IDENTIFYING EXPRESSIVE SEMANTICS IN ORCHESTRAL CONDUCTING KINEMATICS

Yu-Fen Huang[1] **Tsung-Ping Chen**[1] **Nikki Moran**[2] **Simon Coleman**[3] **Li Su**[1]
[1]Music and Culture Technology Lab, Institute of Information Science, Academia Sinica, Taiwan
[2]Reid School of Music, University of Edinburgh, UK
[3]Institute for Sport, Physical Education and Health Sciences, University of Edinburgh, UK
yfhuang@iis.sinica.edu.tw

ABSTRACT

Existing kinematic research on orchestral conducting movement contributes to beat-tracking and the delivery of performance dynamics. Methodologically, such movement cues have been treated as distinct, isolated events. Yet as practicing musicians and music pedagogues know, conductors' expressive instructions are highly flexible and dependent on the musical context. We seek to demonstrate an approach to search for effective descriptors to express musical features in conducting movement in a valid music context, and to extract complex expressive semantics from elementary conducting kinematic variations. This study therefore proposes a multi-task learning model to jointly identify dynamic, articulation, and phrasing cues from conducting kinematics. A professional conducting movement dataset is compiled using a high-resolution motion capture system. The ReliefF algorithm is applied to select significant features from conducting movement, and recurrent neural network (RNN) is implemented to identify multiple movement cues. The experimental results disclose key elements in conducting movement which communicate musical expressiveness; the results also highlight the advantage of multi-task learning in the complete musical context over single-task learning. To the best of our knowledge, this is the first attempt to use recurrent neural network to explore multiple semantic expressive cuing in conducting movement kinematics.

1. INTRODUCTION

During orchestral conducting, conductors use their body movements to guide musicians' expressions of various features such as the tempo, dynamics, and articulation in music. Through these delicate nuances in their body movement, conductors are able to communicate their refined interpretations and expressive intentions of the musical work in question. As documented in pedagogical literature, 'conducting semantics' - a codified repertoire of movements bearing specific musical instructions - are broadly accepted as core knowledge, fundamental to conducting practice. [11,15,21,31]. However, in the scenario of actual conducting performances, conductors' movement styles are remarkably diverse. Moreover, while it is conventional common sense for experienced musicians to identify distinct musical features (e.g. phrasing, dynamics, articulation), these features associate with one another in the complex musical context, and conductors tend to communicate similar musical features using diverse strategies depending on the musical context, e.g., dynamic and phrase cuing may vary when conducted with different articulation patterns. As a result, existing music-movement coupling models are limited in some aspects [16, 28, 33]. The straightforward association which presumes that certain movement features can communicate specific musical traits, as stated in conducting pedagogy, has not been observed in such models.

The major challenges to overcome for current conducting movement research are: 1) to identify interpretable quantitative descriptors to represent movement features; 2) to construct a generalisable model, which is robust in identifying features such as phrase, dynamic, and articulation cuing in complex musical context, and which is tolerant to the flexibility of conducting performance and the differences between individual conductors. In this study, we approach both of the challenges by adopting machine learning algorithms. More specifically, to determine potential movement descriptors relevant to expressive musical features, we first applied a supervised feature selection technique, ReliefF [17–19], to three-dimensional body movement data. Based on such selected features, we then sought to jointly identify different types of expressive cuing in conducting movement, and thus trained a recurrent neural network (RNN) on the body movement data from various conductors to perform multi-task learning (MTL).

This study identifies the effective descriptors in conducting movement which are used to communicate musical features. As the pioneering attempt to apply RNN to music-movement coupling in conducting, we also verify the advantage for RNN framework with MTL to probe potential complex connections between various movement and musical elements, especially compared to previous works using other models. In the subsequent section, re-

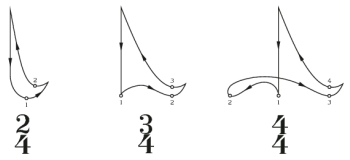

Figure 1. Basic metrical patterns in musical conducting. (reproduced from http://www.purposeful primarymusic.com).

lated research in musical conducting movement will be reviewed. Our dataset and model will be introduced in section 3 and 4 respectively. The experimental designs will be reported and the results will be discussed in section 5, followed by the conclusion in section 6.

