The Rearranger Ball: Delayed Gestural Control of Musical Sound using Online Unsupervised Temporal Segmentation

Juan Ignacio Mendoza
University of Jyväskylä
Finland
juigmend@student.jyu.fi

ABSTRACT

The state-of-the-art recognition of continuous gestures for control of musical sound by means of machine learning has two notable constraints. The first is that the system needs to be trained with individual example gestures, the starting and ending points of which need to be well defined. The second constraint is time required for the system to recognise that a gesture has occurred, which may prevent the quick action that musical performance typically requires. This article describes how a method for unsupervised segmentation of gestures, may be used for delayed gestural control of a musical system. The system allows a user to perform without explicitly indicating the starting and ending of gestures in order to train the machine learning algorithm. To demonstrate the feasibility of the system, an apparatus for control of musical sound was devised incorporating the time required by the process into the interaction paradigm. The unsupervised automatic segmentation method and the concept of delayed control are further proposed to be exploited in the design and implementation of systems that facilitate seamless human-machine musical interaction without the need for quick response time, for example when using broad motion of the human body.

Author Keywords

unsupervised, segmentation, music, gesture, controller

CCS Concepts

•**Human** → **centered computing;** •**Computing methodologies** → **Machine learning;** •**Information systems** → Music retrieval; •**Applied computing** → Performing arts;

1. INTRODUCTION

Musical instruments are usually designed to be controlled with fine movements of hands and fingers, as they afford precision and speed. These qualities are often described as the foundations of responsiveness, believed to be indispensable for musical expression. The instrument thus becomes an extension of the human body.

NIME'23, 31 May–2 June, 2023, Mexico City, Mexico.

These ideas have permeated into the design of digital musical instruments (DMI) [15], and a response time approaching zero has become a standard goal [23, 9, 10]. The challenge extends to the design of DMI that recognise gestures "in the air", using machine learning techniques. For example, a musician wears, holds or stands in front of, a device that may sense position (i.e., static gestures) or motion (i.e., continuous gestures). The musician makes a gesture in free space: describes a circle with the head, wiggles a hand, or stands in a particular pose. The DMI learns these gestures in a process called "training", and it recognises them when they are performed. The recognition of a gesture can be mapped to a musical action, such as triggering a sound, activating an effect, etc. (e.g., [8]).

Two algorithms and variations of them have been extensively used to recognise continuous gestures, regardless of the sensing technology: Dynamic Time Warping (DTW) [7] and Hidden Markov Models (HMM) [1]. Both estimate the likelihood that a gesture being performed corresponds to a gesture that has been learned in the training. However, this likelihood may change while the gesture is executed, therefore recognition is only reliable after the gesture has been completed. This adds time to the recognition, arguably reducing responsiveness. In addition, training requires the beginning and ending of gestures to be explicit.

Given a stream of data from a sensor, individual gestures may be extracted by a process called "segmentation", in which the start and ending points of gestures are identified. For example, when training the algorithm the user presses a button (e.g., [14]) or makes pauses between gestures (e.g., [16]). While this constraint has not prevented the use of the algorithms mentioned above in DMI, the ability of a machine to recognise and learn gestures without explicit training would open new avenues for human-machine musical interaction. Furthermore, the time required for the recognition of continuous gestures might not be a disadvantage if when designing a DMI we don't hold the same standards of responsiveness as for the human voice or other non-electronic instruments. Consider that digital technologies have greatly expanded our possibilities for control of sound, far beyond what is possible with the human voice or with non-electronic devices. Why should we hold ourselves from exploring forms of gestural control that are not quick and precise, but instead slow and imprecise (i.e., delayed detection, perception, action, by the user and the automatic system) such as broad motion of the human body?

This article describes a system that was devised as a proof of concept towards exploring the feasibility of unsupervised learning of patterns in a continuous input signal, in a musical application that doesn't require quick responsiveness. The system is conceptually a musical instrument in a broad sense, for it essentially allows a user to control sound.

