Proceedings of the 25th Anniversary Conference of the European Society for the Cognitive Sciences of Music, 31 July-4 August 2017, Ghent, Belgium
Van Dyck, E. (Editor)

Modelling Perceived Segmentation of Bodily Gestures Induced by Music

Juan Ignacio Mendoza[1], Marc Richard Thompson[2]

Department of Music, Art and Culture Studies; University of Jyväskylä, Finland
[1]juigmend@student.jyu.fi, [2]marc.thompson@jyu.fi

ABSTRACT

This article presents an ongoing investigation whose goal is to model perceived segmentation of music-induced bodily gestures. The investigation consists of three stages. The first stage is a database of multimodal recordings of people moving to music. The data of these recordings are video and motion-capture (acceleration and position at several points of the body). In the second stage the videos produced in the first stage are manually segmented. This is regarded as ground truth for the evaluation of the performance of an automatic gesture segmentation system developed in the third stage of the study. This system extracts kinetic features from motion-captured data. Then a novelty score is computed from the kinetic features. The peaks of the novelty score indicate segmentation boundaries. So far 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, different functions of the kinetic features yield considerably similar results between perceived and computed boundaries at isolated regions of the data. This suggests that each of these functions performs best on a specific kind of gesture. Further work will consider evaluating kinetic features composed of combinations of functions.

I. INTRODUCTION

A. Background

In line with the Embodied Music Cognition train of thought (Leman, 2008), 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 (Clayton & Leante, 2011). Quantitative investigation has shown that bodily movement induced by music relates to features of the music, such as periodicity and kinetic energy (Toiviainen, Luck & Thompson, 2010) or tonality (MacRitchie, Buck & Bailey, 2013). The correspondence between music and bodily movement has been studied under the term *musical gesture* (Schneider, 2010). It has been noted that human beings have a remarkable ability to perceive and understand musical gestures by visual observation (Camurri & Moeslund, 2010). The first stage in perception of a gesture is the identification of when and where it starts and ends, a process called *segmentation* (Kahol, Tripathi & Panchanathan, 2004). 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* (Godøy et al., 2016).

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 (Wanderley et al., 2005; King & Ginsborg, 2011; Luck, 2011; Clayton & Leante, 2011; Trevarthen, Delafield-Butt & Schögler, 2011). 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 generalization. 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. The observation of videos should include precise annotations of where gestures occur and a description of them. Such an endeavor 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.

B. Aim

The purpose of this study is to model perceived segmentation of music-induced bodily movement.

II. METHODS

This section presents the three-stage methodology used in this investigation project. Data is periodically added and methods are refined as the investigation advances. What follows corresponds to the state of the project as of April, 2017.

A. Multimodal Database

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

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 either 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 sign a document giving consent to the use of recorded data for research and communication thereof, including audio and video recordings.

3) *Apparatus.* Data is collected at the motion capture laboratory of the Music, Art and Culture Studies department of the University of Jyväskylä. The apparatus is composed by the following measurement processes:

- Optical Motion Capture: An array of 8 Qualisys Oqus cameras track the position of reflective markers attached to a tight suit that the participant wears. Markers are placed on 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 is syncronised to an SMPTE signal emitted by a second computer. Also the Qualisys system sends back a syncronisation audio signal to the second computer.

- Tri-axial accelerometers: The participant wears a Thalmic Myo armband on one forearm beneath the motion capture suit. Also the participant holds a Nintendo Wii-remote ("wiimote") controller with the hand of the arm that wears the Myo armband. Data from these devices are simultaneously recorded at a rate of 100 Hz in the second computer, using software made with the Pure Data programming environment (Puckette, 1997). This software also simultaneously records audio.

- Audio: Stimuli is presented to the participant using two Genelec 8030-A studio loudspeakers with their base at 110 cm. from the floor. A microphone hanging from the ceiling is connected to the audio system of the second computer, which simultaneously records this audio stream (i.e., room audio) in one audio channel and the audio synchronization signal from the optical motion capture system in a second channel. The starting and ending of the audio recording is set to be at the same time of the accelerometer devices' data recording. The audio signal is used later to set a common starting time for accelerometer, optical motion capture and room audio.