2. RELATED WORK

For orchestral performance, a fundamental function of conducting is to coordinate musicians' playing. Conductors regulate the musical timing using basic beat patterns (e.g. two-beat, three-beat, or four-beat cycles etc., see Figure 1), and such beating movements communicate both the tempo arrangement and the metrical structure of the performed musical piece [11, 15, 21, 31]. The beating pattern in conducting has been substantially investigated, and it is found that the vertical action is the major component to lead the beat timing [4, 22, 24, 34, 35]. The tracking of beat is the most efficient under the constraint of movement bounding box area [35] and using dynamic time warping technique [34]. Based on these findings, interactive systems such as Pinocchio, vMaestro, and Personal Orchestra has built, in which the tempo of recorded orchestral audio is manipulated by the user's body movement, and the music automatically aligns with the beat timing identified in the movement [4, 22, 24].

Yet conducting movement is far more complicated than simple beating. Built on these basic metrical patterns, conductors take a step further to communicate their musical interpretations via refined variations in their movement. Pedagogical sources present a stock of movements frequently used by conductors to instruct their expressive intentions regarding articulation, dynamics, and phrasing [11, 15, 21, 31]. As summarised by the authors' previous work, specific conducting movements carrying expressive intentions, i.e. conducting semantics, are understood to comprise distinct combinations of hand position (high/ low/ away from the body/ close to the body), movement size (large/ small), speed (quick/ slow), acceleration (sudden/ gradual change of movement), smoothness (smooth/ jerky), trajectory shape (straight/ curved), and palm direction (upward/ downward/ facing musicians/ facing the conductor) [13, 14]. These qualitative, subjective descriptions of movement features match with quantitative, empirical analysis of conducting kinematics, in which the palm directions, movement size, hand positions and velocity reflect the dynamic change in music [8, 33, 35].

That conductors use their body movement to communicate musical expressiveness is self-evident, and is a truth demonstrated by both musical pedagogy and empirical analysis of conducting movement. However, existing music-movement coupling models generate only low to moderate correlations between musical and movement features [16, 33], and the computational approach which automatically classifies conductors' expressive intentions based on their movement has also produced unsatisfactory results [28]. These findings demonstrate that the expressive semantics in conducting are highly context-dependent. Different aspects of musical expression, such as dynamics, articulation, and phrasing are entangled in actual instances of musical performance, and their study would benefit from the ecologically-valid, intact musical contexts.

In music information retrieval (MIR) research, vast efforts have been devoted to extracting semantic representations such as pitch, beat, and harmony from audio signals. The machine learning approach has gained great success in modelling the semantics from music audio recordings [7]. During performance, musical sound is usually generated by the execution of body movement. There is a pressing need, therefore, for an equivalent effort to explore the expressive semantics carried by musical movement. The recurrent neural network (RNN) is widely used to analyse music semantics in a sequential way [25, 30, 32, 38]. RNN also allows the multi-task learning (MTL) setting, which has been proven successful in music information retrieval [12, 39]. We therefore propose a RNN framework with MTL approach in this study, and test the model on our conducting movement corpus.

3. DATA

This study is based on a professional conducting movement dataset collected by the authors, which contains conductors' upper body movement recorded by a high-resolution motion capture system, together with annotations of beat timing, phrase, dynamic, and articulation.

3.1 Collection of motion capture data

The motion capture data were recorded in the Biomechanical Laboratory at the Institute for Sport, Physical Education and Health Sciences (ISPEHS), University of Edinburgh, UK. Conductors' movement were collected using a nine-camera optical motion capture system (Qualisys, Pro-Reflex, Sweden) at a sampling frequency of 120 frames per second. The captured area was calibrated using the Qualisys 300 mm wand kit with the average residual being lower than 2 mm. Twenty-seven 12 mm optical markers were attached to the conductor's upper body and baton following the Golem Upper Body Model in Visual3D documentation [6]. The locations of 27 markers were illustrated in Figure 2 and listed in Table 1.

Six conductors (3 professional conductors and 3 advanced conducting students) participating the collection were all right-handed males, with an average conducting experience for 10.6 years (SD = 9.37), and conducted for 4.4 hours per week on average (SD = 2.38) at the time of participation. Five string musicians accustomed to musi-

