AI-based Spectral Analysis on Music Conducting

Fourier-Analytic Expansion of Boulez-Ito's Angular Dynamic Method for Conducting

Ken ITO, Kazuki OTSUKA and Kota HAYASHI

Interfaculty Initiative in Information Studies
The University of Tokyo

Abstract

We present the spectral analysis of conducting physical movements by exploiting machine learning systems pre-trained human pose vector representations. 2 or 3-dimensional dynamic motions are extracted from 2D original video recorded in actual musical live performances. The fundamental movement of conducting are projected into angular velocity and angular acceleration according to the Angular-Dynamics method originated with Pierre Boulez and Ken Ito. Musculoskeletal motion spectrum are calculated when Fourier transform is applied. No previous work discussing spectrums and harmonics of corporal musical movements exists. A new way of systematization on various musical performances would be expected from here.

Index Terms — musical conducting, physical movement, machine learning, musculoskeletal, Fourier transform, spectrum analysis

1 From Berlioz's Geometrical Figures toward Saito Method

The Western "Conductor" are considered to emerge around the 17th century in France which is the age of Jean-Baptiste LULLY (1632-87). Hector BERLIOZ(1803-69) who performed in the early 19th century in France can be referred to as the origin of the modern conducting. The 4-beat's "geometrical figure", depicted in Figure 1, which is estimated to be developed by Berlioz can be used even in the 21st century's performances.

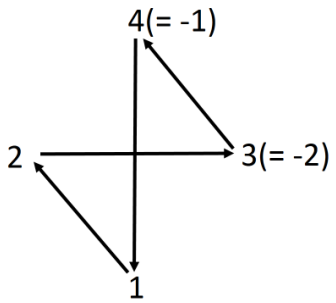


Figure.1 "Berlioz's figure" for quadruple trajectory

This figure shows the 3rd beat as “-2’ and the 4th beat as “-1”. These assignments imply 2 beats and 1 beat before the next bar respectively. Various “conducting methods” are developed around the same time of the appearance of the professional conductor mainly after Hans von Bülow(1832－94) in the later 19th century. Geometrical figures shown in Figure 2 are used in 21st century’s compulsory education for teaching 4-beats.

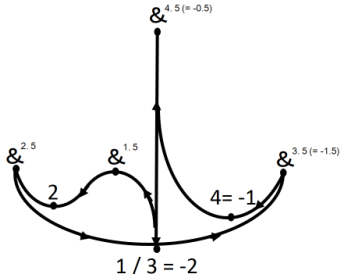


Fig.2 Typical trajectory for quadruple meter from 19th century

A Jewish Polish conductor Joseph ROSENSTOCK(1895-1985) practiced a way of simple performance and interpretation according to original pieces on the contrary to the trend of New Objectivity in the early 20th century. Hideki SAITO(1902-1974) systematized Rosenstock’s conducting technics by developing his original model, to be called “Saito method” generally, when Rosenstock had stayed in Japan between 1936-1946. Figure 3 shows 4-beat geometrical figure from Saito method.

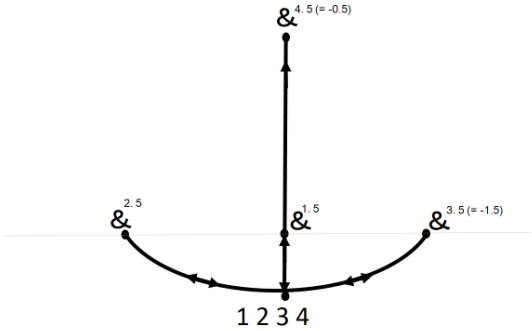


Fig.3 Figure of quadruple meter after the “SAITO method”

Comparing Figure 1~3, beats are represented as lines in Figure 1 while they are represented as curves in Figure 2. The way of representation on Figure 2 also uses points of off-beat dividing a single beat although it has a difficulty to express accurate moments of beats for performers since the position of each point of beats differs. Rosenstock and Saito’s method solved this problem by concentrating every point of beats into a single central position. This method is generally considered to suitable for advanced professional purposes since it enables accurate beat.

Now here looking at the analytical side of these orbits, the “infection point” which appears on Figure 2 while removed on Figure 3 can be pointed out.

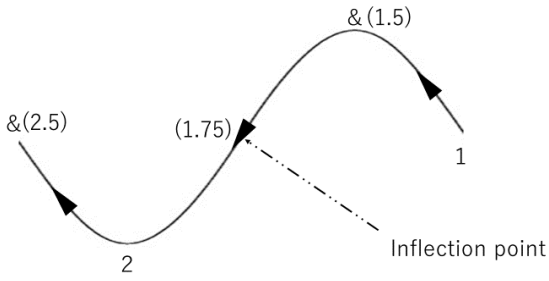


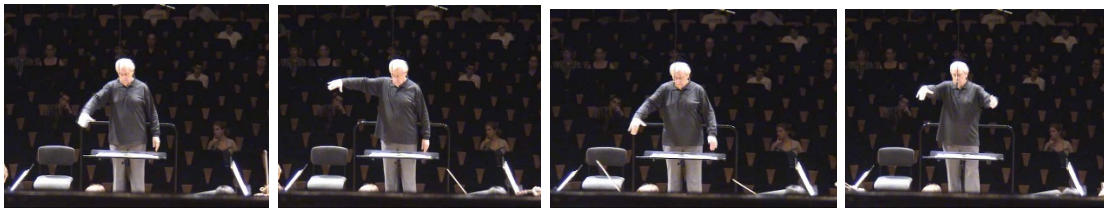
Fig.2 Inflection point in the19th century-type trajectory

