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

June, 2017

UNIVERSITY OF JYVÄSKYLÄ

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. The correspondence between music and bodily movement has been studied under the term *musical gesture* (Schneider, 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). Modelling perceived segmentation of bodily gestures serves to a better understanding of human perception. Also, it allows a computing machine to segment gestures in the same way as a human being.

AIMS

This research project is aimed to observe, model and predict the perceived segmentation of bodily gestures induced by music. The project is composed by three stages: Building a multimodal database, the collection of ground truth and the development of an automatic system that performs segmentation of bodily gesture.

MULTIMODAL DATABASE

Naive participants spontaneously move to music excerpts between 40 and 190 s. For each music excerpt they are recorded in two conditions: free movement and 'dancing with one arm'. In the two conditions they wear an optical motion capture suit with reflective markers. In the second condition they hold an accelerometer with the hand of the arm that moves.

Recorded data modalities are:

- Optical motion capture (3D position)
- Accelerometer
- Video
- Audio (music excerpts)

Free movement

Dancing with one arm

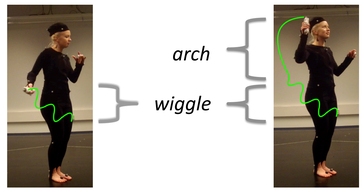
GROUND TRUTH

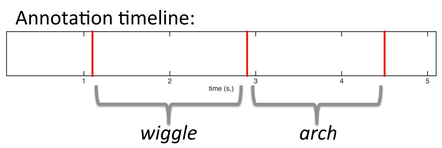
Semi-expert annotators watch videos of the database and indicate where there is a change of gesture. Gesture is defined as a new pattern. Thus, if movement repeats consecutively, no indication shall be made until a new pattern is perceived.

The annotation task is done in two conditions:

Real-time annotation:
Video is presented with audio. While watching and listening, the annotator press a button where a change in gesture is perceived.

Non-real-time annotation:
Video is presented without audio and the annotator can scroll back and forward to accurately place a marker at a perceived change of gesture.

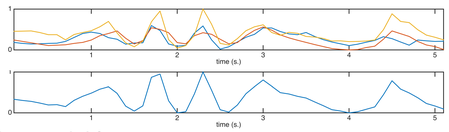




AUTOMATION

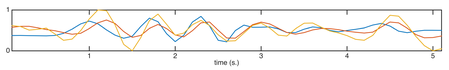
An automatic system to predict annotators' segmentation that takes motion data at one point as input. For example, the input can be acceleration at the hand of the moving arm from the condition 'dancing with one arm'. The system is an adaptation and extension of a method for data segmentation described by Foote and Cooper (2003) and used for audio and video segmentation. The main challenge is to find a combination of kinetic features that are consistent and distinct for each gesture (see Wang et al., 2016). A combination of kinetic features is defined as a combination of free variables as shown below.

Step 1) Downsample and Compute Magnitude. (e.g., acceleration at 100Hz to 10Hz)



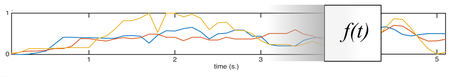
Free variable:
- Dimensions {XYZ, magnitude}

Step 2) Empirical Mode Decomposition



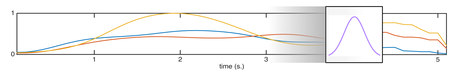
Free variables:
- EMD method {EMD, Ensemble EMD, no EMD}
- Intrinsic Mode Function

Step 3) Windowed Function



Free variables:
- Function {energy, RMS, mean, standard deviation, mean absolute deviation, kurtosis, skewness, interquartile range, centered zero-crossings, local extrema, integration, spectral entropy, spectral centroid, spectral flatness}
- Window length

Step 4) Gaussian Smoothing Filter



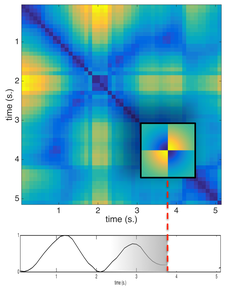
Free variable:
- Window length (0 = no filter)

Step 5) Functions Combination

Free variables:
- Arbitrary amount of outputs from Step 4
- Scaling factors

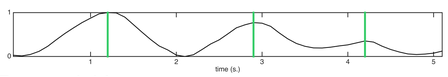
Step 6) Novelty Score

Convolve a Gaussian checkerboard kernel along the diagonal of a self-similarity distance matrix from the output of step 5.



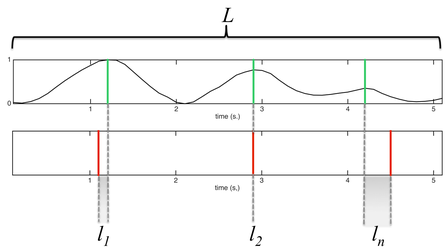
Free variable:
- Gaussian checkerboard kernel length

Step 7) Novelty Peaks



Free variable:
- Threshold

Step 8) Comparison of Result with Ground Truth



Average distance (d):

$$d = mean(l_1 \ldots l_n)$$

Average closeness (c) :

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

Fraction of paired elements:

$$f = \frac{N*}{N''}$$

$N*$ is the least amount of unique elements and N'' is the largest amount of unique elements, in either vector (Result or Ground Truth).

Similarity (S):

$$S = c \cdot f$$

RESULTS AND FUTURE WORK

Preliminary results have been obtained by a constrained brute-force search for greatest similarity between computed and perceived segmentation boundaries. The latter corresponds to responses of one annotator, for a video of one participant in the condition 'dancing with one arm' to one musical excerpt. The search was made without combining functions (Step 5), revealing that kurtosis, skewness, interquartile range and root mean square (RMS) are satisfactory predictors for isolated regions of boundaries. The closest similarity for the full sequence of boundaries was still not satisfactory. This suggest that each evaluated function characterise a specific kind of gesture.

Future work will evaluate the automated system using more of the collected motion data and ground truth data. The search will comprise combinations of functions and will be optimised by means of a genetic algorithm, as done in research on auditory segmentation (Hartmann, Lartillot & Toiviainen, 2016).

REFERENCES

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.

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.

Leman, M. (2008). *Embodied music cognition and mediation technology*. Cambridge, MA: MIT Press.

Schneider, A. (2010). Music and Gestures. In R. I. Godøy & M. Leman (Eds.) *Musical gestures: Sound, movement, and meaning.* New York, NY: Routledge.

Wang, Z., Wu, D., Chen, J., Ghoneim, A., & 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.