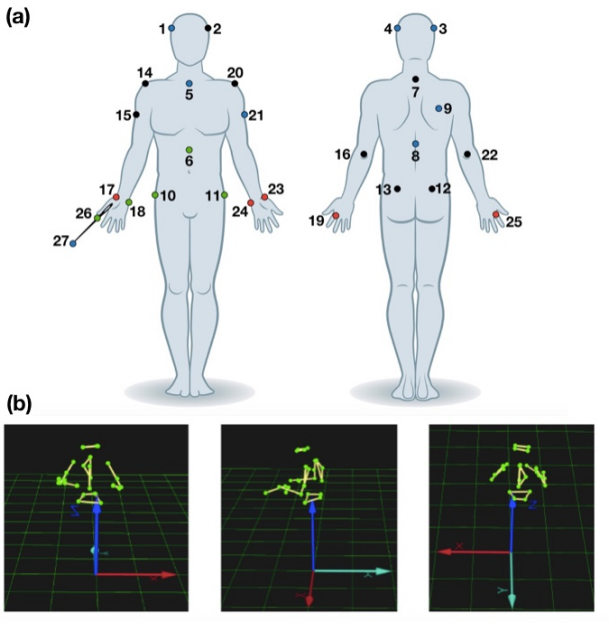

Figure 2. (a) The location of 27 optical markers placed on the conductor's upper body from the frontal and rear viewpoints. The marker colours represent the ranking from the feature selection procedure ReliefF: The top 15 elements (red), no 16-30 elements (green), no 31-45 elements (blue), no 46-81 elements (black). (b) The 27 markers captured by the system from the frontal, lateral, and rear viewpoints.

#	Name	Description
1	RFHD	Right temple
2	LFHD	Left temple
3	RBHD	Right back head
4	LBHD	Left back head
5	CLAV	Jugular Notch
6	STRN	Xiphoid process of the Sternum
7	C7	Spinous process of the 7th Cervical vertebrae
8	T10	Spinous process of the 10th thoracic vertebrae
9	RBAK	Middle of the right Scapula
10	RASI	Right Anterior Superior Iliac Spine
11	LASI	Left Anterior Superior Iliac Spine
12	RPSI	Right Posterior Superior Iliac Spine
13	LPSI	Left Posterior Superior Iliac Spine
14	RSHO	Right Acromio-clavicular joint
15	RUPA	Right upper arm
16	RELB	Right elbow joint
17	RWRA	Right wrist thumb side
18	RWRB	Right wrist pinkie side
19	RFIN	The 2nd Metacarpal of the right forefinger
20	LSHO	Left Acromio-clavicular joint
21	LUPA	Left upper arm
22	LELB	Left elbow joint
23	LWRA	Left wrist thumb side
24	LWRB	Left wrist pinkie side
25	RFIN	The 2nd Metacarpal of the right forefinger
26	BASH	Baton shaft
27	BAEN	Baton end

Table 1. The locations for 27 optical markers placed on the conductor's upper body

cal conductors' directions were recruited from University ensembles; the average number of years' instrumental experience was 15.6 (SD = 2.30) and average experience of playing in orchestra was 12.2 (SD = 3.11).

Each conductor rehearsed with musicians for 30 minutes. In the subsequent recording session, the conductor's movements were collected in a repeated measures design. The 6 conductors performed 3 different excerpts of Western classical orchestral repertoire: 1) W. A. Mozart, *Serenade in G major, K.525* (first movement, bars 1-55), 2) A. Dvořák, *Serenade in E Major, Op.22* (first movement, bars 1-53), and 3) B. Bartók, *Divertimento for String Orchestra, Sz. 133* (third movement, bars 1-183). The excerpts represent different compositional styles [5], yet share aspects of metrical structure, which lends them to comparison. Each excerpt is roughly 1 minute long according to constraints of the motion capture equipment. Individual conductors recorded 3 instances of each excerpt, with excerpt performance order counterbalanced across the 6 individuals. As a result, 54 conducting performances were collected in the corpus in total (3 performances x 3 musical excerpts x 6 conductors). The complete dataset was uploaded as the C3D file format for motion capture analysis to the DataShare open access data repository at the University of Edinburgh: https://datashare.is.ed.ac.uk/handle/10283/2906.

3.2 Data pre-processing and labelling

3.2.1 Pre-processing of motion capture data

The collected motion capture data were exported from Qualisys Tracker Manager (version 2.7, Pro-Reflex, Sweden) and imported to Visual3D (standard version 4.93, C-motion, USA) and Python (version 3.6.8) for further analysis. The original movement data containing the position on x-, y-, and z- axes of 27 markers were smoothed by the fourth-order low-pass Butterworth filter with a cut-off frequency of 10 Hz. The speed on x-, y-, z- axes of each marker was defined as the first derivative of x-, y-, z- position respectively, divided by the sampling interval (1/120 s). Speed was considered to carry important information in previous studies on musical conducting movement [13, 14, 26, 27], and thus was chosen as the variable to be investigated. The speed data were normalised based on the mean and standard deviation in each performance trial, and were converted to the z-scores of speed.