Between the points of 1st and 2nd beat on Figure 2, an infection point equivalents to beat of near around 1.75 can be observed from the fact that the maximal is equivalents to 1.5 and the minimal is 2. These points doesn’t appear as static points in actual performances rather go through at the almost maximum velocity. However it is worth noticing that the four division note from a basic beat is emerged.

2 Boulez-Ito’s Angular Dynamics Method

Pierre BOULEZ(1925-2016), a French composer / conductor who was active in the late 20th and early 21st centuries, pointed out the confused usage of terms identifying conducting motions and the absence of proper methods, and therefore appealed the need of an appropriate measure. To meet his demand, one of the authors of this article, Ken ITO, organized a system of Angular Dynamics, a conducting method which doesn’t use conventional diagrams, by using motion capture image analysis (2004-07). Considering the actual conducting movements as a composition of joint rotations, in a broad sense, driven by skeletal muscles, it can be written only with the angular change, more precisely “angular velocity” and “angular acceleration”.

1 (&1.5) 2 (&2.5)



3 (&3.5= - 1.5) 4 (&4.5= - 0.5)

Fig. 5 Quadruple conducting gesture by Pierre Boulez

Using the Boulez-Ito's method, the "diagram" usually drawn on 2-dimensional plane can be replaced to a degree 2 of freedom derived from the elbow joint and the wrist joint. That is to say, the orbit which are drawn as the projection into x-y plane with a degree 2 of freedom, is represented as the elbow joint flexion angle θ and the wrist joint rotation angle ϕ (and the lower arm length r considered as a constant value). This representation method doesn't only include all the information that "diagram" had basically, but also identifies which part of the body to be used, furthermore it has a potential to develop clearly more advanced performance techniques by positively using the angular velocity $\dot{\theta}, \dot{\varphi}$, the angular acceleration $\ddot{\theta}, \ddot{\varphi}$, and so on.

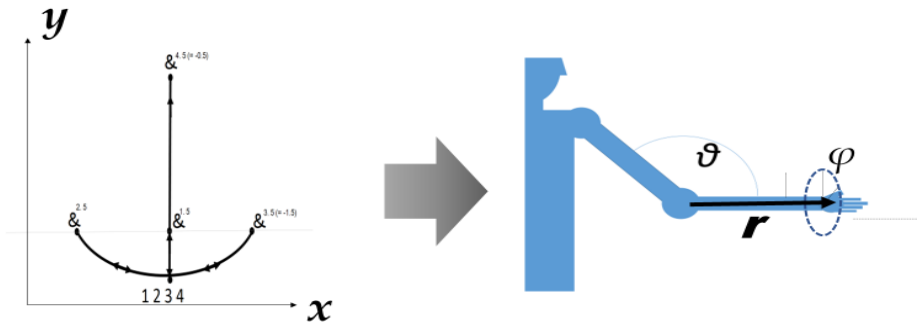


Fig. 6 From x-y plane projection to θ - ϕ -r angular dynamics

From this view point, Boulez and Ito systematically recorded and analyzed motion data from 2005 to 2007, but there was a limitation to effective analysis since the marker-less motion capture technology of the 2000s which can be used without disturbing performances has a limitation of the analytical degree of freedom.

3 Spectrum Analysis using Machine Learning Models (A) 2D

Features of physical movement are extracted from images which keeps the same camera angle as when it was recorded. We used CNN-based multi-layered neural network pre-trained OpenPose [2] model developed in CMU(Carnegie Mellon

University) Perceptual Computing Lab to analyze the motion as PAF(Part Affinity Field) representation which includes coordinates of keypoints within a vector space composed of 25 parts of body. The pre-trained model has been trained by using MP II human multi-person dataset [3] and COCO [4] human pose dataset. To extract features from right arm movements, 3 point's 2-dimensional coordinates which corresponds to shoulder joint, elbow joint, and wrist were estimated on movies recorded from right angle of the performer. The elbow angle θ can be calculated as:

$$\theta = arccos\left(\frac{\vec{v1} \cdot \vec{v2}}{|\vec{v1}||\vec{v2}|}\right) \tag{1}$$

where $\vec{v1}$ is the vector which connects two coordinates of shoulder joint and elbow joint, $\vec{v2}$ is the vector which connects two coordinates of elbow joint and wrist.

