Automatic scoring of synchronization from fingers motion capture and music beats

BAYD Hamza, GUYOT Patrice, BARDY Benoit, SLANGEN Pierre R. L.

EuroMov Digital Health in Motion, Univ. Montpellier IMT Mines Ales, Ales, France
hamza.bayd@mines-ales.fr, patrice.guyot@mines-ales.fr,
pierre.slangen@mines-ales.fr, benoit.bardy@umontpellier.fr

Abstract. Thanks to various technological progresses such as musical rhythm estimation and motion capture systems, the evaluation of synchronization performances between motion and music beats is today possible. In this paper, we propose an innovative playful method to assess synchronization between hand motion and music. In this application, the hand gestures are tracked from live webcam motion capture based on MediaPipe open-source framework. Musical beats are estimated from the Librosa library. The synchronization between finger motion and beats is then computed from the dynamic time warping algorithm. A preliminary study was conducted with two different songs mixed with tempo beats at different intensities. The results are encouraging and show different levels of performance according to the tempo of the songs.

Keywords: synchronization, motion capture, hand, finger, music beats, dynamic time warping

1 Introduction

Synchronization between music and human movements is typically based on the perception of a common rhythm. Humans respond to rhythmic parameters of music by synchronizing movement pattern to the beat, that explain the close connection between the perceptual and audio-motor system. In general, synchronization between individuals are based on predictive abilities that are fed by visual and auditory perception [1]. During the last few years, significant progress has been made in challenging computer vision problems, including human pose estimation, in particular for neural network and motion capture (MoCap) technology. The task of human pose estimation from video, nowadays applied in various domains such as sign language recognition, is usually based on body landmarks or keypoints from each video frame. Moreover high level semantics from motion such as joint collection distance and temporal displacement of the body (visual information) can also be computed [3, 4].

From the music side, beats are one of the most fundamental features linking to the accompanying movements. The ability to synchronize movements and music is primarily analyzed through the task of tapping. In more complex situations, such as walking, synchronization can be analyzed through the impact of

the steps, and their correspondence with the strong beats of the music piece [5]. However, the way in which sound and visual information interact in the perception and production of synchronization, intentional or spontaneous, is still poorly understood [6]. In the present study, we propose a webcam beat-tracking software that measure and quantify in real time a synchronization score between natural movements of hands and the musical beats.The rest of this paper is structured as follows. Section 2 presents a short review of related works for musical rhythm, human pose estimation and synchronization. Sections 3 provides an overview of the developed system. In Section 4, we describe a preliminary experiment and discuss the results.

2 Related works

2.1 Musical rhythmic features extraction

Rhythm in music denotes the element of time. While a series of notes repeats in a regular cycle, they form a rhythmic pattern. Moreover to indicate the annotation timings for the respective notes, musical rhythm can also specify the duration and intensity. The best way to describe beat in real life is the moment when people taps feet or hand while listening to a song. The tempo in music is the number of beats per minute (bpm), and can change over the course of the music audio [9]. Some of the main characteristics that humans can understand and perceive when listening to music are tempo, rhythm, onset and tatum [8], as shown in Table 1.

Table 1. Characteristics of Rhythm

Feature	Definition
Rhythm	The continuous repetition of a musical pattern with its variation as it moves over time
Tempo	The speed at which the music is played and is measured in beats per minute
Onsets	The time instant of the detectable start of a melodic note
Beat	The fundamental feature to the perception of timing in music usually grouped in bars
Downbeat	The first beat of a measure, the strongest in any meter
Tatum	Corresponds to the fastest perceived pulsation in the metrical structure

The music information retrieval is an interdisciplinary research field that enhances a wide range of applications such as beat tracking, structural analysis and music classification. It relies traditionally on features can be extracted in two domains, time domain or frequency domain, within different levels of abstraction. On one hand, low level features make sense for the machine as statistics

features which are extracted directly from the audio such as amplitude envelope, energy, spectral centroid, spectrogram and MFCC (Mel Frequency Cepstral Coefficents) [7, 15]. On the other hand, high-level features are related to the perception of rhythm.

Recent works on beat tracking are based on deep learning models to predict the beat position, using recurrent neural network and LSTM (Long Short Term Memory networks). These models directly process magnitude spectrograms at multiple levels and provide output feature recognizing beats and downbeats [11]. Furthermore, other researches relies on a temporal convolution network framework [10].