The z-scores of speed were then imported in the subsequent feature selection procedure ReliefF. Moreover, a generalisable element is preferable for RNN model. Considering that the minor fluctuations within 1/10 beats (roughly 5 frames) has only trivial effect on our target musical features – which are phrase, dynamic, and articula-

tion structure in bar level, and such minor fluctuations may, on the contrary, have negative effect on generalisation, the simple moving averages of speed data for per 5 frames were thus taken before being imported in RNN model.

In order to align motion capture data with expressive labels extracted from musical scores, the timing of beats and bars in conducting movement was assessed. The beat timing in movement was estimated using the motion capture analysis software Visual3D, and was defined as the time when the lowest position of baton tip on the z-axis (vertical axis) occurred within a beat period according to previous studies [13,14,26,27]. The initiation of a musical bar is then defined as the onset of the first beat in the given bar.

3.2.2 Annotation of expressive labels

The label for dynamics, articulation, and phrasing were annotated by three experienced music theorists (with an average experience for music analysis = 20.7 years, SD = 2.89) based on designated edition of musical scores [2, 9, 29]. Where the annotators disagreed with each other, the opinion of the majority was taken. The dynamic and articulation labels were appointed to each musical bar. Phrasing labels, on the other hand, indicate the initiation timing of phrases, which is dissimilar to dynamic and articulation status that prolongs for a period of time. The phrase initiation was thus determined as the timing of the first beat in each phrase, and then entered the model with a window size of +/- 60 frames, which is equivalent to +/- 0.5 seconds of the phrase initiation in the given sample frequency 120 fps, and roughly +/- 1 beat of the phrase initiation according to 120 bpm instructed in musical scores. The labels are illustrated in Figure 3. The annotation process complied with the following principles:

1) Dynamics: The dynamic annotation contains 6 classes including: *pp, p, mp, mf, f, ff* in the 3 music excerpts. Dynamic levels were labelled according to the expressive terminology in the scores. Where there is no dynamic instruction specified in the given bar, the dynamic level in the previous bar was taken. Where the dynamic level changes within a bar, the dynamic level for the majority of beats was taken. Where different dynamic levels exist in equal number of beats within a bar, the first dynamic level occurring in the given bar was taken. Where crescendo or diminuendo is marked, the gradual change of dynamics was divided into equal levels and was designated to the bars in the crescendo or diminuendo process.

2) Articulation: The articulation annotation contains 3 classes including: *legato, neutral, staccato*. The legato label was assigned to where the slur is printed; the staccato label was assigned to where the staccato mark is printed in the score, or where a note equals to or shorter than a quaver is followed by a rest; the neutral label was assigned to bars where no specific legato or staccato marks was printed. Where there are more than one type of articulation terms printed within a bar, the articulation type for the majority of beats was taken. Where different articulation types exist in equal number of beats within a bar, the first articulation type occurring in the given bar was taken.

3) Phrasing: The phrasing annotation specifies the initiation of phrases, and it contains 2 classes including: *the phrase onset, none*. The phrase structure was determined according to conventional music analysis techniques [3]. Where different melodies interlocked, the phrase in the main melody is taken.

4. DATA ANALYSIS AND MODEL

The aforementioned data set was analysed in two steps: 1) the supervised feature selection technique ReliefF was applied to identify effective descriptors in conducting movement to communicate musical expressiveness; 2) the RNN architecture with MTL setting were constructed to model generalisable rules to associate multiple movement and musical features.

4.1 Data representation

The input data is represented as the z-score of speed on x-, y-, z- axes of 27 markers (3 x 27 = 81 features) with the sample frequency of 120 fps. As shown in Figure 4, each segment fed into the bidirectional long-short-term memory (BLSTM) network contains 81 features with 121 frames (which is equivalent to roughly 1 second or 2 beats); the hop size for consecutive segments is 30 frames (0.25 seconds or 0.5 beats); each input clip contains 64 segments (16.75 seconds or roughly 32 beats; 8 bars in Mozart and Dvořák; 32 bars in Bartók). In addition, we also imported the top 15 movement elements selected by ReliefF (15 features x 121 frames per segment) into RNN, to examine if our recognition model is capable of identifying musical expressions from a small assemble of crucial features.