2. ONLINE UNSUPERVISED TEMPORAL SEGMENTATION

A signal may be segmented using the algorithm described by Foote [5], which has seen application in segmentation of musical audio and video [6, 21], dancing motion captured by an accelerometer [13], and daily activity recorded by wearable accelerometers [12, 17]. Its meta-parameters can be adjusted to detect boundaries of segments at different timescales. The cited sources described the use of the algorithm on recorded data. Conversely, Schätti [18] described an online version of the algorithm, that detects boundaries of audio data while the data is being produced. Later Mendoza [11] reported a study in which the algorithm's segmentation of dancing motion captured by a hand-held accelerometer, was compared to manual segmentation of video recordings of the dancing. The meta-parameters were optimised for each accelerometry recording. The music used for dancing and the person doing the manual segmentation were the main factors affecting the quality of computed segmentation. These results suggest that the algorithm is suitable for gestural control of a DMI, albeit its meta-parameters might need contextual adjustment. Figure 1 succinctly illustrates the online segmentation procedure. It uses the same principle of buffering and computation of a local distance matrix, as described by Schätti [18] and Mendoza [11].

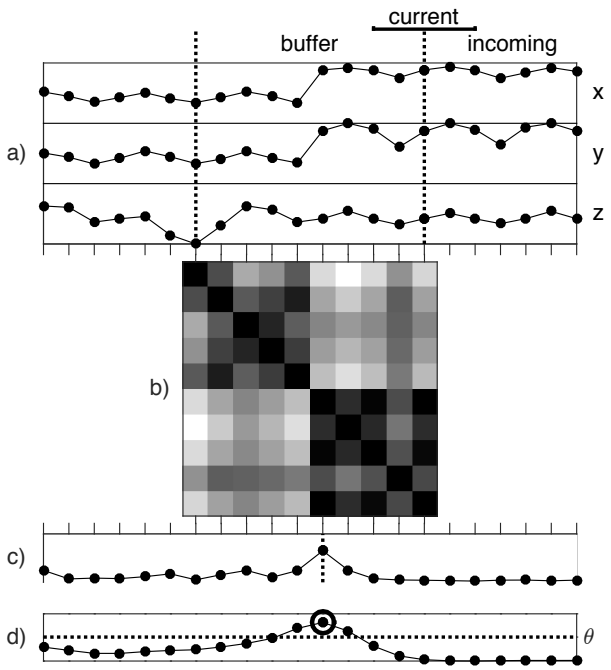

Figure 1: Online temporal segmentation. Horizontal axes represent time. (a) is accelerometer data composed of triaxial frames. (b) is a distance matrix of the data in the buffer having a length of n frames. Lighter shades represent more distance. (c) is a novelty score resulting from the correlation of the distance matrix with a gaussian-tapered checkerboard kernel. The vertical dotted line indicates the current result. (d) is the novelty score after smoothed by a gaussian filter of length n_{filt}, where θ is a threshold and the point in a circle is the selected peak indicating a boundary. Note that this visualisation shows (c) and (d) aligned in time, but in practice there will be a lag because of the filter. The total lag of the process is $(n + n_{filt})/2$ frames plus 3 frames for peak detection.

3. PROOF OF CONCEPT

3.1 Hardware

A polystyrene ball having 12 cm. of diameter was cut in half and the interior was carved to fit a Myo armband controller (Figure 2). The Myo was originally designed by Thalmic Labs to be worn on the forearm. It has several sensors, of which only its triaxial accelerometer was used in the system described here. The two halves of the ball are put together restoring the spherical shape, but it can be easily disassembled to recharge the battery of the Myo. The data from the sensors is broadcast in real time using the Bluetooth Low-Energy (BLE) specification. The BLE signal is captured by a computer nearby, and a piece of software written by Rodrigo Schramm[1] outputs the data in Open Sound Control (OSC) format to a User Datagram Protocol (UDP) port, where it can be accessed by other software. This controller was used for its convenience, as it was available to the researcher along with the software to get the data in real time.

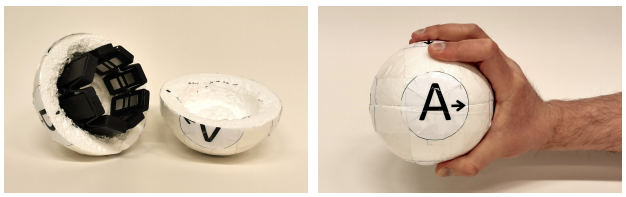

Figure 2: Left – Carved open polystyrene ball with the Myo armband in it. Right – Closed ball.