2.2 Human movement and MoCap

Human movement can be mathematically described in terms of distance, displacement, speed and articulations. Table 2 shows detailed motion features. Skeleton-based action is defined as the problem of localizing human joints also known as 2D and 3D keypoints (hands, elbows, wrists, etc) in images or videos. These features plays a critical role in various applications and has been widely used in multimedia applications such as human computer interaction [13]. For example, they make up the backbone for yoga, dance, and fitness applications [14]. A wide range of systems are available to collect continuous movement data. The most popular in the recent years is MoCap (motion capture), usually referring to motion representations in digital formats to allow quantitative analysis or real-time processing. Approaches can be split into two broad categories: marker-based approaches rely on infrared cameras and reflective markers (single or multiple instances), while marker-less approaches are only based on cameras (RGB-only images vs depth data RGB-D).

Table 2. Visual motion features

Feature	Definition
Body landmarks	The spatial location of human body articulation (keypoints) from visuals such as images and videos, each points described by three coordinates (x, y, z)
Articulation angles and orientation	Rotation between two adjacent body articulation with different orientation
Translation and rotation	translation and rotation related to the world coordinate axis (tx,ty,tz,rx,ry,rz)
Temporal displacement and acceleration	Predict spatiotemporal features and velocity from the previous frame to the current frame
Joint collection distances	Represent the changes over time of joint angles computed at each time frame and a fixed point of the skeleton

Human hand motion is highly articulate, but also highly constrained, which makes it difficult to model [18]. Vision-based hand pose estimation analysis has

been an important research topic as this can play an essential role in enhancing the experience on a variety of technology disciplines and platforms in order to control devices or to interact with computer interfaces. This analysis process can be resumed in two steps: detection and tracking. The first major block will detect hand position that contains palm from frames of video. This can be achieved via an oriented hand bounding box by means of depth based methods [19], or alternatively by RGB based methods [17]. The second stage is the tracking pipeline that can track multiple hands in real-time by detecting the keypoints coordinates using regression model [20]. The hand can be also modelled in several aspects like shape, kinematical structure, dynamics, and semantics.

2.3 Audiomotor synchronization

The human audio-motor systems are practically connected during musical performance. Moreover listening to music can easily encourage the motor system to act in the form of head nodding, foot tapping, and even dancing [27]. Further research on audio-motor systems is trying to test the ways in which audio-motor coordination is influenced by visual cues from a conductor's gestures [28]. Another popular tool for the systematic assessment of audio-motor system is the BAASTA (Battery for the Assessment of Auditory Sensorimotor and Timing Abilities) [29], that uses finger tapping with music and adaptive tapping to a sequence with a tempo change in order to evaluate the synchronization performance of the participants.

Musical cognition refers to the study of musical thinking field, that focuses to understand the mental processes and the way of human brain perception such as cortical activity during listening to the auditory rhythms through the methods of cognitive science, and neurological methods based on auditory motor synchronization [30]. Following this notion, recent research suggests that the processes of action-perception between neural and mechanical activities synchronize in different forms during musical activities, linking body movement and high level musical features [31].

3 Methodology

This section will present our approach to measure and quantify in real time the "synchronization score" between natural movements of hand (fingers) and the musical beats. Its workflow is presented in Figure 1. Inspired by the success attained in hand tracking and music beat tracking methods, we propose to exploit these methods in order to evaluate the synchronization with temporal alignment using dynamic time warping (DTW) algorithms. This will be applied to measure similarity between two temporal sequences of motion and music beat while the participant is listening to the music using human computer interaction based on computer vision.

