Automatic gesture segmentation of a musical performance using a cumulative dissimilarity measure

Hugo B. de Paula[1], Euler C.F. Teixeira[2], Raphael Borges[2], Maurício A. Loureiro[2]

[1]Instituto de Informática – Pontifícia Universidade Católica de Minas Gerais (PUC Minas)

[2]Center for Research on Musical Gesture and Expression (CEGEM)
Universidade Federal de Minas Gerais,

hugo@pucminas.br, {mauricioloureiro, eulercft}@ufmg.br

Abstract. *This paper presents a methodology for gesture segmentation of musical performances. It uses a novel approach based on cumulative dissimilarity (CDIS) to determine the velocity of the movement, as well as the local displacement of the musician during a musical gesture. A segmentation algorithm is applied along the dissimilarity vector to provide information about musical gestures. These movement segments are then compared to previous studies of movement analysis and correlated to onsets, and velocity data. During experiments, acoustical and motion data of clarinet performances were analyzed. Results obtained from motion analysis of the clarinet bell showed that, despite CDIS being a 1-dimensional descriptor, it retains information about movement direction and shape and is correlated to expressive musical content.*

1. Introduction

According to Mitra and Acharya, (2007), *"gestures are expressive, meaningful body motions involving physical movements of the fingers, hands, arms, head, face, or body with the intent of conveying meaningful information or interacting with the environment"*. Gestures are studied in many different fields. In music, effective gestures are those that play a direct role in the production of sound, while ancillary gestures refer to body movements that are not involved in production of sound (Wanderley, 1999). During clarinet performances, it has been shown that ancillary gestures are not only frequent, but correlate with several audio cues that suggest its relation to musical expressive content (Teixeira et al., 2014).

Many studies have tackled expressiveness throughout analysis of acoustical data, such as energy envelope, pitch, note onsets and offsets, attack duration and spectral centroid. Some examples of such an approach can be seen in Campolina et al. (2009), Maestre and Gómez (2005) and De Poli et al. (2004). These studies have shown that musicians make use of small deviations, regarding note durations, articulations, intensity, pitch and timbre, in order to convey their musical intentions (Juslin, 2000).

Fenza et al. (2005) used a 3 layer motion processing to segment movements based on Quantity of Motion, and investigated gesture mapping in a 3D expressive space. Camurri et al. (2004) presented the gesture as a conveyor of information related to the emotional domain in dance or music performances. They identified descriptors related to specific features of trajectory patterns, such as angularity, spatial occupation, symmetry, and others. Several other fields of research have successfully employed multi-modal analysis to study the coupling between the acoustical and visual components, such as in speech analysis (Barbosa et al., 2008) and dance (Naveda and Leman, 2009).

Wanderley et al. (2005) and Vines et al. (2006) showed that body movements are part of a performance, and some gestures are not produced for the purpose of sound generation, which they designated as ancillary gestures. They also observed that these gestures were closely related to the musician's expressive intentions in a particular performance.

Recent studies have searched for methods to extract and analyze these movements in detail. Teixeira et al. (2010) proposes a tangential velocity based model to represent, segment and analyze expressive movements based on local gestural parameters. More recently, Teixeira et al. (2014) presented recurrent sequences of clarinet gestures in regions of the excerpts that related to musical structure. Teixeira et al. (2015) showed direct correlations between the recurrence pattern of clarinetists' ancillary movements and expressive timing associated with melodic phrasing and harmonic transitions. Several studies search for methodologies for segmenting musical gestures, in order to stablish correspondence to musical structure. However, there are still no basic units established to segment body movements of music performances, unlike the audio data.

Caramiaux et al. (2012) used Hidden Markov Models (HMM) to segment movements into primitive shapes, selected from a base dictionary. Desmet, Nijs, Demey, Lesaffre, Martens and Leman (2012) proposed another statistical model for body movement segmentation and pointed to subjective links between these segments and the musical score. Rasamimanana (2012) defined a conceptual framework considering performer–instrument relationships that can provide ground to model expressive gestures using a space of possibilities.

In this study, we present a procedure for automatic movement segmentation in order to extract clarinet players' physical gestures during performances of pieces of the classical repertoire. We aim at evaluating gesture segmentation of several clarinetists, and with the purpose of relating their gestural patterns to the music structure.

