Towards Automatic Unsupervised Segmentation of Music-Induced Arm Gestures from Accelerometer Data

Juan Ignacio Mendoza
Department of Music, Art and Cultures Studies
University of Jyväskylä
Finland
juigmend@student.jyu.fi

Marc Richard Thompson
Department of Music, Art and Cultures Studies
University of Jyväskylä
Finland
marc.thompson@jyu.fi

ABSTRACT

This article presents preliminary results of an ongoing investigation project whose goal is to model perceived segmentation of music-induced gestures. The project consists of three stages. The first stage is a database of multimodal recordings of people moving to music in two conditions: free movement and *dancing with one arm*. The data of these recordings are video, motion (position and acceleration at the hand of the arm that moves in the second condition) and audio (i.e., the music). In the second stage the videos produced in the first stage are manually segmented. These perceived segmentation boundaries constitute ground truth to evaluate an automatic gesture segmentation system developed in the third stage of the project. This system extracts kinetic features from motion data at a single point. Then a novelty score is computed from the kinetic features. The peaks of the novelty score indicate segmentation boundaries. So far only data form the condition dancing with one arm has been considered. The kinetic features that have been evaluated are composed of only one windowed statistical function. None of them yields a reasonable similarity between computed and perceived boundaries. However, some of the functions yield considerably similar results between perceived and computed boundaries at isolated regions. This suggests that each of these functions perform best on a specific kind of motion and thus, of gesture. Further work considers evaluating more kinetic features in combination and the use a genetic algorithm to optimise the search for greatest similarity between computed and perceived segmentation boundaries

KEYWORDS

unsupervised, automatic, segmentation, gesture, music, accelerometer

1 INTRODUCTION

1.1 Background

In line with the Embodied Music Cognition train of thought [1], it has been argued that a person's spontaneous movement when listening to music can reflect the person's perception of the music. Qualitative investigation has observed, for example, that music teachers explain musical sound with bodily movements, especially with their hands [2]. Quantitative investigation has shown that bodily movement induced by music relates to features of the music, such as periodicity and kinetic energy [3] or tonality [4]. The correspondence between music and bodily movement has been studied under the term *musical gesture* [5]. It has been noted that human beings have a remarkable ability to perceive and understand musical gestures by visual observation [6]. The first stage in perception of a gesture is the identification of when and where it starts and ends, a process called *segmentation* [7]. Further phenomenological inquiry has observed that musical gestures are perceived in different time scales and that the grouping of shorter-scale gestures into larger entities depends on musical structure, a phenomenon called *co-articulation* [8].

Several studies have observed the relation between bodily movement of people making music and moving to music (e.g., dancing) using qualitative analysis of video recordings [9-13]. Because the careful observation of video is a time-consuming task, these studies have focused in a few examples. Therefore their results, while being important for advancing knowledge, are not appropriate for generalisation. In contrast, a large-scale experimental investigation that could yield statistically relevant results, would take the effort of people watching many videos. These videos should show a range of individuals moving to different kinds of music and the observation thereof should include precise annotations of where gestures occur and a description of them. Such an endeavour appears to be prohibitive in terms of human resources. Thus, it seems reasonable to automate the process, which requires first to model human perceived segmentation of gestures.

A model of human perceived segmentation of gestures would allow a machine to automatically recognise gestures without having an exact idea of how each individual gesture looks like. This capability might find practical applications not only in the understanding of musical gestures.

1.2 Aims of this Study

This study focuses on *music-induced gestures*, which are here defined as the meaningful movement patterns that a person does when spontaneously moving along when presented with music. The starting point is the observation, modelling and prediction of perceived gestures induced by music made by only one arm. This constrain allows to reduce the complexity of the problem. In what follows the overall three-stage methodology of the project is explained and preliminary results are discussed.