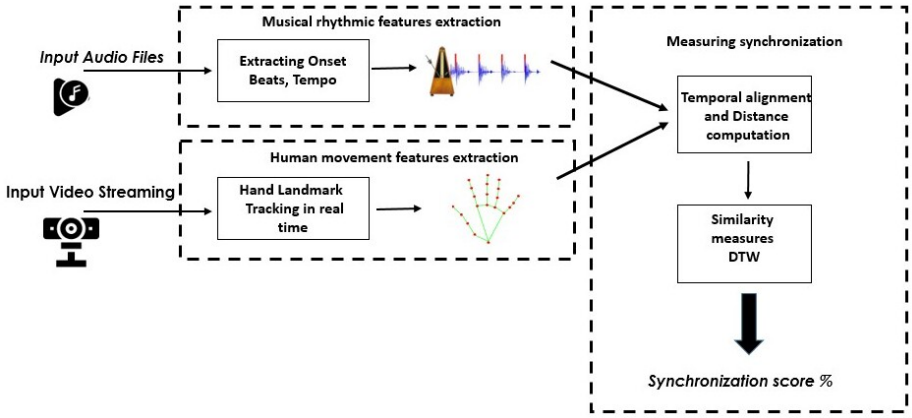

Fig. 1. Workflow Method Research

3.1 Motion parameters

The presented process relies on MediaPipe [21], an open source framework to detect in real time the hand keypoints (x,y) captured by the camera [22]. The MediaPipe module predicts 21 hand landmarks based on three axes: x (horizontal), y (vertical) and z that represents the depth of landmark. Then the hand score, indicating the accuracy of hand presence in the input image is generated. Finally a simple classification occurs between left and right hand. In this study, the landmarks of index fingers are firstly determined. Figure 2 presents hand landmark in MediaPipe [21]. The left hand index finger is shown by coordinates [5,6,7,8] for example. These landmarks are then used to make condition of tapping based on position and the angle of articulation corresponding to image width and height.

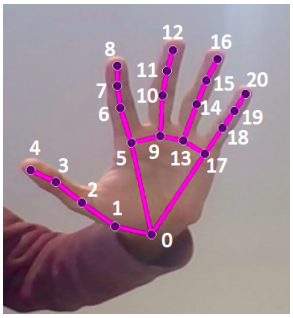

Fig. 2. The labeled keypoints generated by MediaPipe

3.2 Music Parameters

Instead of using low level acoustic features, the high level music features rhythm is adopted. Therefore tempo, onset and beat are used to represent musical rhythm. Beat in music theory is the most used feature in audio conditioned motion [25], that tracks the periodic element in music. Moreover detecting beat and tempo are closely related to onset detection. First the music audio signal is converted into spectrogram, then the coefficients of spectrogram are analyzed through neural network in order to classify each frame as onset or not onset. Finally, the beat location are founded by processing periodic patterns in the onset location. The public Librosa [23] is leveraged to extract the beats from the audio-based music. The beats features is a 1D vector, representing the annotation timings for the respective beats in which we consider it as rhythm information. Intuitively, the motion rhythm has to match the musical rhythm.

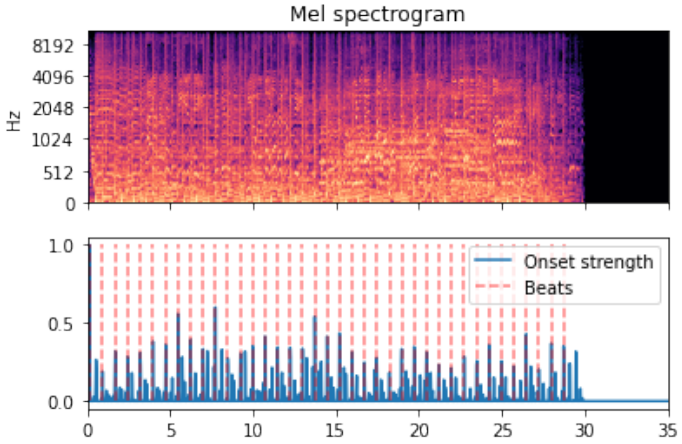

Fig. 3. Audio-based music waveform and the estimated beat positions

3.3 Similarity and score of synchronization

The aim of this paper is to develop more sophisticated methods for automated synchronization scoring between the hands movement and musical beat. As shown in the previous section we take temporal features with aggregation in real time from motion and music. Dynamic Time Warping is an algorithm that produces a better similarity measurement compared to others methods such as Hamming, Euclidean and Manhattan distance. The greatest advantage of DTW is to cope with signals of different length and in synchronization signals which move exactly at the same speed and time. Finally DTW can be used as a score component for synchronization between hand movement and musical beat [24].