2. Methodology

The main goal of this study is to evaluate a new method for movement segmentation of musical performances by a group of clarinet players contained in the same dataset presented by Teixeira et al. (2014). Figure 1 presents a schematic diagram of the procedure. As seen in the diagram, acoustical and motion data were obtained from clarinet performances. Several features were then calculated based on these data. After calculation of the cumulative dissimilarity vector, several procedures such as thresholding and filtering are applied to segment the musicians' body movement into gestures. This section presents how these features are obtained and combined to allow a fully automatic detection of musical gestures.

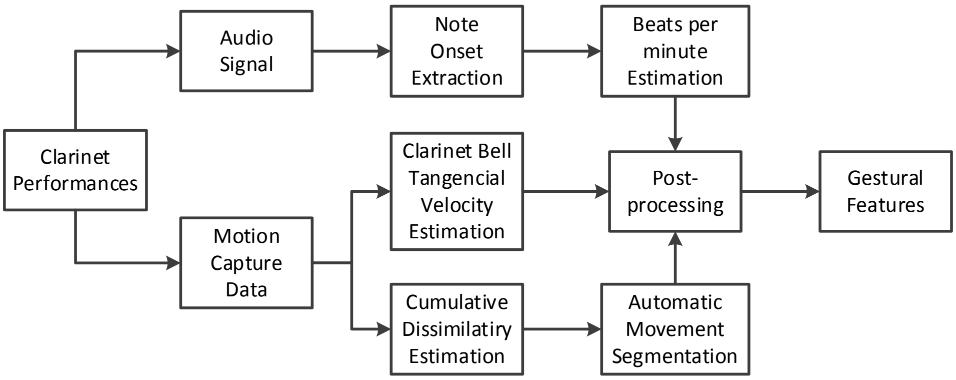


Figure 1. Movement segmentation procedure diagram.

Ten professional clarinet players performed a short excerpt of six bars extracted from the first movement of the Quintet for Clarinet and Strings in A Major, Kv 581 by W.A. Mozart (Figure 2). The musicians were asked to play according to two distinct experimental conditions: expressive performances as in a real concert situation (standard condition) and following a metronome, set to a tempo estimated from his/her previous standard performances (metronome condition). Each of the ten clarinet players performed this excerpt six times without accompaniment, three performances for each of the two experimental conditions. Motion tracking was done with high-end 3D motion capture devices, the NDI Optotrak Certus and the NDI Optotrak 3020. Motion capture markers were placed on their bodies and instruments. Detailed information and pictures about experiment setup, such as marker positions can be obtained in Teixeira et al. (2014).

Motion was captured at a sampling rate of 100 frames per second. Audio was recorded synchronously at a sampling rate of 44.1 kHz using a condenser microphone positioned one meter away from the clarinet. Pitch and energy envelope curves were extracted, and from these all note onsets and offsets were detected using the system described in Campolina, Loureiro and Mota (2009).

Figure 2. Main theme from first movement of Mozart's Quintet for Clarinet and Strings in A Major, Kv 581, performed in the first experiment.

2.1. Movement representation

Movement analysis in this study is based on the clarinet bell movement. The clarinet movement has been the object of previous studies (Caramiaux, Wanderley and Bevilacqua, 2012; Wanderley, 2002; Wanderley, Vines, Middleton, McKay and Hatch, 2005) and it is believed to be an important indicator of expressive movements made by the musician. The segmentation methods presented by these approaches are mainly based on mapping or recognition from a dictionary of shapes (Caramiaux et al., 2012; Vatavu et al., 2009), or low level features such as velocity or Quantity of Motion (Fenza et al., 2005).