2 METHODOLOGY

2.1 Multimodal Database

2.1.1 Aim. This stage of the investigation consists in collecting multimodal data, which allows to observe people's spontaneous movement to music. The data modalities are:

- Tri-axial position
- Tri-axial acceleration
- Video
- Audio

2.1.2 Participants. N = 12, of which 7 (58.3%) are female and 6 (41.7%) are male. Their range of ages is 23 to 53, median 33. All of them are degree students, researchers or other staff at the University of Jyväskylä. None of them is associated with the Music, Art and Culture Studies department or with research in musicology. All participants signed a document giving consent to the use of recorded data for research and communication thereof, including audio and video recordings.

2.1.3 Apparatus. Data was collected at the motion capture laboratory of the Department of Music, Art and Culture Studies at the University of Jyväskylä. The following measurement devices and processes compose the apparatus:

- Optical Motion Capture: An array of 8 Qualisys Oqus cameras track the position of reflective markers attached to a tight suit that the participants wear. Markers were placed at every articulation and ending point of limbs, as well as on the head. Optical motion capture data is recorded using the Qualisys Track Manager software running in a personal computer. This system was syncronised to an SMPTE signal emitted by a second computer. Also the Qualisys system sent back a syncronisation audio signal to the second computer.
- Tri-axial accelerometer: The participant held a Nintendo Wii-remote ('Wiimote') controller with one hand, as explained in paragraph 2.1.5. Data from this device was recorded at a rate of 100 Hz in the second computer, using software made with the Pure Data programming environment [14]. This software also simultaneously recorded audio.
- Audio: Stimuli were presented to the participants using two Genelec 8030-A studio loudspeakers with their base at 110 cm. from the floor. A microphone hanging from the ceiling was connected to the audio system of the second computer, which simultaneously recorded this audio stream (i.e., room audio) in one audio channel and the audio syncronisation signal from the optical motion capture system in a second channel. The starting and ending of the audio recording was set to be at the same time of the Wiimote's data recording. The audio signal is used later to synchronise the recored time series, by trimming their beginnings to the beginning of the corresponding music.
- Video: Two small digital cameras (Vivitar DVR-786 and Sony DSC-W610) recorded video and room audio. They were placed together, pointing perpendicular to the wall. The room shape is a rectangle. The image shows the participant's full-body against a white wall. Redundancy of video recordings serves as a backup strategy. Using the room audio, the video stream is syncronised to data from the Wiimote and optical motion capture. The use of mall video cameras allowed flexibility when positioning them, opposed to the use of cameras fixed to the wall or mounted on cumbersome rigging.

2.1.4 Stimuli. The list below shows the excerpts of music that were used and a brief description that explains the choice.

- 'Bouzouki Hiphop' [15] from the beginning to 45.7 s. with no fade-in or fade-out. This is Rembetiko instrumental music mixed with Hip-hop bass and drums, published on the Internet by an independent artist. Tempo is 90 BPM and meter is binary. All participants declared to not know this piece.
- 'Minuet in G Major' [16]. MIDI rendition with piano sound, from beginning to end (104 bars, 93 s.) with no fade-in or fade-out. Tempo is ca. 128 BPM and meter is ternary. All participants declared to know this piece.
- 'Ciguri' [17] from 56 to 180 s. with fade-out the last 5 s. This is an electroacoustic piece that has no perceivable beat indicating tempo and that has *'an insistent and virtually isochronic rapid percussion attack, together with one or more streams of sustained electroacoustic sound with somewhat clear pitch structure'* [18]. All participants declared to not know this piece.
- 'Stayin' Alive' [19] from the beginning to 108 s. with fade-out the last 2.3 s. Tempo is 104 BPM and meter is binary. All participants declared to know this piece.

2.1.5 Procedure. Data recording was done with one participant at a time. Participants were asked to move spontaneously to the stimulus when it started sounding through the loudspeakers. They were not asked to dance as it was observed in pilot experiments that if they are asked to dance they feel inhibited because they are afraid to fail. This fear derives from the association of the word 'dance' with movements that have to be done correctly, as inferred from participants' accounts. However, if participants are asked to *move to music* this inhibition disappears. In fact, participants usually ask '*Do I have to dance?*'. When they ask this question, they are told that they can dance if they want, otherwise they can move freely.