The distance between two time series data n-dimensional space can be computed via the Euclidean distance $\mathbf{x} = [x_1, x_2, ..., x_n]$ and $\mathbf{y} = [y_1, y_2, ..., y_n]$

$$dist(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (1)$$

However, if the length of x is different from y, then the euclidean formula cannot be used to compute the similarity distance. Instead, a more flexible method must be developed to find the best mapping with the minimum distance from elements in x to those in y in order to compute the similarity synchronous.

Let x and y two vectors of lengths m and n :

$$D(i, j) = |x(i) - y(j)| + min \left\{ \begin{matrix} D(i-1, j) \\ D(i-1, j-1) \\ D(i, j-1) \end{matrix} \right\} \quad (2)$$

DTW is used to align the signals and computes the euclidean distance at each frame across every other frames to reach the minimum path that will match the two signals.

4 Experiments

4.1 Set-up

The objective of this demo showcase is to present a real application based on our innovative and robust hand-music synchronisation system. The main objective is to study a playful configuration coupled with an innovative method using the webcam sensor, in order to improve the automatically generated synchronisation score between the music rhythm and the hand movements (see Figure 4).

To test the synchronization, participants were asked to tap as regularly as possible with their two index fingers (Right and Left) in both directions and on the specific red button, in order to be synchronized to a rhythmic input music. We propose to the participants to test our demo in four difficulty levels to study the perception rhythm ability of each participant. The music was delivered over headphones at a comfortable sound volume level. The first "easy" level contains a clear beats pulse that were added in the background of the music, that can help the participants to match the beats rhythm easily. For the second level "moderate", the same musical audio is used but with low volume of the short pulses annotation... until level 4 "insane" that contains the audio music without beat pulses added to the original music. For this experiment two music clips of 35 seconds were selected, with different tempo and different style of music i.e. Ruby Baby *Act One* with 80 beats/minute including 39 beats, then Daft Punk *One More Time* but with 125 beats/minute and 65 of the placement beats.

Our algorithm was implemented in Python. All experiments are running on DELL with a Intel Core i5-8400H CPU 2.50GHz × 8 processor and 16 GB of RAM, and the Ubuntu 20.04.2 operating system. The lived processed images video is refreshed at 25 frames per second, with 1280×720 resolution.

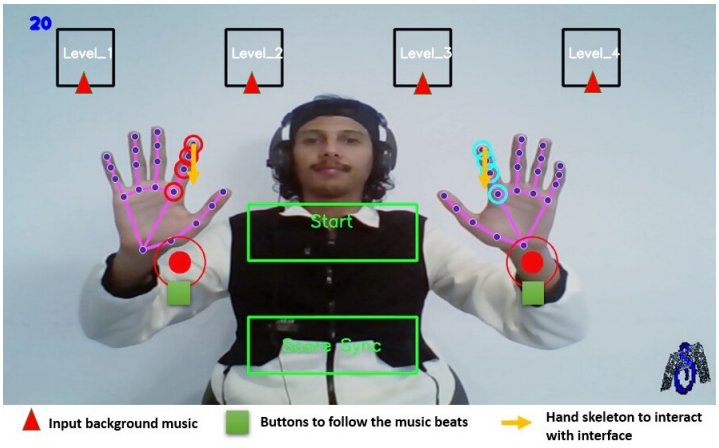

Fig. 4. The screen contains four **black boxes** to choose the music. 4 options can be selected with audio music that has annotated timings for the respective **beats annotations** with lower pitched pulses in order to help the participant to match the beats rhythm on four difficulty levels. Then there are two **red buttons** in the center of the screen to follow the beats with the fingers, and it can be done by moving the index finger in both directions on the specific red button, in order to synchronize with the rhythms of the audio music in the background.

4.2 Results

The present study was carried out on 3 non-musicians students (2 males and 1 female, average age 21). This experiment was designed to investigate the level of synchronization at different intensity of beats annotation and to examine the impact of the tempo on the performance (slow-rhythm and fast-rhythm).

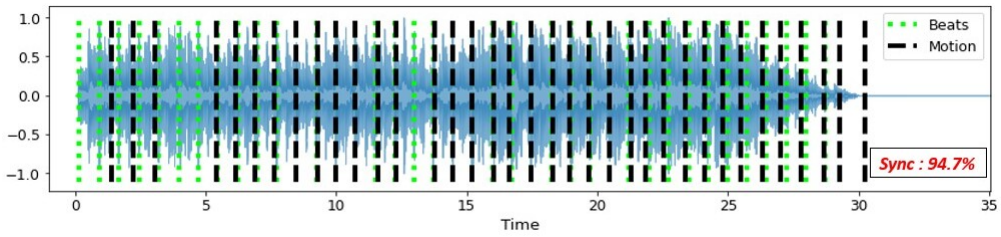

Fig. 5. Beat tracking for music waveform and motion beat, participant #2, lev moderate.