4.2 Feature selection for movement data

Orchestral conducting consists of movements from different body parts. Conducting pedagogy mainly describes the expressive guidance instructed by hand movements, yet no evidence has been provided by previous motion capture studies to verify such emphasis on hands. To this end, it is essential to know which parts of the body and what kinds of movements are relevant to the expressive intentions in music. The process of finding such relevance can be regarded as a feature selection problem. The ReliefF algorithm, one of the most commonly used supervised feature selection algorithms for music related movement [23], is applied here as an exploratory study. ReliefF is an extension from the original Relief [17–19] with higher reliability and is applicable to multi-class datasets. The algorithm searches every class of features x_i for their k-nearest neighbours from the same class (nearest hit) and from a different class (nearest miss) to score how well such feature to distinguish data from different classes [19]: $\text{Score}_i \sim -\sum_k d(x_i, x_{\text{NearHit}}) + \sum_k d(x_i, x_{\text{NearMiss}})$, where $d(\cdot,\cdot)$ is a distance measure. The exact form of ReliefF can be found in [23]. In this research, we set $k = 20$ as suggested by Urbanowicz *et al.* [37].

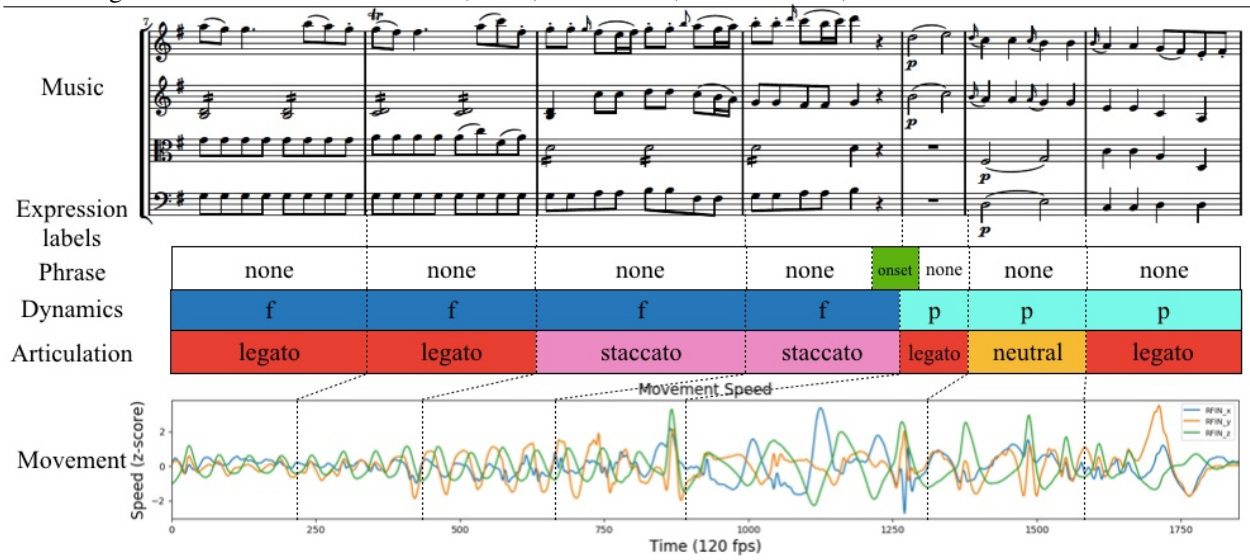

Figure 3. The example of musical scores, 3 types of expression label (phrase onset, dynamic level, articulation), and the movement data from right finger marker.

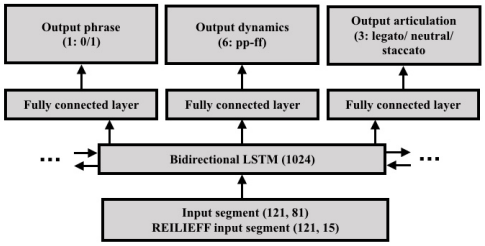

Figure 4. The multi-task model built on the RNN

4.3 Recognition model for music expression

We applied a recurrent neural network (RNN) with bidirectional long-short-term memory (BLSTM) cells to carry out multi-task learning (MTL) for expressive cuing in conducting movement. Such architecture is advantageous to determine sequential features in musical signals [7, 30, 32, 38], yet to the best of our knowledge, this approach has not been applied to musical movement in any existing study.

As shown in Figure 4, the model has a shared BLSTM layer with 1024 hidden units, and a task-specific fully-connected layer. The outputs from forward and backward LSTMs are concatenated as a 2-by-1024 matrix in the segment level, which is then flattened to link the fully-connected layer. The output layer is a 10-D vector containing the classes for the three tasks. Sigmoid is applied to output phrase initiation (1D); Softmax is applied to output the dynamic level (6D) and the articulation type (3D).