Each stimulus was presented twice. Participants were asked on the first presentation to move to the music without any constraint other than an area of approximately 9m^2, which corresponds to the bounds of the optical motion capture and video recording systems. The second time, participants were asked to hold the Wiimote with one hand and *dance* only with that arm. In this condition participants were asked to remain at the center of the area facing to a corner of the room. This was done to get in the video recording the most complete visualisation of the arm's movement. In this condition participants were allowed to move the rest of the body naturally as long as the previous constraints were not violated. This procedure (a 'trial') was repeated for each stimulus.

Stimuli are presented in the order of the list above (*4. Stimuli*). However, participants were told that the first stimulus (Bouzouki...) was *just for practice*. Indeed that trial was intended to be a practice so that the participant could get familiarity with the procedure. Still, data for this stimulus is recorded and kept. Participants were allowed to rest as much as needed between trials.

2.2 Ground Truth

2.2.1 Aims. In this stage the videos from the Multimodal Database were manually segmented in two conditions: *real-time* and *non-real-time*. In each condition the time location of segmentation boundaries was recorded. This task is called *annotation*.

- Real-time annotation: Videos with their corresponding audio were segmented as they were watched.
- Non-real-time annotation: Videos without audio were segmented as they were watched, with the option of scrolling back and forth to refine the annotation.

2.2.2 Participants. N = 7, of which 3 (42.9%) are female and 4 (57.1%) are male. Their range of ages is 26 to 39, median 27. All of them are students at the University of Jyväskylä that have completed at least an introductory course in music psychology comprising the Embodied Music Cognition and Segmentation. 5 of them (71.4%) are degree students of the programme 'Music, Mind and Technology'. Two of them are doctoral students in musicology, out of which one is the author of this article. These participants are regarded to have an expertise to perform the required tasks ranging from semi-expert to expert, with a median of semi-expert.

2.2.3 Apparatus.

- Real-time annotation: A personal computer running a custom-made piece of software made with the Pure Data programming environment, which automatically presents the video and records the elapsed time when depressing a key of the computer's keyboard. These times are recorded in a CSV (comma-separated-values) text file.
- Non-real-time annotation: A personal computer running the Reaper digital audio editing software [20] (Cockos Reaper, 2010). This system allows video playback at different speeds, scrolling through the video and accurately placing markers. These markers are exported as a CSV file.

2.2.4 Stimuli. After trying different alternatives in pilot experiments, it was deemed convenient to use only two videos in a single annotation session. Each of these videos shows a different person (one male, other female) moving their arms to the same music excerpt: the song Stayin' Alive. By using only two stimuli it is possible to complete the procedure (described in the next paragraph) in less than an hour (most participants completed the procedure in around 40 minutes) to prevent cognitive overload and fatigue.

2.2.5 Procedure.

- Real-time annotation: Participants were presented with the following instructions on the computer screen:

> *You will be presented with two videos, each lasting around two minutes. Each video shows a person 'dancing' with an arm. When doing this the person does distinct patterns with the arm. A pattern is composed by one distinct movement or several repetitions of the same movement. When the video is playing press the space bar to indicate a change in pattern. Focus in the movement of the arm holding the white device (it is a sensor*).

These instructions are a revised version after the pilot experiments, in which participants were asked to indicate 'a change of gesture'. As a result of this request the participants attempted to indicate every single movement, which indeed highly correlates with the beat of the song. However, it was deemed desirable to observe the grouping of motion into larger structures and likewise, their correspondence with musical structures larger than the beat. The constraints of the revised instructions allow the participants to judge the grouping of observed movements. Thus, the recorded responses account for gestures at a resolution level defined by the grouping constraints. Before performing the task while watching the videos described in section *2.2.4*, the participant performed the task while watching a video of a person moving to the 'Bouzouki' stimulus, as a practice.

- Non-real-time annotation: Participants were asked to do the same task as in the real-time condition with the only difference that in this condition is possible to scroll back and refine the placement of the markers.

2.2.5 Data Analysis. Responses by all participants are summarised into a single compound response for each condition. This is done using Kernel Density Estimation, which produces a curve of density. The peaks of this curve, over a threshold, indicate the segmentation boundaries of the annotators as a group. The amount and salience of peaks can be adjusted by the size of the kernel and the peak threshold. Additionally, the results of the two conditions are compared with segmentation of the digital audio file of the corresponding stimulus obtained with Music Information Retrieval techniques [21].