The derivatives as time changing rate, the velocity $\dot{\theta}$ and the acceleration $\ddot{\theta}$ can be referred to as:

$$L = (\text{duration sec}) \times (\text{sampling rate}) \tag{2}$$

$$\dot{\theta}_i = \theta_{i+1} - \theta_i \{i \in \mathbb{N} \mid 1 \le i \le L - 1\} \tag{3}$$

$$\ddot{\theta}_i = \dot{\theta}_{i+1} - \dot{\theta}_i \{i \in \mathbb{N} \mid 1 \le i \le L - 2\} \tag{4}$$

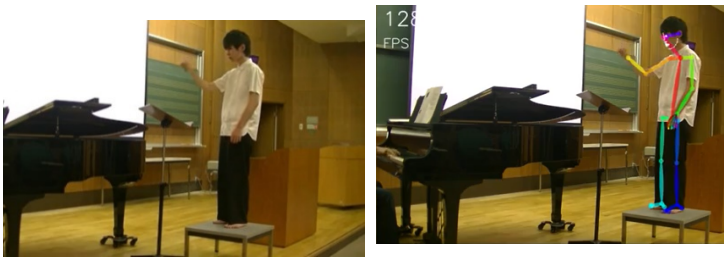


Fig. 7 Estimated joint parts from 2-D motion pictures

Figure 8 (a), (b), (c) shows examples of 2D projection $\theta, \dot{\theta}, \ddot{\theta}$ obtained from estimated joint movements.

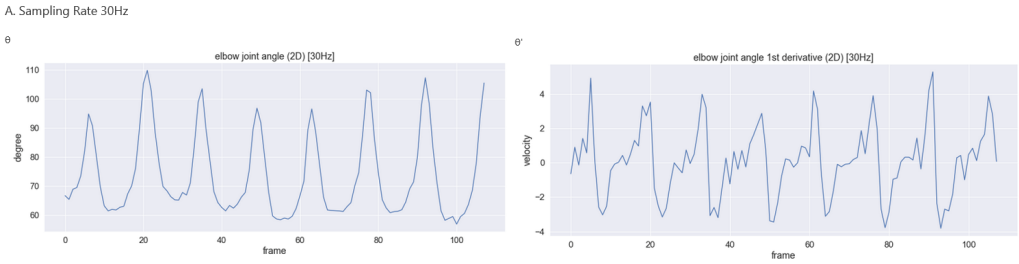
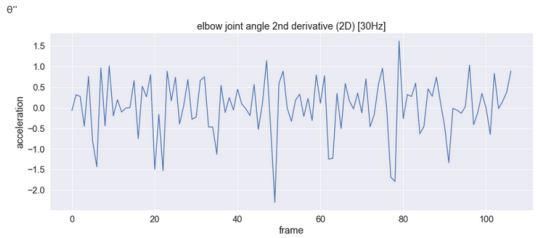


Fig.8 (a) A sample of estimated θ values. (b) A sample of estimated $\dot{\theta}$ values.



(c) A sample of estimated $\ddot{\theta}$ values.

It is taken as the basic characteristics to be that the transition of the angle draws upwardly convex curves, the first derivative looks nearly rectangular waves behavior, and the second derivative is close to a constant value in which spikes or beginner's noises derived from the extra forces are folded.

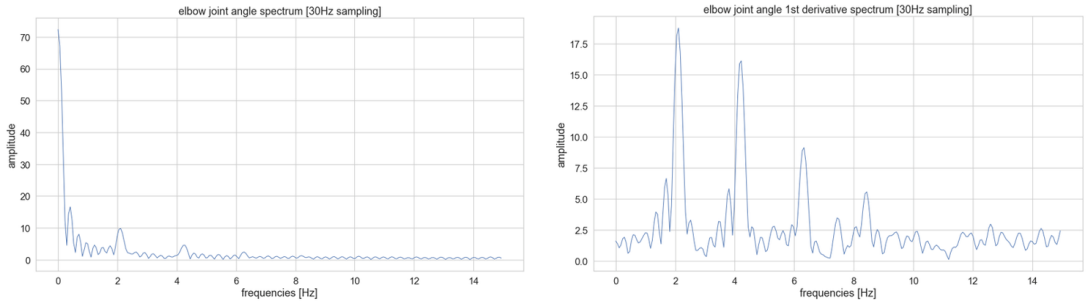
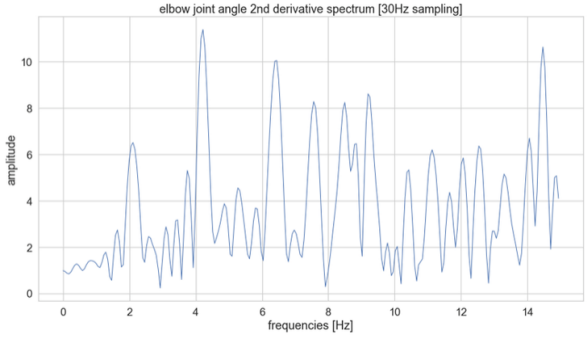


Fig.9-a. A sample of estimated θ spectra. b. A sample of estimated $\dot{\theta}$ spectra.