To examine the advantage of MTL, we also constructed single-task learning (STL) models, where the same RNN framework and BLSTM cells are applied to phrase, dynamics, and articulation tasks respectively.

5. EXPERIMENT

5.1 Experimental settings

The 54 conducting performance trials (810 input clips) are divided into training set (48 trials, 720-722 clips) and testing set (6 trials, 88-90 clips). The 9-fold cross-validation is performed in a way that per 6 trials are in turns assigned to the testing set in 9 experiments. Each clip fed into the BLSTM network contains 64 segments (# of feature x 121 frames per segment); the hop size for consecutive clips are 32 segments. To prevent the problem of over-fitting, data augmentation is performed in a way that random Gaussian noise is added to the original data with the level of 0.1 standard deviation of the input data.

All networks are implemented using TensorFlow [1], and are trained using stochastic gradient descent with Adam optimiser. The cross-entropy between output and labels are computed for training purpose. To avoid over-fitting, L2 regularisation is applied. The dropout rate for both the input and output of BLSTM cells is 0.7. The learning rate is 0.0001 with 4800 training steps. This procedure is performed on 4 models: the MTL model, STL models for phrase, dynamics, and articulation tasks respectively.

5.2 Evaluation metrics

For the phrase recognition task, The Precision, Recall, and F1 measures (the balanced harmonic mean of Precision and Recall) [10] are computed in segment level (hop size = 30 frames; 0.25 seconds; roughly 0.5 beats). A detected phrase cuing event is considered as a true positive if it lies within a tolerance window +/- 60 frames (0.5 second and roughly 1 beat) from the ground truth annotation. If there are two or more phrase cuing detected within this tolerance window, one of the detections is considered as a true positive, and others are considered as false positives. For dynamic and articulation recognition tasks, the accuracies (acc) are computed in segment level: acc = (# of true posi-

tive segments/ # of total segments).

5.3 Result

5.3.1 Results of movement feature selection

The top 15 relevant movement features selected by ReliefF algorithm to fit 3 types of musical features (phrase, dynamics, articulation) are illustrated in Figure 2. It appears that the elements from right and left fingers, and right and left wrists are prominent in communicating all three types of musical expressions. These top 15 elements are inputted in the subsequent RNN model.

It is an unexpected outcome that the elements from the baton end are not included in the top list. As suggested in conducting pedagogy, the baton end has been considered as an important reference to communicate expressive information in conducting [11, 15, 21, 31]. In the ReliefF analysis, the elements from baton end are ranked as no 31-33 within 81 elements. It could be the effect that our target musical features are in bar level rather than in beat level. According to eigenvalue and eigenvector analysis in previous research on musical movement, the movements in extremity usually correspond to lower-level musical elements with a shorter period (such as per beat), whereas movements in torso tend to associate with higher-level musical features with a longer period (such as per bar or longer) [36]. The right finger and wrist are in the intermediate location between the body and baton tip, and are the key body part for conductors to manipulate the baton, which could be the reason that why they contribute the most to our expressive targets.

5.3.2 Results of music expression recognition

The results of using RNN framework to perform MTL and STL are reported in Table 2. It is manifest that the accuracy produced by all models is higher than the random incidence (0.166 for 6-class dynamics; 0.333 for 3-class articulation). All the t-tests comparing the models with the feature selection procedure (# of feature = 15) and with full feature sets (# of feature = 81) do not reach the significant p-value of 0.05, which suggests that the top 15 features selected by ReliefF are effective descriptors regarding the expressiveness in conducting movement. It is evident from t-tests that the MTL model shows its advantage over STL models. It appears that these three musical features are intertwined in the musical context. Identifying phrase cuing can be the most challenging one among the three tasks. Considering the characteristics of movement kinematic signal, the phrase cuing can be easily confused with local-level beat cuing. Yet in the multi-task model, the dynamic and articulation information can help the recognition of phrase in one way or another.