2.3 Automation

2.3.1 Aim. In this stage an automated system is developed with the goal of predicting human perceived boundaries. The system takes as input the accelerometer or optical motion-capture data from the Multimodal Database. To assess the performance of the system, its output is compared with the corresponding annotations obtained in the Ground Truth stage. The main challenge is to find an appropriate combination of kinetic features that are consistent and distinct for each gesture.

2.3.2 Procedure. For now only accelerometer data from the Wiimote is being considered. This means that data is comprised of tri-axial acceleration at a single moving point. The core of the system was developed by Foote and Cooper [22] for audio and video segmentation. In this study that method is adapted and expanded to be used for segmentation of kinetic data. The procedure involves the choice of multiple *free variables*, which determine the system's performance. In its current state of development, the procedure is as follows:

Step 1) Downsample raw acceleration data from 100 Hz to 10 Hz. This sampling rate is enough to achieve satisfactory results at a lower computational cost than using full resolution.

Step 2) Compute magnitude (Euclidean norm). This is a free variable, here called 'Input Data Type', as either the tri-axial acceleration signal or its magnitude may be used as input for the next step.

Step 3) Compute windowed functions. Statistical functions are computed individually over a sliding window with hop of a single sample. The functions currently used are a subset of functions

evaluated by previous investigation on medical surveying of physical activity using accelerometers [23-24]. To minimize distortion at the borders, the signals are extended at the beginning with the value of the first sample and at the ending with the value of the last sample. The length of each of these extensions is half of the sliding window. The width of the window is a free variable. Also the choice of functions is a free variable.
The functions currently used are the following:

- root mean square
- mean
- standard deviation
- mean absolute deviation
- kurtosis
- skewness
- interquartile range
- centered zero-crossings count

Step 4) Convolve the output of the previous step with a Gaussian kernel and rescale to a range between 0 and 1. The same extension procedure of the previous step is applied to the input of this step before convolution. The window of the kernel is a free variable. If the window length is set to zero, then convolution is not done but only rescaling.

Step 5) Compute a distance matrix of a single function or combined functions. Here the outputs of one or more functions are dimensions of a matrix. Euclidean distance between each point with all the other points is computed to obtain the distance matrix. Additionally, for each function output there is a scaling factor $C\ \{0 < C \leq 1\}$, which determines the contribution (i.e., 'weight') of a function to the computed distances.

Step 6) Compute a Novelty Score by convolving a Gaussian-smoothed Checkerboard Kernel with volume V=1, along the diagonal of the distance matrix. Before performing the convolution, the matrix is extended to half the length of the kernel. The extension section at the beginning is set to the mean value of the section of the kernel that is in the non-extended distance matrix. The same procedure is done at the ending. These extensions with mean values help to reduce the distortion at the beginning and ending. Here the free variable is the length of the kernel.

Step 7) Extract peaks from the novelty score over a threshold. Here the threshold factor $T\ \{0 < T \leq 1\}$ is a free variable. These peaks indicate the computed segmentation boundaries.

Computed segmentation boundaries are then compared with perceived segmentation boundaries (i.e., ground truth) of the corresponding videos, by means of a similarity measure. An earlier version of this measure was used to assess similarity of computed and perceived segmentation boundaries of electroacoustic music [25]. In this study an updated version is used, which is computed as follows:

Step 1) a and b are vectors containing indexes (i.e., time location) of segmentation boundaries, at the downsampled rate. One of them contains perceived boundaries (ground truth) and the other contains computed boundaries (novelty peaks). L is the length of the downsampled data. L_a=L_b. N is the amount of indexes. $N_a \geq N_b$

Step 2) Compute a distance matrix M_{jk} of vectors a and b:

$$M_{jk} = |a_j - b_k|$$

Step 3) Find the minima (m) of rows (r) and columns (c):

$$m_r(j) = \text{argmin} M_{jk} \qquad k \in [1, n]$$
$$m_c(k) = \text{argmin} M_{jk} \qquad j \in [1, n]$$

Step 4) The values of a and b at the intersection minima become vectors a' and b', which are the closest paired elements from a and b.