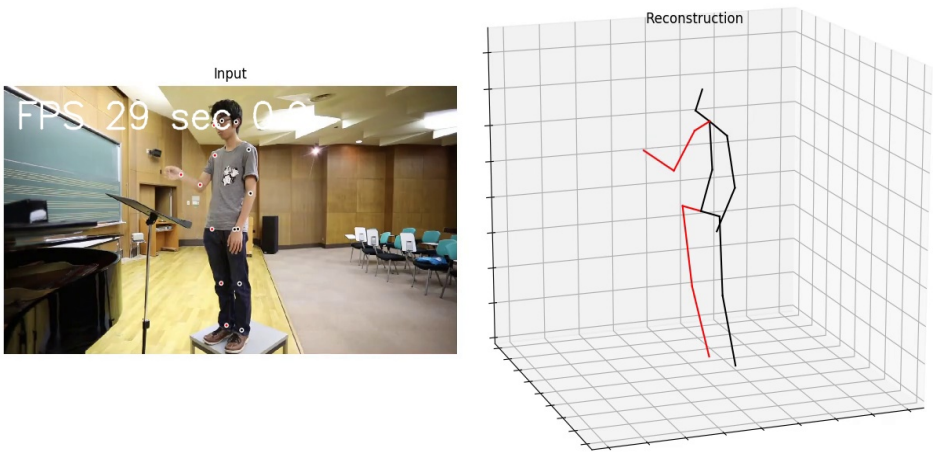


c. A sample of estimated $\ddot{\theta}$ spectra.

Fig.9-a represents a variety of elbow angle's changes such as approximately 2Hz, 4Hz and 6Hz equal to doubled series spectral structure from the repetitive movement. When the 2Hz corresponds to the fundamental beat, the 4Hz explicitly indicates the "half beat" going upward or downward, and the 6Hz indicates a hidden component dividing a beat into 3.
Fig.9-b shows a spectrum of angular velocity which clearly displays harmonic structure.
Fig.9-c shows a spectrum of angular acceleration explaining the beginner's movement which cannot have achieved well controlled motions requiring development of specific muscles yet, seeing a wide bandwidth components from low to the highest observable area.

4 Spectrum Analysis using Machine Learning Models (A) 3D

We also carried out the extraction of 3 dimensional physical movements from the same motion pictures. A neural network model VideoPose3D [5] developed by Facebook Research[1] is used for the 3 dimensional analysis. VideoPose3D inputs temporal sequence of 2D keypoints generated out from them model Detectron [7,8,9,10,11,12,13,14,15,16,17,18,19] which is also developed by Facebook Research, and it learns a mapping from the 2D keypoints into 3D vector space considering temporal dependencies by using CNN Dilated Temporal Convolution [20] architecture. The pre-tained model created by Facebook Research from Human3.6 [6] and COCO [4] human pose datasets was used. Similar to the 2-dimensional analysis, the angle θ, the angular velocity $\dot{\theta}$ and the angular acceleration $\ddot{\theta}$ were obtained from the time sequence of the joint positional keypoints, and the spectrum was obtained by Fourier transform on the sampling frequency.



[1] https://research.fb.com/

Fig.1 Estimation of 3-D corporeal motion data

Figure.11 shows the results of the 3D analysis over the 2D results on θ, $\dot{\theta}$, $\ddot{\theta}$ from the same video used in the previous chapter.

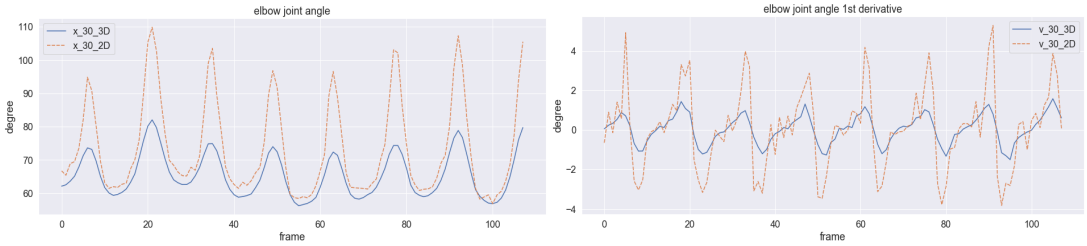
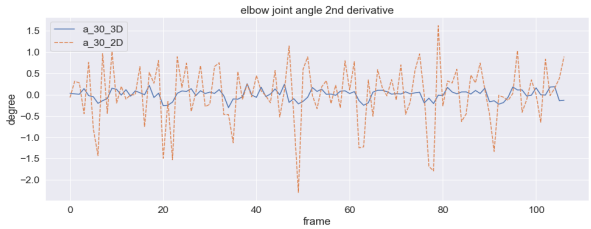


Fig.11-a. A sample of estimated 3 D- θ values. b. A sample of estimated 3 D $\dot{\theta}$ values.



c. A sample of estimated 3D- $\ddot{\theta}$ values.

The advantage of the 3-dimensional analysis is that the joint angle can be evaluated not by the geometrical projection but by an absolute value while losing a part of information obtained in 2-dimensional analysis. The output accuracy of machine learning system can be expected to enhanced by increasing the sampling frequency, that is, the frame rate of video recording when we newly shoot a performance in a video. Below are the results from motion pictures recorded at 120 fps frame rate (= 120Hz sampling frequency) on a scene that beginner(top)/intermediate(middle)/ advanced(bottom) conducted the same part in a piece. We used the beginning of P.Tchaikovsky String Serenade Op.48 II Valse: Moderato

5 Spectrum Analysis using Machine Learning Models (C) Historical Film

These analyses can be performed not only on newly recorded audio-video data but also on data recorded in the past, especially historical performance recordings. We experimented the same analysis using audio-video gathered in the early period when this author and P.Boulez established Angular Dynamics theory from 2004 to 2007. Suggestions by Carlos Agon from IRCAM[2] have had a great influence on it. We would like to express our sincere gratitude to him.