Figure 5 shows one of the 35s music clip with aligned musical beats and motion beats for participant #2. Then the global synchronization score for every

frame in the audio-video, the music beats marked by the green dashed line and the motion with the black dash-dotted line. It can be observed that the motion fingers tapping occur in very similar ways where the musical beats occur. The consecutive matched beats correspond to tapping index finger on left and right alternatively. For beat matching, the motion beats extracted from each frame and their corresponding input music beats are compared using euclidean distance and dynamic time warping. Then the score of synchronization is computed and displayed.

Table 3. Performance synchronization score (in percent)

Ruby Baby - ***Act One*** (80 bpm, 39 beats)

Difficulty levels	Participant #1	Participant #2	Participant #3
Easy	93.2	**97.8**	95.4
Moderate	88.5	94.7	89.6
Hard	83.2	90.5	82.3
Insane	76.1	88.7	78.9

Daft Punk - ***One more time*** (123 bpm, 65 beats)

Difficulty levels	Participant #1	Participant #2	Participant #3
Easy	**95.1**	86.8	88.7
Moderate	91.3	85.4	86.6
Hard	86.5	79.1	83.9
Insane	80.02	72.5	**70.4**

Results show fairly good performances of the participants. The best synchronization score (97.8%) is achieved for the song *Act One* at 80 bpm with an *easy* level of difficulty, and the smallest score (70.4%) for *One ore time* at 123 bpm with an *insane* level of difficulty. The average synchronization score is decreasing proportionally to the difficulty level. The overall average of synchronization score is higher (88.48%) for the first song at 80 bpm than for second one the (83.87%) at 123 bpm. This shows that the tempo impacts the ability of synchronization, in line with previous studies [32]. The overall best performance is achieved by participant #2 with 97.8% synchronization score. However the best performance with fast tempo is performed by participant #1 with 95.1% synchronization score.

Interviews were conducted with each participants after the experiments. The main feedback is that participants really appreciated the synchronization experience through this playful developed tools, and founded the application very easy to use. In addition, Participant 1's comments reveal that he is a fan of the *One more time* song. This could explain why he scores better on this song than on the first, unlike the other participants.

5 Conclusion

In this paper, we propose an efficient and accurate playful method for the automatic notation of the synchronization of hand movements and musical rhythms, using the real-time webcam display and musical audio as input.

This also enables to test the level of synchronization for a musical piece by tapping fingers with the beat times position. Moreover, an experiment was conducted trough four difficulty, with beat pulses added to the music and different mixing levels. The results show that the synchronization score decreases proportionally as the level of difficulty increase.

Future work will consider increasing the number of participants in the experiment to achieve statistical processing of the data. We will also integrate novel multi-scale music features contribution, such as downbeat and tatum. In order to increase the robustness of our synchronization score method we will extract the peaks and calculate the phase difference between the musical beat and finger movements at each cycle.

Finally, motion capture will be generalized to the whole body using different features such as articulation, joint collection distance and temporal displacement of the body landmarks. This will produce accurate body movement descriptors leading to fine analysis of synchronization between the human body and fully-described music features of the listened song.