Step 5) Find the mean distance d from the intersection of minima:

$$d(a, b) = \text{mean}(\ m_r \cap m_c)$$

Step 6) Compute average closeness (c) of paired elements:

$$c = 1 - \frac{d}{L}$$

Step 7) Compute fraction of paired elements:

$$f(a, b) = \frac{N^*}{N''}$$

N^* is the least amount of unique elements and $N''$is the largest amount of unique elements, in either vector a' or b'.

Step 8) Compute similarity (S):

$$S(a, b) = c \cdot f$$

This measure is used because it is a single value that encompasses the hit and misses given by the fraction of paired elements and closeness of those elements. In the context of this study these elements are the time locations of segmentation boundaries. In this way it is not necessary to specify a vicinity of annotated boundaries in which a computed boundary has to be to be considered a match. The method used by MIREX for segmentation of musical audio considered a vicinity of 0.5 s. [26-28]. This is problematic if applied to the segmentation of bodily gestures, as the transition from one gesture to another might take different times at different time-scales. Therefore the vicinity should be dynamically adjusted to those transition times. It is not clear how this can be done, so the similarity measure described above avoids the problem. However, it has the disadvantage that a visual comparison of very high values of S (e.g., over 0.9) might not appear to be reasonably similar and a very small difference in S might be visually perceived as a considerably different. This drawback is a perceptual scaling problem that does not affect the computational effectiveness of the similarity measure, given that precision is set to a substantial amount of decimal places.

The selection of features (i.e., combinations of free variables) that yield results most similar to the ground truth is an optimization problem in a highly dimensional space. The amount of possible combinations might range from very high to astronomical. An extensive search (i.e., by brute force) for the highest S value is therefore impractical. To overcome this difficulty, the solution space is explored by brute-force with constraints that reduce the fee-variable space (see Table 1).

Table 1. Free variables used in the constrained brute-force search.

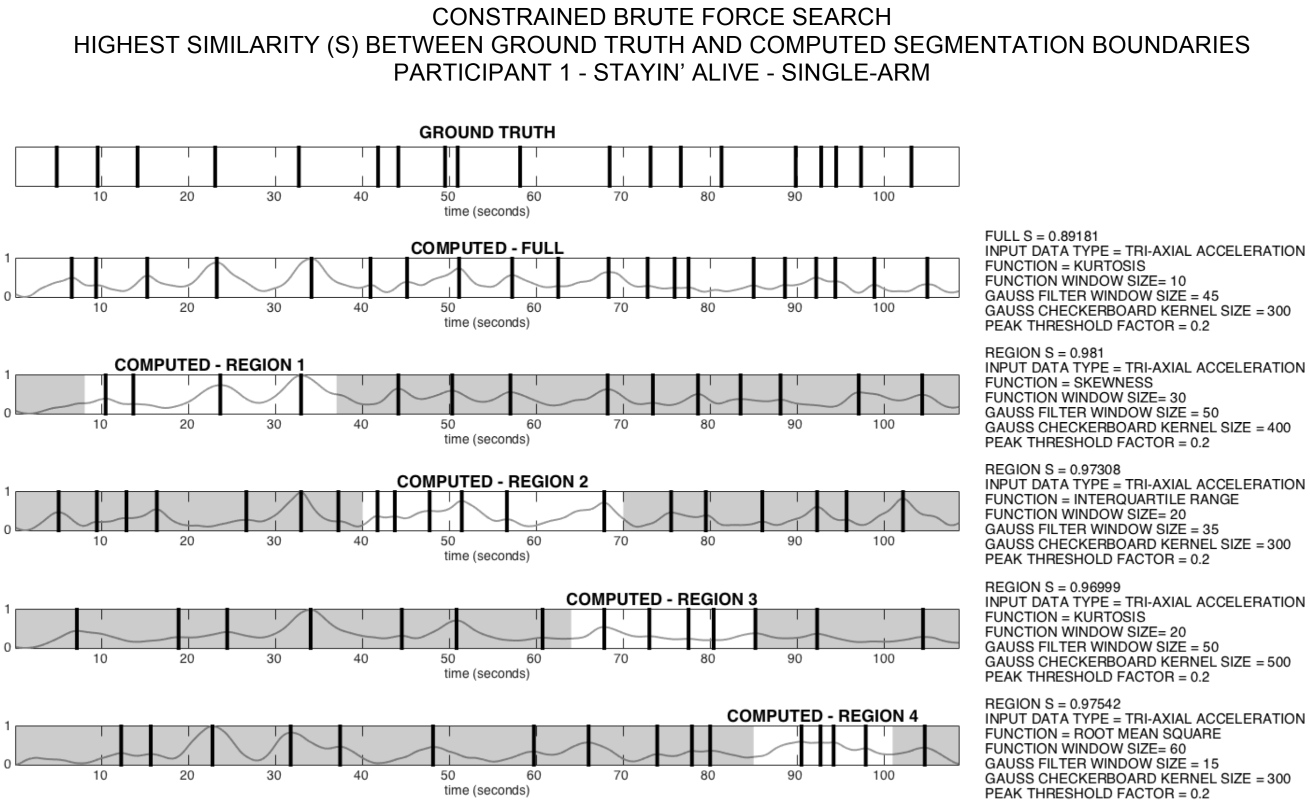
FREE VARIABLES	VALUES
Input Data Type	{Tri-axial Acceleration, Acceleration Magnitude}
Function Window Size (samples)	{10,20,..,60}
Gaussian Filter Window Size (samples)	{5,10,15,..,60}
Gaussian Checkerboard Kernel Size (samples)	{200,300,...,600}
Peak Threshold Factor	{0.1,0.2,...,1}