[2] https://www.ircam.fr/

Fig.13
Motion data estimation of P. Boulez's conducting and polygon synthesis

By synthesizing polygon meshes from physical movement and mapping them to model body, to analyze and reconstruct corporeal motions of historical masters who have already passed away, and to establish a next generation's orthodox performing method can be projected. We are investigating such direction of approach at the same time. In the following, let's examine the example of Boulez's performance obtained by the 3 dimensional analysis the same as previous sections. The music being played is the beginning of Luciano Berio "Chmin".

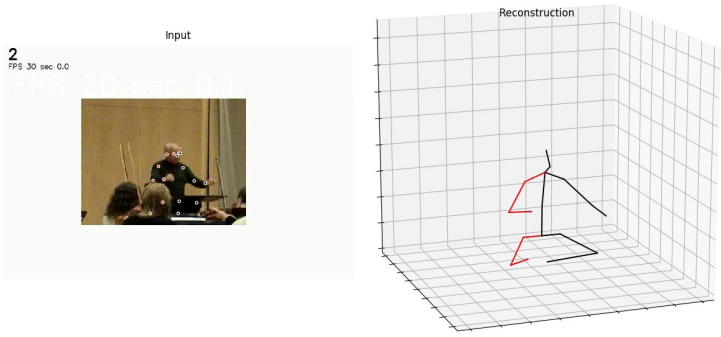


Fig.14　3D Motion data estimation of P. Boulez's conducting

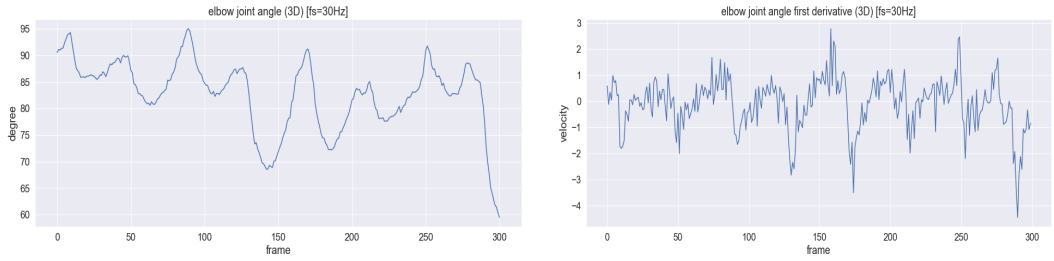
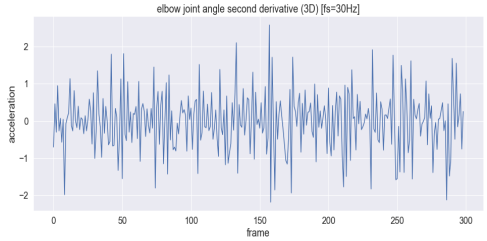


Fig.15-a. Boulez's sample of estimated 3 D- θ .　　b. Boulez's sample of estimated 3 D $\dot{\theta}$.



c. Boulez's sample of estimated 3D- $\ddot{\theta}$.

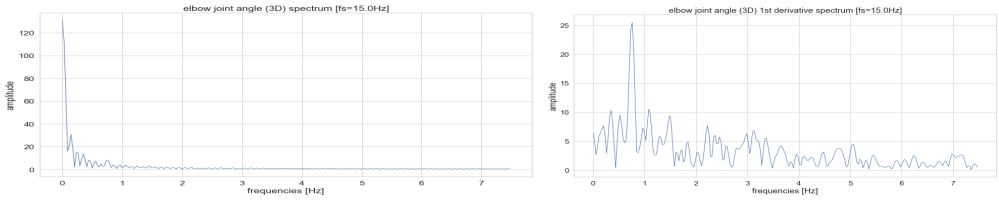
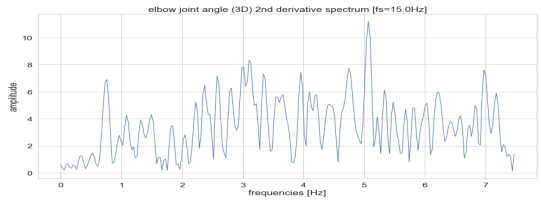


Fig.16-a. Boulez's sample of estimated 3 D- θ spectra. b. Boulez's sample of estimated 3 D $\dot{\theta}$ spectra.



c. Boulez's sample of estimated 3D- $\ddot{\theta}$ spectra.