References

1. Bardy, Benoît G., et al. "Moving in unison after perceptual interruption." Scientific reports 10.1 (2020)
2. De Cock, V. Cochen, et al. "Rhythmic abilities and musical training in Parkinson's disease: do they help?." NPJ Parkinson's disease, (2018)
3. Kadir, Md Eusha, et al. "Can a simple approach identify complex nurse care activity?." Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers, (2019).
4. Sun, Xiao, Chuankang Li, and Stephen Lin. "Explicit spatiotemporal joint relation learning for tracking human pose." Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. (2019).
5. De Cock, V. Cochen, et al. "Rhythmic abilities and musical training in Parkinson's disease: do they help?." NPJ Parkinson's disease, (2018)
6. Ipser, Alberta, et al. "Sight and sound persistently out of synch: stable individual differences in audiovisual synchronisation revealed by implicit measures of lip-voice integration." Scientific Reports 7.1 (2017)
7. Somesh Ganesh, Alexander Lerch. Tempo, Beat and Downbeat estimation for Electronic Dance Music, (2018)
8. Fuentes, Magdalena. Multi-scale computational rhythm analysis : a framework for sections, downbeats, beats, and microtiming, (2019)
9. Sogorski M, Geisel T, Priesemann V. Correlated microtiming deviations in jazz and rock music, (2018)
10. M. E. Davies and S. Böck, "Temporal convolutional networks for musical audio beat tracking," inProceed-ings of the 27th European Signal Processing Confer-ence (EUSIPCO), (2019)
11. Böck, Sebastian, Florian Krebs, and Gerhard Widmer. "Joint Beat and Downbeat Tracking with Recurrent Neural Networks." ISMIR, (2016)
12. Böck, Sebastian, Matthew EP Davies, and Peter Knees. "Multi-Task Learning of Tempo and Beat: Learning One to Improve the Other." ISMIR. 2019.
13. Ren, Zhou, et al. "Robust hand gesture recognition with kinect sensor." Proceedings of the 19th ACM international conference on Multimedia, (2011)
14. Zou, Jiaqi, et al. "Intelligent fitness trainer system based on human pose estimation." International Conference On Signal And Information Processing, Networking And Computers. Springer, Singapore, (2018)
15. Jensenius, Alexander Godøy, Rolf Wanderley, Marcelo. Developing Tools for Studying Musical Gestures within the Max/MSP/Jitter Environment. Proceedings of the International Computer Music Conference, (2011)
16. Huang, Ruozi, et al. "Dance revolution: Long-term dance generation with music via curriculum learning." arXiv preprint arXiv:2006.06119 (2020).
17. Sridhar, Srinath, et al. "Real-time joint tracking of a hand manipulating an object from rgb-d input." European Conference on Computer Vision. Springer, Cham, (2016)
18. WU, Ying et HUANG, Thomas S. For vision-based human computer interaction. studies, (2001)
19. Gattupalli, Srujana, et al. "Towards deep learning based hand keypoints detection for rapid sequential movements from rgb images." Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference, (2018)

20. Zhang, Fan, et al. "Mediapipe hands: On-device real-time hand tracking." arXiv preprint arXiv:2006.10214 (2020).
21. MediaPipe: Cross-platform ML solutions made simple. https://google.github.io/mediapipe/.2020
22. Zhang, Fan, et al. "Mediapipe hands: On-device real-time hand tracking." arXiv preprint arXiv:2006.10214 (2020).
23. McFee B, Raffel C, Liang D, Ellis DP, McVicar M, Battenberg E, et al. librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference, (2015)
24. Cheong, J. H. (2020, December 8). Four ways to quantify synchrony between time series data. https://doi.org/10.17605/OSF.IO/BA3NY
25. Huang, Ruozi, et al. "Dance revolution: Long-term dance generation with music via curriculum learning." arXiv preprint arXiv:2006.06119 (2020).
26. WU, Ying et HUANG, Thomas S. For vision-based human computer interaction. studies, 2001, vol. 5, p. 22.
27. Matt D. Schalles, Jaime A. Pineda, "Musical Sequence Learning and EEG Correlates of Audiomotor Processing", Behavioural Neurology, vol. 2015, Article ID 638202, 11 pages, (2015)
28. Colley, Ian D., et al. "The influence of visual cues on temporal anticipation and movement synchronization with musical sequences." Acta Psychologica 191 (2018)
29. Dalla Bella, Simone, et al. "BAASTA: Battery for the assessment of auditory sensorimotor and timing abilities." Behavior Research Methods 49.3 (2017)
30. DAMM, Loïc, VAROQUI, Déborah, DE COCK, Valérie Cochen,B, Bardy, Benoit et al. Why do we move to the beat? A multi-scale approach, from physical principles to brain dynamics. Neuroscience and Biobehavioral Reviews, (2020)
31. https://mutor-2.github.io/MUTOR/units/12.html
32. Repp, B.H. Sensorimotor synchronization: A review of the tapping literature. Psychonomic Bulletin Review 12, 969–992 (2005).