ_h) at downbeat. This forward-moving action is done within the beat interval before that downbeat, as shown in Fig. 3. Take 4-beat music as an example: given the arousal values between the kth downbeat d_k and its precedent beat $d_{k-0.25}$, the head tilt angle at d_k is:

$$\theta_h[d_k] = -\beta_h \frac{\sum_{t=d_k-025}^{d_k} r[t]}{d_k - d_{k-0.25}}$$

where $\beta_h = \pi/3$ is the maximal tilt angle for head. By assuming that $\theta_h[d_{k-0.25}] = 0$, the expressive head movement over time can be animated by interpolation.

4. Experiment

4.1. Data and preprocessing

The URMP dataset [26] is applied for body movement modeling. The dataset comprises 43 music performance video with individual instruments recorded in separate tracks. Among the tracks containing solo violin performance, we select 10 pieces (total length = 15m42s) as the training set, and 4 pieces (total length = 8m14s) for testing. The DEAM dataset [23] is applied to train our expression model. It consists of 1,802 music excerpt recordings with annotations of arousal and valence values per second. We use all the audio and arousal annotations in the dataset for training. For both datasets, audio signals are sampled at f_s = 22.05 kHz, and mel-spectrogram with dimension 128 is the input for all the three models. For the bowing and position models, the frame rate is 30 fps. For the expression model, the frame rate is set to 60 fps and the mel-spectrogram is pooled by averaging and standardization for every non-overlapped 500ms segments [24]. The list of used data and other details will be released afterwards.

The OpenPose library [12] and the COCO body model [27] are adopted to extract the 2-D position of the violinists' 10 upper body joints. The joints are extracted frame-wisely at the video's frame rate of 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 using a median filter with the window size of 5 frames.

Since the ground truth labels of bowing attack timing are not provided in the dataset, we retrieve them based on the right hand movement along the vertical axis.

More specifically, we compare the average velocities over 10 frames before and after every time instance t. Then, t is identified as a bowing attack if the two velocities are of opposite direction and their magnitude are both larger than a threshold. This is a simple yet efficient way to obtain bowing attack labels that facilitates model training and evaluation.

4.2. Objective evaluation

To evaluate how similar the bowing directions of the generated body movements are to the ground truth ones, we evaluate the performance of bowing direction and bowing attack on the four testing data. We re-implemented the end-to-end RNN model [9] for baseline comparison. First, we compare the predicted bowing direction $d[t]$ to the ground truth and report the frame-level accuracy. Results on the 4 testing pieces show that the average accuracy of the proposed model is 0.782, while the one of baseline method is 0.764. In the evaluation of bowing attack, we assume that a bowing attack is correctly predicted if the predicted time a_p is within a tolerance window $\varphi_e = 0.3$s from a ground truth a_l; $|a_p - a_l| \leq 0.3$. We report the F1-score, which is computed in the same way as in MIREX onset detection task. The F1-score of our method is 0.777, while the F1-score of the baseline is 0.760. In summary, the proposed model can better capture the behavior of bowing direction and bowing attack timing than the baseline model. Note that the objective evaluation does not reflect the overall quality of generation since the bowing movement is not uniquely determined by music.

4.3. Subjective evaluation

	Musician	Non-musician
Reasonableness	0.507±1.072	0.620±1.036
Naturalness	0.652±0.960	0.669±1.054

Table. 1. Mean (±standard deviation) scores of the subjective test. The score rates the improvement of the *reasonableness* and *naturalness* of the expressive body movement at the scale between -2 and 2. See Section 4.3 for details.

The subjective test is conducted to evaluate the effectiveness of the expression model. To do this, we compare the generated body movements with the expression model to the one without that model. In the test, every subject needs to watch the animation video showing the music and the generated movements. Every subject needs to answer two questions: 1) whether the movements with the expression model look more reasonable (i.e. the bowing attack and direction is compatible with the music), and 2) whether the movements with the expression model look more natural (i.e. more like human) than those without expression. We asked the subjects to score their answers in five levels: much more unreasonable/unnatural (-2); more unreasonable/unnatural (-1); no change (0); more reasonable/natural (1); and much more reasonable/natural (2).

Table 1 shows the mean and standard deviation of the scores rated by two groups of subjects: 185 subjects having no experience in professional music training, and 28 subjects having professional music training. On average, both groups of subjects rate the generated body movements with the expression model to be more reasonable and more natural than the one without the expression model. It could be note that the musician group still rates lower scores to both questions. Another interesting

phenomenon is that the two scores rated by the musicians differs 0.145 (0.507 and 0.652), while the scores rated by non-musicians are almost the same (0.620 and 0.669). This is probably because non-musicians are not able to catch the precise meaning of `reasonable' and `natural,' while musicians are more able to distinguish the two. This is an important lesson which suggests the need to revise the contents of our subjective tests in the future.