- Video: Two small digital cameras (Vivitar DVR-786 and Sony DSC-W610) on flexible portable tripods record video and room audio. They are 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. Later the video stream is synchronized to the accelerometer and optical motion capture using the room audio. This method allows flexibility when positioning the cameras, opposed to having cameras fixed to the wall or mounted on cumbersome rigging.

4) *Stimuli.* The list below shows the excerpts of music that have been used and a brief description that explains the choice.

- "Bouzouki Hiphop" (Tetarto Hood, 2014) 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 4/4. All participants declared to not know this piece.

- "Minuet in G Major" (Petzold, ca. 1725). 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 3/4. All participants declared to know this piece.

- "Ciguri" (Otondo, 2008) from 56 to 180 s. with fade-out the last 5 s. This is an electroacoustic piece that has no perceivable beat that indicates 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"* (Olsen, Dean & Leung, 2016). All participants declared to not know this piece.

- "Stayin' Alive" (Bee Gees, 1977) from the beginning to 108 s. with fade-out the last 2.3 s. Tempo is 104 BPM and meter is 4/4. All participants declared to know this piece.

5) *Procedure.* Data recording is done with one participant at a time. Participants are asked to move spontaneously to the stimulus when it starts sounding through the loudspeakers. They are 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 do ask this question, they are explained that they can dance if they want, otherwise they can move freely.

Each stimulus is presented twice. Participants are 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 Capture systems. The second time participants are asked to hold the Wii-mote with one hand and *dance* only with that arm (this arm is also wearing the Myo armband). In this condition participants are asked to remain at the center of the area facing to a corner of the room. This is done to get in the video recording the most complete visualization of the arm's movement. In this condition participants are allowed to move the rest of the body naturally as long as the previous constraints are not violated. This procedure (called "trial") is 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 are allowed to rest as much as needed between trials.

B. Ground Truth

1) *Aims.* In this stage the videos from the Multimodal Database are manually segmented in two conditions. In each condition the time location of segmentation boundaries is recorded. This task is called *annotation*.

- Real-time annotation: Videos with their corresponding audio are segmented as they are watched.

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

2) *Participants and Stimuli.* Participants of this experiment are called *annotators*, to differentiate them from the participants in data collection for the Multimodal Database. So far two annotators have performed only the Non-real-time task upon the video corresponding to single-arm movement to the "Stayin' Alive" stimulus. These annotators are doctoral students of musicology, one of them the first author of this article. This data has been regarded as preliminary.

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 comma-separated-values text file.

- Non-real-time annotation: A personal computer running the Reaper digital audio editing software (Cockos Reaper, 2010). This system allows video playback at different speeds, scrolling through the video and accurately placing markers, which can be assigned different colors. These markers are exported as a comma-separated-values text file.

4) *Procedure.*

- Real-time annotation: The participant is presented with a video of the Multimodal Database and asked to depress a key when noticing "a change of movement". This wording is meant to indicate a change in bodily gesture without giving an extensive explanation of the concept.

- Non-real-time annotation: The participant is asked to place markers where there is a change of movement. Additionally, the participant is asked to group the annotated markers into larger structures, without further explanation of what this means. To indicate the boundaries of these bigger structures a new set of markers is placed on top of the existing ones, with a different colour.

5) *Data Analysis.* Responses by all participants are summarized 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. Additionally, the digital audio file of the corresponding stimulus is segmented using Music Information Retrieval techniques (Lartillot, Toiviainen & Eerola, 2008).

C. Automation

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. Performance of the system is assessed by comparing its output with the corresponding annotations obtained in the Ground Truth stage. The main challenge is to find an appropriate combination of kinetic features and their parameters that are consistent and distinct for each gesture.

2) *Procedure.* For now only accelerometer data from the Wii-mote is being considered. This means that data consists of tri-axial acceleration of a single moving point. This is a starting point and it is thought that the same method could be applied for data of any of the optical-motion-capture markers individually or in combination. The core of the system was developed by Foote and Cooper (2003) for media segmentation. This method has been adapted and expanded to be used in this investigation for the 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:

- 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.