Then the computed boundaries that have highest similarity with ground truth are manually inspected to find constraints that would facilitate the search by a genetic algorithm. A genetic algorithm has previously been used for a similar problem by an investigation oriented to find the audio features that yield a novelty score that has highest correlation with Kernel Density Estimation of perceived audio segmentation [29].

3 RESULTS

Non-real-time annotation from one participant has been used as ground truth to test the automated segmentation procedure. A brute-force search was done for the highest similarity values between annotated boundaries (ground truth) and computed boundaries, for isolated time regions of the stimulus. This search consisted of 4900 sequences of computed boundaries, produced with single (non-combined) functions and permutations of free variables having the constraints shown in Table 1.

Visual inspection was made of the computed boundaries that have highest similarity with the perceived boundaries. This inspection revealed that while some computed boundaries are remarkably close to perceived boundaries, there are some computed boundaries that do not have any matching perceived boundary or are too far to be considered as matching. However, considering only isolated regions it is possible to observe remarkable closeness between perceived and computed boundaries, only within those regions. Figure 1 shows the highest similarity values within regions, when computed boundaries were compared to perceived boundaries provided by one annotator.



FULL S = 0.89181
INPUT DATA TYPE = TRI-AXIAL ACCELERATION
FUNCTION = KURTOSIS
FUNCTION WINDOW SIZE= 10
GAUSS FILTER WINDOW SIZE = 45
GAUSS CHECKERBOARD KERNEL SIZE = 300
PEAK THRESHOLD FACTOR = 0.2

REGION S = 0.981
INPUT DATA TYPE = TRI-AXIAL ACCELERATION
FUNCTION = SKEWNESS
FUNCTION WINDOW SIZE= 30
GAUSS FILTER WINDOW SIZE = 50
GAUSS CHECKERBOARD KERNEL SIZE = 400
PEAK THRESHOLD FACTOR = 0.2

REGION S = 0.97308
INPUT DATA TYPE = TRI-AXIAL ACCELERATION
FUNCTION = INTERQUARTILE RANGE
FUNCTION WINDOW SIZE= 20
GAUSS FILTER WINDOW SIZE = 35
GAUSS CHECKERBOARD KERNEL SIZE = 300
PEAK THRESHOLD FACTOR = 0.2