Fig.15 shows the behavior of θ and the temporal change rate. Fig.16 shows the spectrum. From each graph of Fig.15, it can be confirmed that the conducting gesture itself is gentle. Fig.16-b indicates that the angular velocity has a clear peak around 0.8 Hz, that is, MM=75 as metronome tempo. At the same time, by observing the distribution of spectrum constituting the motion it is also verified that the motion is strictly controlled by a little more than 5Hz, that is, MM=300 which divides a beat into 4 counts, and a variety of cues are given from the motion. These analyses lively describe the tempo felt inside the internal mind of Boulez during his lifetime.

The correlation between these various cues and the player's specific performance can be analyzed through many-body analysis. We are currently working on these as ongoing issues.

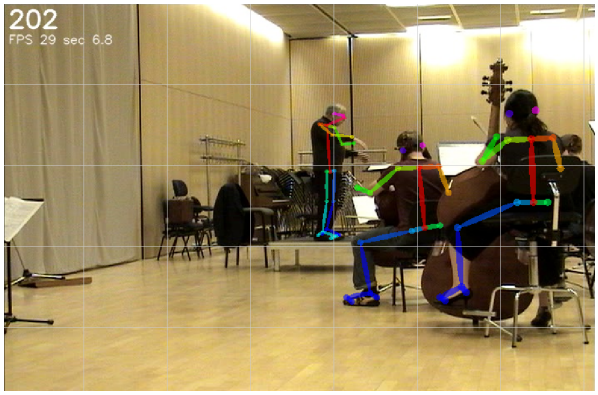


Fig.16 Example of many-body machine

6 Angular dynamics and Spectral methods : Towards a new conducting method

So as not to be misunderstood, we must emphasize that to create a robot performing as historical masters is not our goal. We are exclusively interested in providing an appropriate performance method for younger generations who create a music in the present and the future. On reflection, the language of music used in lesson rooms or rehearsals contains a great deal of rudimentary errors from a scientific point of view. To correct such errors, these AI-based analysis method has a great advantage of the ability to observe performance characteristics which is unable to be perceived subjectively by the performer himself. Although the same benefits can be applied for the improvement of sport technics or rehabilitation since the method is universal, musical technic is the finest example in the meaning that the merit can be utilized most delicately.
Neither the conductor himself nor players who see the conducting notices a remarkable spectrum peak in the beginner's baton movements as shown in Fig.9-b,c. However it becomes possible to acquire more advanced performance techniques easily and accurately by inspecting his motion through these analytical method.
The fact that the angular acceleration of the elbow can be resolved into clear peaks of frequency components is not well known to musicians in general, and the performing musician cannot recognize it.
But these can be understood naturally by considering the musculoskeletal system and the neural control.
Our musical performance is achieved by contraction of individual muscles that drives the skeletal system, and each muscular contraction has a time constant. The elbow joint consists of three borns, the humerus of the upper arm side, the ulna and the radius of the lower arm side, where many muscular starts and stops are combined.
Regarding the stretching of the elbow, the triceps brachial muscle (extension), the biceps brachial muscle (flexion) and the ancones muscle (extension), which are in an antagonistic relationship with each other, operate cooperatively. As for elbow flexion, many more muscle groups: the biceps brachial muscle, the brachial muscle, the brachioradial muscle, the circular prosthesis, the radial coral flexor, the ulnar carpal flexor, are involved.
These muscle groups are cooperatively related to the stretching of a single joint, and operates with unique time constants of each, thereby form a motion which seems to be simple. By dividing these into spectral components, we will introduce some techniques related to the independence of movement and especially decoupling force, which are difficult to be perceived subjectively for us.

1 Beat Division by Controlling Angular Momentum

Now we examine the spectral components of the Boulez conducting gesture shown in Fig.16. As mentioned earlier, the orbital motion describes a quite general beat and the acceleration components which form this movement show the remarkable peak of MM=300, that is, the 4 times of the basic beat. This motion can be explained as the diagram below.

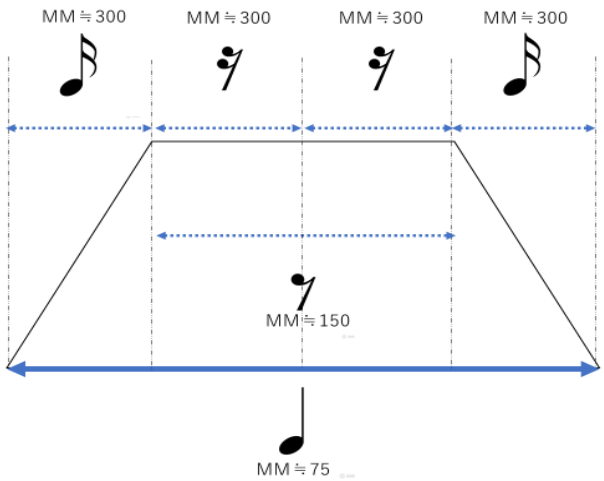


Fig.17 Example of many-body machine learning analysis and corporeal interaction study

The velocity of motion is not indicated generally in the orbital instruction using “geometric figure”. In actually the beats are divided into various patterns, and the velocities or accelerations are not equable. One example is shown as Fig.18.