The present study proposes a novel approach that abstracts from shape mapping but considers movement direction in a cumulative dissimilarity measure of movement which might be more effective for stablishing correspondence between gesture and audio musical content. Previous studies suggested the clarinet bell movement as an indicator of expressive movements of clarinet performance (Wanderley, 2002; Wanderley et. al., 2005). Thus, the movement of the clarinet bell was taken relative to a coordinate system, located in the center of the Optotrak tracker. With a static reference, instead of a dynamic one, such as the mouthpiece, the clarinet bell movement incorporates any movement performed by the musicians with their feet, knees, torso, neck and arms, and can thus be seen as a general indicator of the players' movements. Optical flow techniques were already used to define a general motion indicator (Barbosa, Yehia and Vatikiotis-Bateson, 2008). The analysis of a single tracking point, the clarinet bell, offered more precision and allowed to include its 3D trajectory, in order to define recurrent gestures and many associated gestural features.

The clarinet bell's tridimensional motion is represented by a matrix of dimension *T* x *M*, where *T* represents the number of time frames and *M* represents the 3-dimensional markers' positions. We used three markers (top, left and right) to capture bell's translation and rotation in X, Y, and Z coordinates (Figure 3). Previous analyses showed that these markers are highly correlated and do not add in precision to the expressive gesture segmentation, probably due to the fact that the extent of rotation is small compared to the extent of translation. X and Y axes presented a correlation coefficient of 99.6% among markers and Z axis 92%. Correlation index of combined X and Y axes ranged between ±4%.

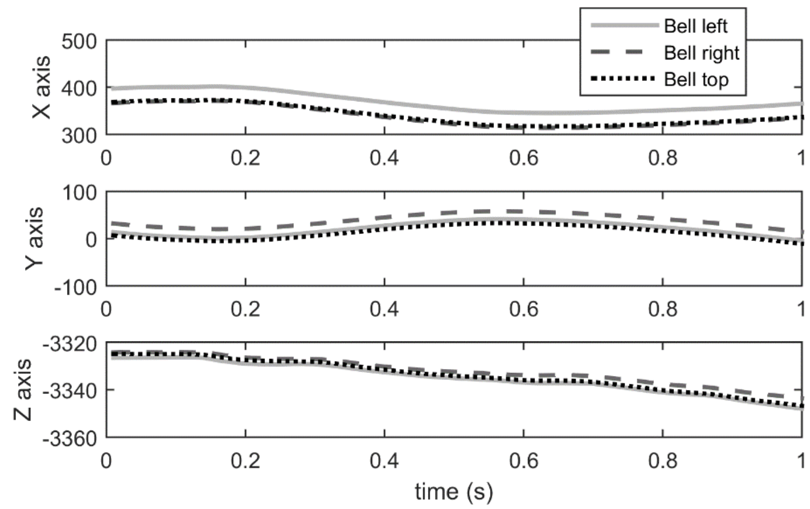


Figure 3. Trajectories of clarinet bell's top, left and right markers.

2.2. Movement processing

In order to analyze the evolution of the clarinet bell's tridimensional motion in conjunction with the acoustical data, we need an effective scalar representation of the motion data in time. Fenza et al. (2005) call these low-level features related to motion description. A simple solution presented by Teixeira, Loureiro and Yehia (2010) is to use the tangential velocity of the clarinet bell marker's trajectory, estimated by the Euclidian distance between subsequent samples of the positions of this marker. Although this unidimensional parameter captures a large amount of information from the musician's movements (Teixeira, Loureiro and Yehia, 2010), it does not contain information about the direction or the shape of the movement. For instance, the difference between an up and down movement and a long ascending trajectory made by the clarinet player would not be revealed solely by the tangential velocity. The same problem would happen with the low-level feature Quantity of Motion (QoM) from Fenza et al. (2005). However, Fenza and colleagues combine other low-level features to help describe shape and direction of movements, which are: Contraction Index (CI), Movement Length (ML), Straight trajectory Length (SL), and Directness Index (DI).

While the estimation of note onsets and offsets from pitch and energy envelope curves extracted from the audio signal, allows the segmentation of the acoustical data into musical notes and phrases, there are no basic units established to segment the movements into gestures. It is possible to develop a procedure to segment the movement data accordingly, by subdividing those movements into representative segments, based on their geometrical and temporal attributes.

2.3. r-Cumulative dissimilarity (r-CDIS)

To measure the amount of movement within a timeframe we present a new strategy based on a cumulative dissimilarity measure. Patrocínio Jr. et al. (2010) originally presented cumulative dissimilarity to detect gradual transitions on video sequences. The main idea behind cumulative dissimilarity is that small movements around the same region in the feature space are meaningless, but small movements consistently driving towards a new region in that space might mean a transition between different scenes.