- 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.

- Compute windowed functions. A set of statistical functions is 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 (Lara, & Labrador, 2013; Machado et al., 2015). 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:

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

- 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.

- 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.

- 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.

- Extract peaks from the novelty score over a threshold. Here the free variable is the factor of the threshold $T\{0 < T \leq 1\}$. 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 (Mendoza, 2014). In this study an updated version is used, which is computed as follows:

- 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. $N_a \geq N_b$

- Compute a distance matrix M_{jk} of vectors a and b:

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

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

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

The values of a and b at the intersection minima become vectors a' and b', the closest paired elements from a and b.

- Find the mean distance d from the intersection of minima:

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

- Compute average closeness (c) of paired elements:

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

- 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'.

- Compute similarity (S):

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

This measure is used because it gives 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, as is the case of the method used by the MIREX structural segmentation evaluation (MIREX Structural Segmentation, 2016; Turnbull et al., 2007; Levy & Sandler, 2008). The MIREX 2016 structural segmentation evaluation considered a vicinity of 0.5 s. This is problematic as the transition from one gesture to another might take different times at different time-scales. Therefore the vicinity should be 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 0.8) 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 only a perceptual scaling problem that does not affect the computational effectiveness of the similarity measure. 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 is astronomical and 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. 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 mixed-integer constrained 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 (Hartmann, Lartillot & Toiviainen, 2016).

III. RESULTS

Data collected so far for the ground truth has been deemed not enough to make the analysis that compares real-time perceived segmentation, non-real-time perceived segmentation and computed audio segmentation. Nonetheless, the available non-real-time grouped annotated boundaries of

single-arm gestures have been used as ground truth in the development of the automated segmentation procedure. A brute-force search was done for the highest similarity values between annotated boundaries given by each annotator (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.