The majority of previous research on conducting movement focuses on the beating pattern in basic level [24, 34, 35], and in this study, we make effort to explore the higher-level expressive semantics in conducting. There are several previous attempts to tackle the semantic-level in conducting, but exclusively target on the dynamic instructions, and tend to consider such gestures as isolated events regardless

Model	#	Phrase			Dync.	Artc.
		P	R	F	acc	acc
MTL	81	**60.39** **	**42.86** *	**48.48**	71.49 ***	76.45 ***
MTL	15	59.94 ***	42.10	47.63	**73.03** ***	**76.97** ***
STL	81	52.35 **	38.15 *	44.07	65.16 ***	70.25 ***
STL	15	48.05 ***	40.78	43.87	64.46 ***	68.95 ***

Table 2. Recognition results (in %) using RNN with multi-task setting (MTL) and single-task setting (STL): The precision (P), recall (R), F1 (F) measures for the phrase task, and the accuracy (acc) for dynamics (dync.) and articulation (artc.) tasks, comparing all movement features (# = 81) and the top 15 features selected by ReliefF (# = 15). Asterisks indicate the p-value of t-test of MTL and STL counterparts: * for $p < 0.05$; ** for $p < 0.01$; *** for $p < 0.001$.

the musical context [4, 8]. Previous studies examined the correlations of movement-music dynamic feature pairs and yield moderate r^2 ranging from 0.4 - 0.5 [33]. Our model takes another approach and is competent to investigate the complex inter-connections among multiple factors, and is able to produce further and solid results. As we expect from previous MIR research using MTL settings [12, 39], our MTL model demonstrates the advantage to consider multiple musical and movement elements together to investigate the complex inter-connections in the communication process during conducting performance. Moreover, as suggested by the previous research, musicians may perform the dynamics and articulation differently from the notated scores [20], the connection between the conducting movement and the performed sound can be further investigated using a similar approach presented in this study.

6. CONCLUSION AND FUTURE WORK

In this paper, we describe results from our investigation into the expressive semantics in conductors' movement used to communicate the phrase, dynamics, and articulation in music. The supervised feature selection technique ReliefF provides insights into effective descriptors in elementary kinematic signals of conducting movement. The movement elements from the right and left hand, and right and left wrists appear to carry important information regarding the conductor's expressive intentions. These selected descriptors are further investigated using recurrent neural network with multi-task learning. The RNN architecture yields improved results compared to previous works using other analysis techniques. Particularly, the multi-task learning model demonstrates a promising approach to examine the complex interactions among multiple musical and movement elements.

As the pioneering investigation on conducting movement using RNN, we highlight the potential for this framework to be applied to further explore other issues in music conducting, such as the connection between the conducting movement and the performance sound. The findings of such studies can enlighten the musical education for both conductors and orchestral musicians.

7. ACKNOWLEDGEMENT

Yu-Fen Huang is supported by the Postdoctoral Fellows Program of Academia Sinica, Taiwan. This work is partially supported by the Automatic Music Concert Animation (AMCA) project of the Institute of Information Science, Academia Sinica, Taiwan.

8. REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[2] B. Bartók. *Divertimento for String Orchestra, Sz.113*. Boosey & Hawkes, 1999.

[3] I. Bent. *Music Analysis in the Nineteenth Century: Volume 2, Hermeneutic Approaches*, volume 2. Cambridge University Press, 2005.

[4] B. Bruegge, C. Teschner, P. Lachenmaier, E. Fenzl, D. Schmidt, and S. Bierbaum. Pinocchio: conducting a virtual symphony orchestra. In *Proc. of the international conference on Advances in computer entertainment technology*, pages 294–295, 2007.

[5] P. Burkholder and D. Grout. *A History of Western Music: Ninth International Student Edition*. WW Norton & Company, 2014.

[6] C-motion. *Visual3D Wiki Documentation*. https://www.c-motion.com/v3dwiki/index.php?title=Tutorial:_Plug-In_Gait_Full-Body, access 2018.

[7] T.-P. Chen and L. Su. Functional harmony recognition of symbolic music data with multi-task recurrent neural networks. In *Proc. of the 19th International Society for Music Information Retrieval Conference, ISMIR*, pages 90–97, 2018.

[8] D. Dansereau, N. Brock, and J. Cooperstock. Predicting an orchestral conductor's baton movements using machine learning. *Computer Music Journal*, 37(2):28–45, 2013.

[9] A. Dvořák. *Serenade No.1, Op.22*. Dover Publications, 2001.

[10] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European Conference on Information Retrieval*, pages 345–359. Springer, 2005.

[11] E. Green, M. Gibson, and N. Malko. *The Modern Conductor*. Pearson Education, Upper Saddle River, NJ, 2004.

[12] P. Hamel, M. Davies, K. Yoshii, and M. Goto. Transfer learning in mir: Sharing learned latent representations for music audio classification and similarity. In *Proc. of the 14th International Society for Music Information Retrieval Conference, ISMIR*, pages 9–14, 2013.