REGION S = 0.96999
INPUT DATA TYPE = TRI-AXIAL ACCELERATION
FUNCTION = KURTOSIS
FUNCTION WINDOW SIZE= 20
GAUSS FILTER WINDOW SIZE = 50
GAUSS CHECKERBOARD KERNEL SIZE = 500
PEAK THRESHOLD FACTOR = 0.2

REGION S = 0.97542
INPUT DATA TYPE = TRI-AXIAL ACCELERATION
FUNCTION = ROOT MEAN SQUARE
FUNCTION WINDOW SIZE= 60
GAUSS FILTER WINDOW SIZE = 15
GAUSS CHECKERBOARD KERNEL SIZE = 300
PEAK THRESHOLD FACTOR = 0.2

Figure 1. The top panel shows perceived segmentation boundaries (ground truth). The panels below it show the computed segmentation boundaries and novelty scores that have highest similarity with ground truth, at the non-shaded regions.

4 CONCLUSIONS AND FUTURE WORK

This article has presented an ongoing investigation project towards the modelling of perceived segmentation boundaries of bodily gestures induced by music. Preliminary results have been obtained to predict perceived segmentation of the movement of a person's arm moving to a stimulus (a section of the song Stayin'Alive). Windowed statistical functions were applied to tri-axial accelerometer data from a sensor held by the hand of the moving arm. The functions kurtosis, skewness, interquartile range and root mean square returned very close segmentation boundaries compared to perceived boundaries, considering specific regions of the stimulus. However, no function returned a sequence of boundaries reasonably close to the perceived boundaries considering the full length of the stimulus.

Further work in this project will focus in finding an appropriate combination of functions and their parameters that yield computed boundaries reasonably similar to perceived boundaries. The automatic system will be improved mainly by:

a) Adding Ensemble Mode Decomposition [30] before computing windowed functions.
b) Adding these windowing functions: Energy (mean square), count of local extrema, numerical integration, spectral entropy, spectral centroid and spectral flatness.
c) Combine features after the output of the Gaussian filter and rescaling step.
d) Optimise the search using a genetic algorithm.

The resulting model shall predict bodily gesture boundaries with data from a single point of the body. Nevertheless, the procedure could be used to process multiple points. This method can be combined with an unsupervised machine-learning technique that clusters the segments, completing an automatic unsupervised system for automatic gesture recognition. Such a system will be useful for studying relationships between musical sound and bodily movement. Furthermore, a real-time implementation of this system could be integrated into the design of electronic musical instruments, as a high-level feature for mapping movement to sound. Overall, this automated system provides a cost-effective solution as it can take advantage of cheap accelerometer sensors and computing technology.

REFERENCES

[1] Leman, M. 2008. *Embodied music cognition and mediation technology*. Cambridge, MA: MIT Press.

[2] Clayton, M., and Leante, L. 2011. Imagery, melody and gesture in cross-cultural perspective. In A. Gritten and E. King (Eds.), *New perspectives on music and gesture*, 203. Farnham, England: Ashgate.

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

[4] MacRitchie, J., Buck, B., and Bailey, N. J. 2013. Inferring musical structure through bodily gestures. *Musicae Scientiae*, *17*(1), 86-108.

[5] Schneider, A. 2010. Music and Gestures. In R. I. Godøy and M. Leman (Eds.) *Musical gestures: Sound, movement, and meaning*. Routledge.

[6] Camurri, Antonio, and T. Moeslund. 2010. "Visual gesture recognition." *Musical Gestures-Sound, Movement, and Meaning*.

[7] Kahol, K., Tripathi, P., and Panchanathan, S. 2004. Automated gesture segmentation from dance sequences. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004 Proceedings.* (pp. 883-888). IEEE.

[8] Godøy, R. I., Song, M., Nymoen, K., Haugen, M. R., and Jensenius, A. R. 2016. Exploring Sound-Motion Similarity in Musical Experience. *Journal of New Music Research*, *45*(3), 210-222.