3.2 Software

The segmentation procedure described in section 2 can detect in real-time boundaries between gestures performed with the hand-held controller continuously, without indicating their start or end. The effect of its meta-parameters are as follows: n sets the timescale of gestures to detect, n_{filt} sets the smoothness of the novelty score, θ is a factor of the maximum novelty score and sets a threshold below which novelty peaks are rejected (e.g., noise). A further meta-parameter was incorporated to prevent detection of segments of less than a given length n_{min}, such as transitions between gestures. The segmentation procedure, as well as the musical application and its graphical user interface, were implemented in the Pure Data programming environment, which receives the accelerometry data using OSC as described in the previous subsection. The software is free and available (see Appendix).

The detected segments, each being a gesture, may be fed to a machine-learning process for training (i.e., gesture learning) and classification (i.e., gesture recognition). The DTW algorithm was chosen for this purpose, as it is available in the easy-to-use software Wekinator [4, 3], which communicates with Pure Data using OSC over a UDP port. However, another algorithm could be used (e.g., HMM). As with segmentation, the result of the recognition has lag due to buffering and latency due to logical processing.

[1] See [22]. Software available: https://github.com/federicoVisi/KineToolbox/blob/master/input%2BML/DaemonMYO

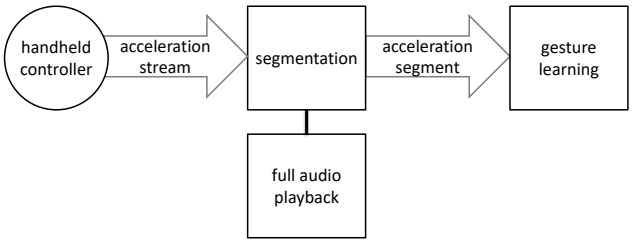

Figure 3: Cut stage

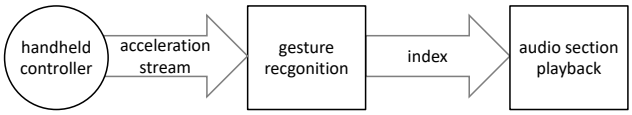

Figure 4: Perform stage

The segmentation and machine-learning processes are incorporated into a system that allows the user to reorder sections of an audio file. The use of the system has two stages: Cut and Perform. In the Cut stage (Figure 3) the audio file is played in its entirety while the user performs distinct gestures. The boundaries between gestures are detected in real time by the segmentation process and their time location is stored and labelled with a sequential index. The segments are fed as individual training examples to the gesture learning process. Also, in the graphical user interface a green vertical line is placed over a plot of the accelerometry signal, to indicate a successfully segmented gesture (Figure 5).

In the Perform stage (Figure 4) the gesture recognition process is continuously comparing the incoming accelerometry signal, to all the segments previously stored in the Cut stage. The segment that is closest to a stored one is deemed a match and its corresponding audio section plays in a loop. If a gesture different than the current is recognised, then the corresponding audio section will be played once the current audio section reaches its end.

3.3 Testing

During the implementation of the system, the author of this article conducted iterative testing using an upbeat electronic dance music piece, as it has been observed that this kind of music stimulates bodily motion [2]. Static gestures achieved by only changing the ball's orientation, and gestures involving repetitive motion, were well segmented and recognised. Figure 6 shows a sequence of gestures that worked well with the following setting of meta-parameters, which was kept throughout the testing: $n = 80$, $n_{min} = 28$, and $n_{filt} = 24$, at a sampling rate of 20 frames per second yielding $lag = 55$ frames (0.4 seconds, not including logical processing latency), and $\theta = 0.03$. Parameters of the DTW process were also adjusted, but are not discussed as that algorithm is well documented [7, 3]. Since the ball is fully symmetrical, letters (A to F) were put on the orthogonal points to aid visually in manipulation. Later a small arrow was put next to each letter pointing to the next one (Figure 2, Right).

Additionally, extraction of features (e.g., amplitude, zero-crossings) from the triaxial accelerometry signal and its magnitude, was implemented. They did not improve segmentation but, because of being windowed processes, they did increase lag (i.e., frames needed for computation) and computation cost (i.e., logical processing). Therefore, development and testing continued using only raw acceleration, to demonstrate what is possible without using extracted features.