5. Conclusion

We have proposed and verified the use of bowing attack prediction, fingerboard position mapping and music arousal prediction to generate instrument and expressive body movement. Based on the domain knowledge of violin performance, the proposed framework performs more stable than the fully data-driven approach, and provides extra flexibility in controlling musical expression. Future work includes 3-D body movement and fingering generation, and the incorporation of more complicated expression such as musical valence.

6. References

[1] M. M. Wanderley, B. W. Vines, N. Middleton, C. McKay, and W. Hatch, “The musical significance of clarinetists’ ancillary gestures: An exploration of the field,” *Journal of New Music Research*, vol. 34, no. 1, pp. 97–113, 2005.

[2] J. MacRitchie, B. Buck, and N. J. Bailey, “Inferring musical structure through bodily gestures,” *Musicae Scientiae*, vol. 17, no. 1, pp. 86–108, 2013.

[3] H. F. Mitchell and R. MacDonald, “Listeners as spectators? audio-visual integration improves music performer identification,” *Psychology of Music*, vol. 42, no. 1, pp. 112–127, 2014.

[4] G. Castellano, M. Mortillaro, A. Camurri, G. Volpe, and K. Scherer, “Automated

analysis of body movement in emotionally expressive piano performances," *Music Perception: An Interdisciplinary Journal*, vol. 26, no. 2, pp. 103–119, 2008.

[5] S. Dahl, F. Bevilacqua, and R. Bresin, "Gestures in performance," in *musical Gestures*, pp. 48–80. Routledge, 2010.

[6] J. MacRitchie and M. Zicari, "The intentions of piano touch," in *12th ICMPC and 8th ESCOM*, 2012.

[7] E. Haga, "Correspondences between music and body movement," 2008.

[8] M. R Thompson and G. Luck, "Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music," *Musicae Scientiae*, vol. 16, no. 1, pp. 19–40, 2012.

[9] Eli Shlizerman, Lucio Dery, Hayden Schoen, and Ira Kemelmacher-Shlizerman, "Audio to body dynamics," in *IEEE CVPR*, 2018, pp. 7574–7583.

[10] B. Li, A. Maezawa, and Z. Duan, "Skeleton plays piano: Online generation of pianist body movements from midi performance.," in *ISMIR*, 2018, pp. 218–224

[11] B. Burger, M. R. Thompson, G. Luck, S. H. Saarikallio, and P. Toiviainen, "Hunting for the beat in the body: on period and phase locking in music-induced movement," *Frontiers in human neuroscience*, vol. 8, pp. 903, 2014.

[12] Z. Cao, G. Martinez, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE TPAMI*, 2019.

[13] J. M. Wang, D. J. Fleet, and A. Hertzmann, "Gaussian process dynamical models for human motion," *IEEE TPAMI*, vol. 30, no. 2, pp. 283–298, 2007.

[14] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton, "Dynamical binary latent variable models for 3D human pose tracking," in *IEEE CVPR*, 2010, pp. 631–638.

[15] S Toyer, A Cherian, T Han, and S Gould, "Human pose forecasting via deep markov models," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2017, pp. 1–8.

[16] D. Holden, J. Saito, and T. Komura, “A deep learning framework for character motion synthesis and editing,” *ACM TOG*, vol. 35, no. 4, pp. 138, 2016.

[17] J. Martinez, M. J. Black, and J. Romero, “On human motion prediction using recurrent neural networks,” in *IEEE CVPR*, 2017, pp. 2891–2900.

[18] C. J. Gaudet and A. S. Maida, “Deep quaternion networks,” in *IJCNN. IEEE*, 2018, pp. 1–8.

[19] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *IEEE CVPR*, 2017, pp. 3444–3453.

[20] R. Kakitsuka et al., “A choreographic authoring system for character dance animation reflecting a user’s preference,” in *ACM SIGGRAPH*, 2016.

[21] K. Kasi, Yet Another Algorithm for Pitch Tracking, Ph.D. thesis, Old Dominion University, 2002.

[22] B. Burger et al., “Relationships between perceived emotions in music and music-induced movement,” *Music Perception: An Interdisciplinary Journal*, vol. 30, pp. 517–533, 06 2013.

[23] M. Soleymani, A. Aljanaki, and Y.-H. Yang, “DEAM: Mediaeval database for emotional analysis in music,” 2016.

[24] M. Malik et al., “Stacked convolutional and recurrent neural networks for music emotion recognition,” *arXiv:1706.02292*, 2017.

[25] S. Bock, F. Korzeniowski, J. Schl ¨ uter, F. Krebs, and ¨ G. Widmer, “madmom: a new Python Audio and Music Signal Processing Library,” in *ACM MM*, 2016.

[26] B. Li et al., “Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications,” *IEEE TMM*, vol. 21, no. 2, 2018.

[27] T.-Y. Lin, M. Marie, S. Belongie, P. Perona, D. Ramanan, P. Dollar, and C. Zitnick, “Microsoft COCO: ´ Common objects in context,” 2014.

【評語】190019

- 本項研究的主題新穎有趣且完成度高。

- 對於作品的研究成果評估，除了質化的主觀評估外，可考慮同時提供較為客觀的量化結果。