[9] Wanderley, M. M., Vines, B. W., Middleton, N., McKay, C., and Hatch, W. 2005. The musical significance of clarinetists' ancillary gestures: An exploration of the field. *Journal of New Music Research*, *34*(1), 97-113

[10] King, E., and Ginsborg, J. 2011. Gestures and glances: Interactions in ensemble rehearsal. In A. Gritten and E. King (Eds.), *New perspectives on music and gesture*, 177-201. Farnham, England: Ashgate.

[11] Luck, G. 2011. Computational analysis of conductors' temporal gestures. In A. Gritten and E. King (Eds.), *New Perspectives on Music and Gesture*, 159. Farnham, England: Ashgate.

[12] Clayton, Martin, and Laura Leante. 2011. "Imagery, melody and gesture in cross-cultural perspective." *New perspectives on music and gesture*: 203.

[13] Trevarthen, C., Delafield-Butt, J., and Schögler, B. 2011. Psychobiology of Musical Gesture: Innate Rhythm, Harmony and Melody. In A. Gritten and E. King (Eds.), *New perspectives on music and gesture*, 11-43. Farnham, England: Ashgate.

[14] Puckette, M. 1997. Pure Data. *International Computer Music Conference.* Thessaloniki, Greece: Michigan Publishing

[15] Tetarto Hood 2014. Bouzouki Hiphop Instrumental - Rempetila. Retrieved on the 23 August of 2016 from https://www.youtube.com/watch?v=mMWMS6VqXTg

[16] Petzold, C. ca. 1725. Minuet from The Anna Magdalena Bach Notebook, Anh. 114.

[17] Otondo, F. 2008. Ciguri. On *Tutuguri*. Sargasso.

[18] Olsen, K. N., Dean, R. T., and Leung, Y. 2016. What Constitutes a Phrase in Sound-Based Music? A Mixed-Methods Investigation of Perception and Acoustics. *PloS one*, *11*(12): e0167643. https://doi.org/10.1371/journal.pone.0167643

[19] Bee Gees 1977. Stayin' Alive. On *Saturday Night Fever, The Original Movie Soundtrack*. RSO.

[20] Cockos Reaper [Computer software] 2010. Retrieved from http://www.cockos.com/reaper

[21] Lartillot, O., Toiviainen, P., and Eerola, T. 2008. A matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, and R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 261-268). Springer Berlin Heidelberg.

[22] Foote, J. T., and Cooper, M. L. 2003. Media segmentation using self-similarity decomposition. In *Electronic Imaging 2003* (pp. 167-175). International Society for Optics and Photonics.

[23] Lara, O. D., and Labrador, M. A. 2013. A survey on human activity recognition using wearable sensors. *IEEE Communications Surveys and Tutorials*, *15*(3), 1192-1209.

[24] Machado, I. P., Gomes, A. L., Gamboa, H., Paixão, V., and Costa, R. M. 2015. Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization. *Information Processing & Management*, *51*(2), 204-214.

[25] Mendoza, J. I. 2014. *Self-report measurement of segmentation, mimesis and perceived emotions in acousmatic electroacoustic music*. Master's Thesis. University of Jyväskylä. Retrieved from http://urn.fi/URN:NBN:fi:jyu-201406192112

[26] MIREX Structural Segmentation 2016. Retrieved from http://www.music-ir.org/mirex/wiki/2016:Structural_Segmentation

[27] Turnbull, D., Lanckriet, G. R., Pampalk, E., and Goto, M. 2007. A Supervised Approach for Detecting Boundaries in Music Using Difference Features and Boosting. In *ISMIR* (pp. 51-54).

[28] Levy, M., and Sandler, M. 2008. Structural segmentation of musical audio by constrained clustering. *IEEE transactions on audio, speech, and language processing*, *16*(2), 318-326.

[29] Hartmann, M., Lartillot, O., and Toiviainen, P. 2016. Interaction features for prediction of perceptual segmentation: Effects of musicianship and experimental task. *Journal of New Music Research*, 1-19.

[30] Wang, Z., Wu, D., Chen, J., Ghoneim, A., and Hossain, M. A. 2016. A triaxial accelerometer-based human activity recognition via EEMD-based features and game-theory-based feature selection. *IEEE Sensors Journal*, *16*(9), 3198-3207.