[13] Y.-F. Huang. *Connecting Conductor Gesture to Compositional Features and Conductors' Expressive Intentions: An Exploratory Kinematic Study*. PhD Thesis, University of Edinburgh, 2018.

[14] Y.-F. Huang, S. Coleman, E. Barnhill, R. MacDonald, and N. Moran. How do orchestral conducting movement kinematics communicate musical structures and interpretational intentions. *Psychomusicology*, 27(3):148–157, 2017.

[15] D. Hunsberger and R. Ernst. *The Art of Conducting*. McGraw-Hill, New York, NY, 1992.

[16] T. Kelkar, U. Roy, and A. Jensenius. Evaluating a collection of sound-tracing data of melodic phrases. In *Proc. of the 19th International Society for Music Information Retrieval Conference, ISMIR*, pages 74–81, 2018.

[17] K. Kira and L. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Proc. of AAAI*, volume 2, pages 129–134, 1992.

[18] K. Kira and L. Rendell. A practical approach to feature selection. In *Machine Learning Proc.*, pages 249–256. Elsevier, 1992.

[19] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Proc. of European Conference on Machine Learning*, pages 171–182. Springer, 1994.

[20] K. Kosta, O. Bandtlow, and E. Chew. Dynamics and relativity: Practical implications of dynamic markings in the score. *Journal of New Music Research*, 47(5):438–461, 2018.

[21] J. Labuta. *Basic Conducting Techniques*. Pearson Education, Upper Saddle River, NJ, 2003.

[22] E. Lee, H. Kiel, S. Dedenbach, I. Grüll, T. Karrer, M. Wolf, and J. Borchers. Isymphony: an adaptive interactive orchestral conducting system for digital audio and video streams. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, pages 259–262. ACM, 2006.

[23] J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2018.

[24] Y. K. Lim and W. S. Yeo. Smartphone-based music conducting. In *Proc. of the International Conference on New Interfaces for Musical Expression , NIME*, pages 573–576, 2014.

[25] Z. Lipton, J. Berkowitz, and C. Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[26] G. Luck and S. Nte. An investigation of conductors' temporal gestures and conductor—musician synchronization, and a first experiment. *Psychology of Music*, 36(1):81–99, 2008.

[27] G. Luck and P. Toiviainen. Ensemble musicians' synchronization with conductors' gestures: An automated feature-extraction analysis. *Music Perception: An Interdisciplinary Journal*, 24(2):189–200, 2006.

[28] P. Modler and T. Myatt. Recognition of separate hand gestures by time-delay neural networks based on multistate spectral image patterns from cyclic hand movements. In *Proc. of IEEE International Conference on Systems, Man and Cybernetics*, pages 1539–1544, 2008.

[29] W. Mozart. *Eine Kleine Nachtmusik K.525*. Dover Publications, 2012.

[30] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan. This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, pages 1–13, 2018.

[31] M. Rudolf. *The Grammar of Conducting: A practical Study of Modern Baton Technique*. Schirmer, New York, NY, 1995.

[32] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee. Recent advances in recurrent neural networks. *arXiv preprint arXiv:1801.01078*, 2017.

[33] Á. Sarasúa and E. Guaus. Dynamics in music conducting: A computational comparative study among subjects. In *Proc. of the International Conference on New Interfaces for Musical Expression , NIME*, pages 195–200, 2014.

[34] R. Schramm, C. Jung, and E. Miranda. Dynamic time warping for music conducting gestures evaluation. *Proc. of IEEE Transactions on Multimedia*, 17(2):243–255, 2015.

[35] L.-W. Toh, W. Chao, and Y.-S. Chen. An interactive conducting system using kinect. In *Proc. of IEEE International Conference on Multimedia and Expo, ICME*, pages 1–6. IEEE, 2013.

[36] P. Toiviainen, G. Luck, and M. Thompson. Embodied meter: hierarchical eigenmodes in music-induced movement. *Music Perception: An Interdisciplinary Journal*, 28(1):59–70, 2010.

[37] R. Urbanowicz, M. Meeker, W. La Cava, R. Olson, and J. Moore. Relief-based feature selection: introduction and review. *Journal of biomedical informatics*, 85:189–203, 2018.

[38] C. Walder and D. Kim. Neural dynamic programming for musical self similarity. *arXiv preprint arXiv:1802.03144*, 2018.

[39] M.-H. Yang, L. Su, and Y.-H. Yang. Highlighting root notes in chord recognition using cepstral features and multi-task learning. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA*, pages 1–8, 2016.