When a functional version was completed, researchers and students of Musicology, Music Therapy and Music Education at the University of Jyväskylä were invited to evaluate the functionality of the system. With this group the following protocol was developed:

1. The researcher demonstrates the task comprising Cut and Perform stages, using the upbeat electronic dance music piece and the tested gestures sequence. The enclosed rectangle shown in Figure 6 is is displayed on a paper.

2. The participant is invited to do the task. If in the Cut stage not all gestures were segmented successfully, the participant is invited to repeat the Cut, as many times as they want. Then, they are invited to try the Perform stage.

3. The participant is invited to freely improvise and/or to use another piece of music.

4. The participant is invited and encouraged to express their opinion on the experience. The researcher shall take observational notes such as number of gestures correctly segmented in a trial, comments and ideas expressed by and discussed with the participant, and if a new gesture is discovered.

The protocol described above was incorporated to a 7-hour presentation in an outreach event at the University of Jyväskylä. The following data was collected of 23 participants: age, gender, number of gestures successfully segmented consecutively from the first, and observations. Further notes were taken of more more visitors. All participants used the upbeat electronic dance music, except one discarded for homogeneity. 17 participants (10 female, 7 male) performed the task as intended. Only six tried a second time, improving segmentation (see Figure 7). The medians of correctly segmented gestures was 4 for first time, 6 for second time and 5 for maxima. No correlation between number of correct segments and age or gender was observed. Most participants under 10 years old could not correctly perform all gestures, albeit they could successfully use the system by only changing the orientation of the ball.

3.4 Overall Assessment

Any set of orientations being different enough will work, but the 6 orthogonal orientations work flawlessly. Also, any combination and variation of repeated movements along the 3 orthogonal axes of the ball will work well. Sudden and energetic movements work best, as they are better measured by the accelerometer. Smooth movements are less likely to be detected by the system. Participants discovered a variety of gestures beyond those in the task. One of them is the "baby rocking", consisting in holding the ball with two hands and moving it describing an upwards concave curve. Other semi-circular and circular motions, and "8" figures were successfully detected, inasmuch as the speed, and therefore radial acceleration, was powerful enough to produce a novelty score above the set threshold (θ).

If the transition from one gesture to the next is slow enough to have a duration equal or greater than n_{min} (minimum duration for gestures to be detected), the transition will be identified as a segment. In the Perform stage the system might get stuck looping these very short segments, due to the characteristics of the DTW algorithm (i.e., computation time is proportional to the length of the segment, parameter sensitivity). However, interestingly, two participants mentioned that they liked the result. One of them referred to it as "a DJ effect". Another participant explored

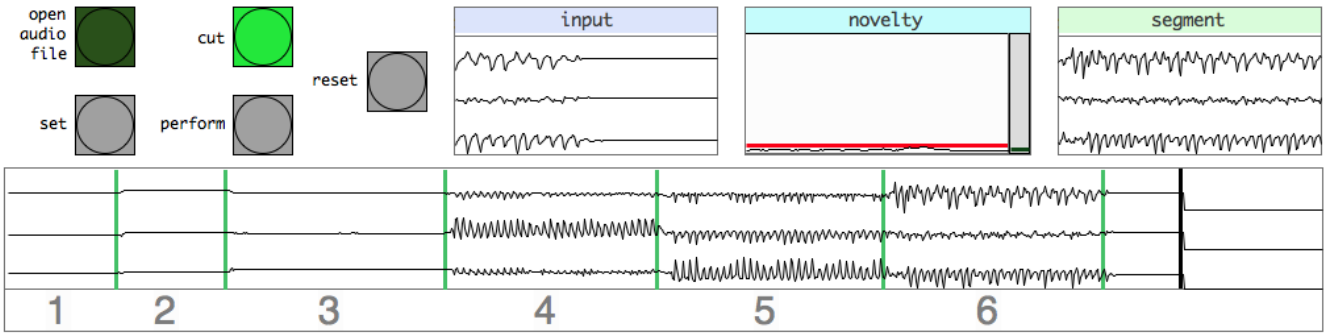

Figure 5: Graphical user interface

order	orientation	gesture	description
1	A	·	do nothing
			rotate left
2	B	·	do nothing
			rotate forward
3	C	·	do nothing
			rotate left
4	D	...	move up-down
			rotate forward
5	E		hit right
			rotate left
6	F		hit twice to each side
		·	do nothing
7	F	·	do nothing