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}

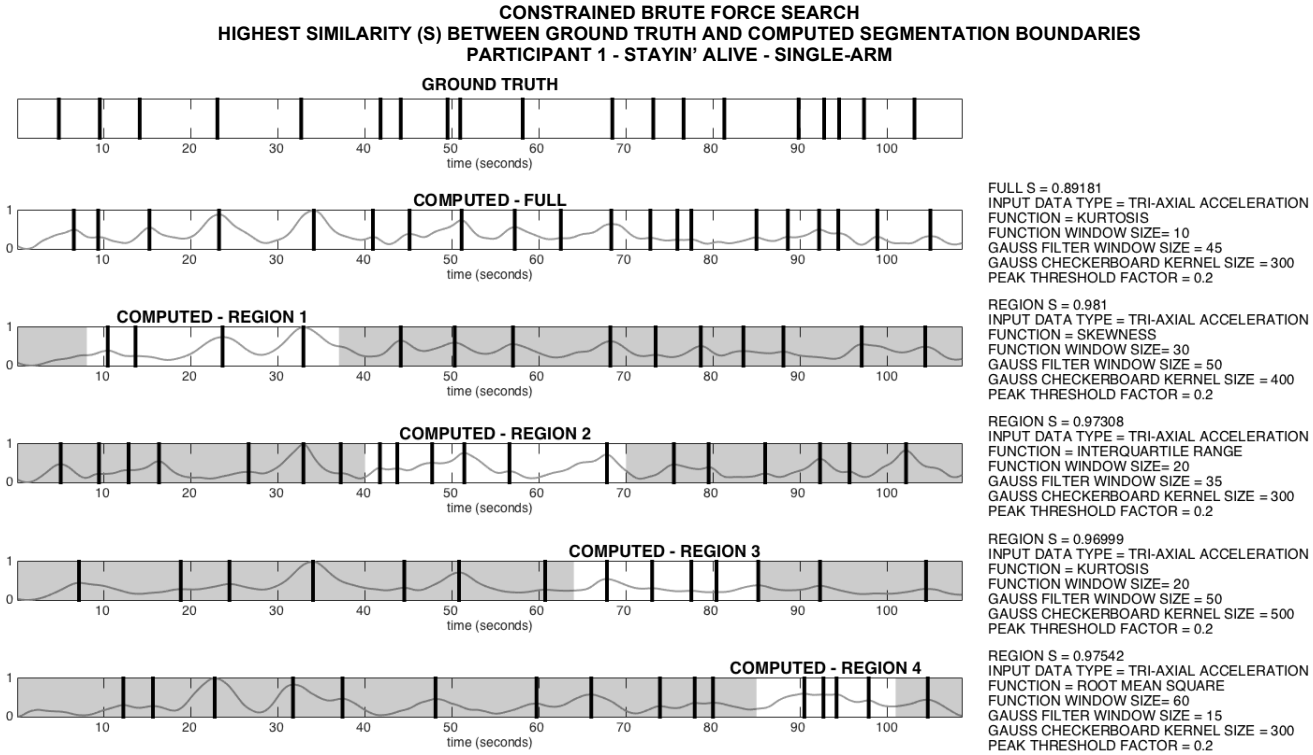

Figure 1. The top panel shows perceived segmentation boundaries (ground truth). The panels below it show the computed segmentation boundaries and novelty scores that have highest similarity with ground truth, at the non-shaded regions.

Visual inspection of the computed boundaries that have highest similarity with the perceived boundaries reveals that while some boundaries are remarkably close, there are some computed boundaries that do not have any matching annotated 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, compared to the annotation of perceived boundaries provided by one annotator (i.e., ground truth).

IV. CONCLUSIONS

This article has presented an ongoing investigation project towards the modeling 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 functions and their parameters that yield computed boundaries reasonably similar to perceived boundaries. Also the collection of more multimodal and perceptual data will contribute to improve the automated system's performance.

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 system 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

Bee Gees (1977). Stayin' Alive. On *Saturday Night Fever, The Original Movie Soundtrack*. RSO.

Clayton, M., & Leante, L. (2011). Imagery, melody and gesture in cross-cultural perspective. In A. Gritten & E. King (Eds.), *New Perspectives on Music and Gesture*, 203. Farnham, England: Ashgate.

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

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

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

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

Kahol, K., Tripathi, P., & 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.

King, E., & Ginsborg, J. (2011). Gestures and glances: Interactions in ensemble rehearsal. In A. Gritten & E. King (Eds.), *New Perspectives on Music and Gesture*, 177-201. Farnham, England: Ashgate.

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

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, Machine Learning and Applications* (pp. 261-268). Berlin, Heidelberg: Springer.

Leman, M. (2008). *Embodied Music Cognition and Mediation Technology*. Cambridge, MA: MIT Press.

Levy, M., & Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 318-326.

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

Machado, I. P., Gomes, A. L., Gamboa, H., Paixão, V., & 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.

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

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

MIREX Structural Segmentation (2016). Retrieved from http://www.music-ir.org/mirex/wiki/2016:Structural_Segmentation

Otondo, F. (2008). Ciguri. On *Tutuguri*. Sargasso.

Olsen, K. N., Dean, R. T., & Leung, Y. (2016). What constitutes a phrase in sound-based music? A mixed-methods investigation of perception and acoustics. *PloS One*, *11*(12): e0167643.

Petzold, C. (ca. 1725). Minuet from The Anna Magdalena Bach Notebook, Anh. 114.

Puckette, M. (1997). Pure Data. *International Computer Music Conference.* Thessaloniki, Greece: Michigan Publishing

Schneider, A. (2010). Music and gestures. In R. I. Godøy & M. Leman (Eds.), *Musical Gestures: Sound, Movement, and Meaning*. London: Routledge.

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

Toiviainen, P., Luck, G., & Thompson, M. R. (2010). Embodied meter: Hierarchical eigenmodes in music-induced movement. *Music Perception*, *28*(1), 59-70.

Trevarthen, C., Delafield-Butt, J., & Schögler, B. (2011). Psychobiology of musical gesture: Innate rhythm, harmony and melody. In A. Gritten & E. King (Eds.), *New Perspectives on Music and Gesture*, 11-43. Farnham, England: Ashgate.

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

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