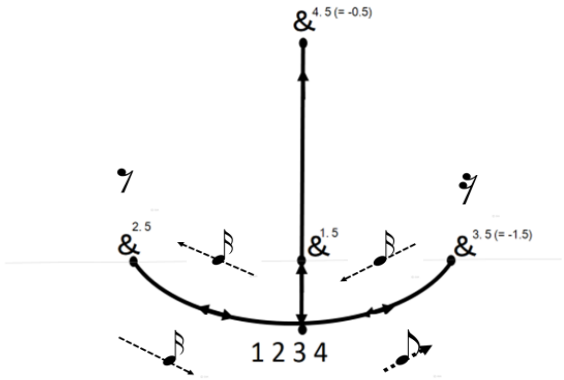


Fig.18 An example of inhomogeneous division of a quadruple meter

In the case of Fig.18, the second beat are indicated by the way of division shown in Fig.17. This instruction is effective for indicating how to divide the beat following the 3rd beat. Here, we assume the case that the 3rd beat moves along an 8th note in the first half and along an 16th note in the last half, and stops along a 16th note in a single point. The angular momentum and the changes showing these individual “orbital motion” can be observed by a spectrum as shown in Fig.16.

The diagrams in Fig.16 and Fig.17 are simplified explanations. Even so, the language used in a lesson room has only the same degree of accuracy as these, or even less accurate in general. Now we remember the schematic infection point of physical orbital motion.
In general, downward motions towards the direction which θ increases should be let go with the arm's weight or gravity by relaxing arms throughly when we perform a gentle conducting movement. The "infection point" of orbital motion indicates the second derivative of the orbit, that is, the point where the acting force becomes zero, in other word, a vicinity where the resultant force becomes zero due to antagonize with gravity.
As already mentioned, individual muscular contractions are considered to have time constants. To control them according to musical requirements, especially to alignments with beat division, gives accurate instructions to performers as a variety of cues.
The term "hit point" or "sharpness" in a lesson room all represents phenomena with a duration. The musical beat itself also have a duration rather a "point".
It should be noted that a simple and consistent acceleration and deceleration of the orbital motion are achieved as a whole by controlling individual muscle contraction time constants correspondingly to them.
The actual performance movement should be approximated not with lines as Fig.17 but with a sigmoid curve for instance. Considering the derivative and the fourier transform, it is obvious that beats have a spectral structure derived from multiple muscular contractions, and the musically consistent components result in an accurate music instruction. But it is difficult for performers to observe such facts subjectively. We named the development of the conducting method based on the harmonic analysis of physical movements, Spectral Conducting. More conducting methods of Spectral Conducting and the training methods to realize them will be developed.

2　Perturbation ——— Carpus × Elbow joint's Interaction, Higher-order Spectrum

In Boulez-Ito's Angular Dynamics method, wrist movements — especially rotations — play a crucial role. Therefore, the possibility of extracting wrist rotational motion is also examined in our machine learning analysis of motion pictures, but it is not always easy, especially when it comes to historical films. However, several facts appear when we pay attention to muscles which governs wrist motions.
For example, a flexion of wrist called "the palm flexion" relates to the superficial the finger flexor muscle, the deep finger flexor muscle, the radial coral flexor, the long palm muscle, the ulnar carpal flexor, the long mother finger flexor and so on. Among them, the radial carpal flexor is also involved in the flexion of the elbow joint as described above, and the deep finger flexor muscle and the deep finger flexor muscle are also adjacent to the elbow joint of the forearm, controlling the wrist motion, at the same time, it has a structure that affects the elbow joint. For this reason, wrist

movements often physiologically induce elbow joint movements, and for conducting and playing piano, it is considered that wrist movements can be detected as fine higher-order components of the elbow joint movement, that is, perturbations.

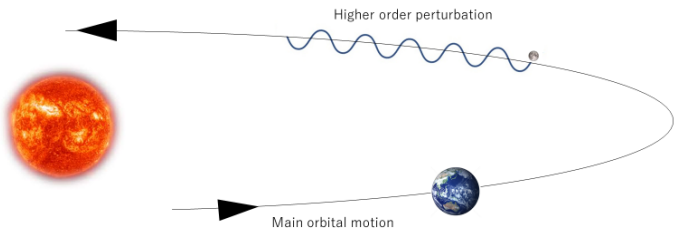


Fig.19 Main orbital motion and higher order perturbation

In classical mechanics, the two-body problem such as the Sun and the Earth, or the Earth and the Moon can be solved exactly, but the many-body problem which handles more than 3 objects such as the sun, the Earth and the Moon simultaneously. However, it may be possible to solve by the appropriate approximation assuming that the third and subsequent forces are sufficiently weak. If the existence of the Moon is ignored, the Earth is thought to move along an elliptical orbit around the Sun, but in actually a slight shift is caused by the interaction with the Moon. The perturbation method has been developed to calculate this deviation in classical mechanics, and has played an important role especially for the development of quantum mechanics and quantum field theory. The higher-order components found in the angular acceleration spectrum of the elbow joint movement, such as the Boulez's performance data shown in Fig.16-c, can be interpreted in principle rather as the higher-order perturbation consequently observed in the elbow joint movement as the result of the activation of the wrist or finger governing muscle, than the macro motion of the elbow itself. We continue systematic research on actual physiological details.