Figure 6: Segmentation task

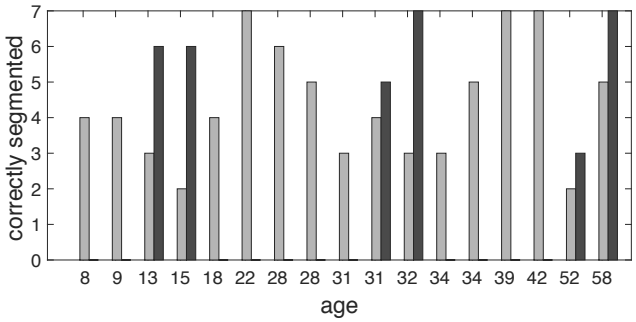

Figure 7: Data collected at the outreach event. Second trials are shown in darker shade.

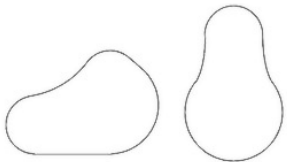

Figure 8: "Boot" form with rotational asymmetry. Left – lateral view. Right – zenithal view.

the possibility of not having to look at the ball when manipulating it. A discussion ensued leading to conclude that, since the ball is fully symmetric, it is not possible to be aware of its orientation without looking at it.

The task was challenging to different extents. Some participants wanted to try again to improve the number of correctly segmented gestures. All participants showed engagement and enjoyment. However, it is to expect that researchers and students have interest as the experience is related to their profession and studies. Likewise, visitors at the outreach event most probably attended because of curiosity.

4. DISCUSSION AND FUTURE WORK

The system described in this article demonstrates the feasibility of unsupervised learning of patterns in a continuous input signal, for gestural control, within a musical application. The process ineluctably produces a lagged response and therefore it is not suitable for the execution of fast notes or rhythmic patterns. Nonetheless, the proposed musical application conforms to this constraint, supporting the concept of delayed control of musical sound. Participants of the assessment tended to regard the task as a challenge, which in combination with the discovery of new meaningful gestures, and the sense-making of the constraints, turned the experience into a ludic one. The system appears promising, offering opportunities for further research:

I. The reported assessment used recorded music, but any audio file may be used, and the meta-parameters may be tweaked for further exploration that may lead to unexpected yet interesting results.

II. The hand-held device will benefit from having rotational asymmetry, such that there is no need of looking at it for manipulation (Figure 8).

III. Using the raw accelerometry signal has established a baseline. Future research could evaluate the impact of features extracted from the raw signal. The computation of such features will impact the overall latency (lag plus logical processing), and the detection of novelty (and therefore the setting of meta-parameters) because of the information that the features carry.

IV. Incorporation of more sensors or sensing technologies other than accelerometry. Besides, several sensors may be used by more than one person simultaneously, as a group activity (e.g., [19, 20]).

V. Implementation of online multigranular segmentation, meaning the detection of gestural boundaries at different timescales.

VI. Current limitations to achieve **III**, **IV**, and **V**, are algorithmic complexity, processing power and software efficiency. Solutions may include low-level programming (possibly embedded software) and faster hardware (possibly parallel computing of several features and timescales).

VII. The setting of meta-parameters generalised well, which is unexpected as perceptual evaluations have suggested the adjustment of meta-parameters for each user [11]. A different setting might be needed when using other configurations of hardware, software, music, user, etc. Future research may assess the effects of meta-parameters on segmentation and user experience.

VIII. The methods described in this article have potential beyond the described application, in which the online segmentation procedure only contributes to display on the screen an indication when a gesture has been successfully segmented in the Cut stage. This allows the user, for example, to stop the Cut and restart if a gesture change was not detected. While this might be an advantage to the user, the online segmentation capability and its further possibilities for near-real-time interaction could be exploited more. For example, a musical system (e.g., a DMI, a sonic installation, a sonification) may learn gestures as they occur. This may be incorporated to interactive systems where both the user and the system discover and learn gestures at the same time, leading to a seamless process of human-machine musical interaction.

5. ETHICAL STANDARDS

All participants gave verbal informed consent for the use of their anonymous collected data, following the research ethics guidelines by the University of Jyväskylä.

6. REFERENCES

[1] F. Bevilacqua, B. Zamborlin, A. Sypniewski, N. Schnell, F. Guédy, and N. Rasamimanana. Continuous realtime gesture following and recognition. In *Gesture in Embodied Communication and Human-Computer Interaction*, pages 73–84. Springer, 2010.

[2] B. Burger and P. Toiviainen. Embodiment in electronic dance music: Effects of musical content and structure on body movement. *Musicae Scientiae*, 24(2):186–205, 2020.

[3] R. Fiebrink. http://www.wekinator.org/

[4] R. Fiebrink, D. Trueman, and P. R. Cook. A meta-instrument for interactive, on-the-fly machine learning. In *Proceedings of NIME*, 2009.

[5] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of ICME2000*, vol. 1, pages 452–455, 2000.

[6] J. Foote and M. L. Cooper. Media segmentation using self-similarity decomposition. In *Proc. SPIE, Storage and Retrieval for Media Databases*, vol. 5021, pages 167–175., 2003.

[7] N. Gillian, B. Knapp, and S. O'modhrain. Recognition of multivariate temporal musical gestures using n-dimensional dynamic time warping. In *Proceedings of NIME*, 2011.

[8] N. E. Gillian. *Gesture recognition for musician computer interaction*. PhD thesis, Queen's University Belfast, 2011.

[9] R. H. Jack, A. Mehrabi, T. Stockman, and A. McPherson. Action-sound latency and the perceived quality of digital musical instruments: Comparing professional percussionists and amateur musicians. *Music Perception*, 36(1):109–128, 2018.

[10] A. McPherson, R. Jack, and G. Moro. Action-sound latency: Are our tools fast enough? In *Proceedings of NIME*, 2016.

[11] J. I. Mendoza. Segmentation boundaries in accelerometer data of arm motion induced by music: Online computation and perceptual assessment. *Human Technology*, 18(3):250–266, 2022.

[12] J. I. Mendoza, A. Danso, G. Luck, T. Rantalainen, L. Palmberg, and S. Chastin. Musification of accelerometry data towards raising awareness of physical activity. In *Proceedings of SoniHED*, 2022.

[13] J. I. Mendoza and M. R. Thompson. Modelling perceived segmentation of bodily gestures induced by music. In *Proceedings of ESCOM*, pages 128–133. 2017.

[14] D. J. Merrill and J. A. Paradiso. Personalization, expressivity, and learnability of an implicit mapping strategy for physical interfaces. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 2152–2161, 2005.

[15] F. R. Moore. The dysfunctions of midi. *Computer music journal*, 12(1):19–28, 1988.

[16] D. Murad, F. Ye, M. Barone, and Y. Wang. Motion initiated music ensemble with sensors for motor rehabilitation. In *2017 international conference on orange technologies*, pages 87–90. IEEE, 2017.

[17] J. Rodrigues, P. Probst, and H. Gamboa. Tssummarize: A visual strategy to summarize biosignals. In *International conference on Bio Signals, Images, and Instrumentation*, pages 1–6, 2021.

[18] G. Schätti. Real-time audio feature analysis for decklight3, 2007.

[19] E. Staudt, Pascal; Sarigöl, M. Lussana, M. Rizzonelli, and J. Hyun Kim. Automatic classification of interactive gestures for inter-body proximity sonification. In *Proceedings of SoniHED*, 2022.

[20] K. Tahiroğlu, N. N. Correia, and M. Espada. Pesi extended system: In space, on body, with 3 musicians. In *Proceedings of NIME*, 2013.

[21] D. Tardieu, R. Chessini, J. Dubois, S. Dupont, S. Hidot, B. Mazzarino, A. Moinet, X. Siebert, G. Varni, and A. Visentin. Video navigation tool: Application to browsing a database of dancers' performances. In *5th International Summer Workshop on Multimodal Interfaces*, pages 35 – 40. 2009.

[22] F. Visi. *Methods and Technologies for the Analysis and Interactive Use of Body Movements in Instrumental Music Performance*. PhD thesis, Plymouth University, 2017.

[23] D. Wessel and M. Wright. Problems and prospects for intimate musical control of computers. *Computer music journal*, 26(3):11–22, 2002.

APPENDIX

Software and documentation: https://gitlab.jyu.fi/juigmend/temporal_segmentation_gestural_control