3　Tact, Keyboard, Binding Frequency ——— Towards an Essentially Unknown Realm

P.Boulez didn't use a tact and stood on a podium with empty hands throughout his life. However, keeping more general conducting techniques in mind, the motion of tact is necessary to be considered. To solve this problem, it is possible to use a model learned a dataset including batons achieved by extending human body, but when we think about the perturbation idea, the possibility of analysis using a normal physical motion dataset without batons can be expected.
As for the performances of historical masters who have passed away, such as Boulez, the frame rate of remaining films is limited, and it is generally difficult to expect a temporal resolution of more than 30fps. Even if the apparent time resolution is

increased by a method such as linear interpolation, it would be difficult or impossible to extract non-linear informations.

In this point, it is important to point out that increasing the frame rate allows us to approach a new problem, with respect to newly recordable musical performance videos. Most historical video materials are recorded at a frame rate of about 30fps because recording equipments have been designed according to the fact that we human cannot recognize contexts in a pattern that changes at a high speed clearly exceeding 20Hz due to the limitation of our time discrimination resolution which is up to around 15Hz.
On the other hand, data recorded at a frame rate of 120fps or 240fps can display motion at a frequency of 60Hz or 120Hz, meanwhile these movements are belonging to the audible range when considered as sound waves. We cannot distinguish 60Hz vibrations, which is already recognized as a pitch, as a rhythm considering contexts. Trills, tremolo, or percussion roles are no longer what allows us to hear the context, therefore they are recognized as a kind of continuous sound.
Similarly, conductor's queuing with high-speed movements in excess of 20Hz gives qualitative cognition to performers and listeners. To give an actual example, a quick palm vibration reminding vibrato corresponds to it.
A realm of frequency which is hard to recognize, in other word, a cognitive blind spot of hearing exists, between the high-band frequency limit of rhythmic time discrimination, 16-20Hz, and the low-band frequency limit of so-called audible sound. This realm might correspond to the "binding frequency of apperception" which integrates human perceptions such as visual, auditory and tactile sensation. It has the frequency of the region, and is thought to govern nerve motive by γ wave shared in the entire brain.
From motion picture data recorded at 120fps, it is possible to physically extract the motion spectrum in a realm that humans cannot recognize both as rhythm and as pitch. Thereby, the possibility to capture a phenomenon that slips through our recognition and replace it with a wisdom of musical performance is expected.
In this context, it is expected to capture phenomena in the physiologically unrecognizable realm not only from the analysis of the conducting movement, but also rather from the analysis of keyboards or other instruments using high-speed camera.
Although it is difficult to apply learning data specialized for the movement of palm or fingers to the analysis of conducting movement, the new knowledge is expected to be gained by recording and analyzing finger motions at a higher frame rate on musical instrument performances.
By recording and analyzing with a high-speed camera of about 1000fps on pianos, bowed instruments, plucked instruments such as harp and koto, and many percussion instruments, sound generation and physical dynamics in both linear and non-linear audible ranges is realized especially focusing on the cognitive blind spot realm. From these approach, we continue to examine capturing usually overlooked phenomena and projecting it to a new way of performances and creations.

References

[1] https://googleblog.blogspot.com/2012/06/using-large-scale-brain-simulations-for.html
[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh,"OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields" In CVPR, 2017. https://github.com/CMU-Perceptual-Computing-Lab/openpose
[3] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields" in CVPR, 2017.
[4] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis" in Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, 2014, pp. 3686–3693.
[5] Dario Pavllo, Christoph Feichtenhofer, David Grangier, Michael Auli,"3D human pose estimation in video with temporal convolutions and semi-supervised training" In CVPR, 2019. https://github.com/facebookresearch/VideoPose3D
[6] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. "Hu- man3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments." Transaction on Pattern Analysis and Machine Intelligence (TPAMI), 2014.
[7] Data Distillation: Towards Omni-Supervised Learning. Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Tech report, arXiv, Dec. 2017.
[8] Learning to Segment Every Thing. Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Tech report, arXiv, Nov. 2017.
[9] Non-Local Neural Networks. Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Tech report, arXiv, Nov. 2017.
[10] Mask R-CNN. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. IEEE International Conference on Computer Vision (ICCV), 2017.
[11] Focal Loss for Dense Object Detection. Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. IEEE International Conference on Computer Vision (ICCV), 2017.
[12] Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Tech report, arXiv, June 2017.
[13] Detecting and Recognizing Human-Object Interactions. Georgia Gkioxari, Ross

Girshick, Piotr Dollár, and Kaiming He. Tech report, arXiv, Apr. 2017.
[14] Feature Pyramid Networks for Object Detection. Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[15] Aggregated Residual Transformations for Deep Neural Networks. Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[16] R-FCN: Object Detection via Region-based Fully Convolutional Networks. Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. Conference on Neural Information Processing Systems (NIPS), 2016.
[17] Deep Residual Learning for Image Recognition. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[18] Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Conference on Neural Information Processing Systems (NIPS), 2015.
[19] Fast R-CNN. Ross Girshick. IEEE International Conference on Computer Vision (ICCV), 2015.
[20] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. *Wavelets, Time- Frequency Methods and Phase Space*, -1:286, 01 1989. 3