Given a vector of elements (v) and a $2r$-sized sliding window centered at position k, the *r-cumulative dissimilarity* $r\text{-CDIS}_k$ can be calculated as:

$$r\text{-CDIS}_k = \sum_{i=k-r+1}^{k} \sum_{j=k+1}^{k+r} \text{DIS}(v_i, v_j),$$

where $\text{DIS}(v_i, v_j)$ is a dissimilarity measure between vector elements v_i and v_j.

Several different dissimilarity measures can be employed to calculate cumulative distance curve. In this work, Euclidean distance will be used as a dissimilarity measure calculated as:

$$\text{DIS}(v_i, v_j) = \sum_{m=1}^{M} (v_{j,m} - v_{i,m})^2$$

where M is the number of marker positions as stated in Section 2.1.

The main advantage of using a cumulative dissimilarity approach is that it can detect long gradual displacements during a fixed period and is resilient to small constant movements that go back and forth around the same region. Figure 4 clarifies this character. Figure 4 (upper line) presents four of the ten common gestures defined by Vatavu, Grisoni and Pentiuc (2009). All movements were designed to have exactly the same overall length (100 points) and same tangential velocity. The only parameter that changes among them is the direction of the movement. Lower plots show tangential velocity and cumulative dissimilarity measures for all the four movements. Cumulative dissimilarity was calculated based on a 20-point sliding window and first and last points were replicated to avoid border effects.

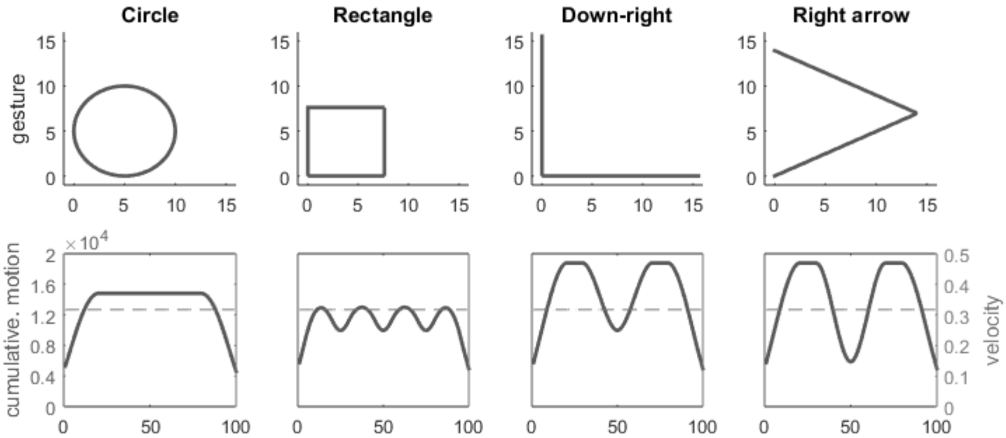


Figure 4. Lower panel show cumulative dissimilarity (10-CDIS) and tangential velocity (dashed line) of common gestures (top panel). Adapted from Vatavu, Grisoni and Pentiuc (2009), Fig. 7, p. 9.

It can be seen in Figure 4 that longer movements result in higher cumulative dissimilarity on average. This feature also captures the change in direction: the higher the change in direction the lower the cumulative dissimilarity value. These properties allow for better characterization of movements and provide a base for gesture segmentation.

2.4. Movement segmentation

Movement segmentation is performed in four steps. Firstly, r-CDIS_k of Euclidean distance is estimated in every point of the T x M matrix of markers positions. The result is a vector of length T of r-CDIS_k, for $k \in [1..\mathrm{T}]$. The second step is the thresholding: every point in vector below a certain threshold is set to zero. This step is important to eliminate small movements that are constrained in space, especially those movements from balance and posture control (Winter, 1995). Afterwards, the 1st derivative is calculated and used to estimate limits, considered as movements onsets and offsets, analogous to notes onsets and offsets, upon which movement lengths are estimated. A final step filters movement segments that were considered too short according to a pre-defined note duration related to the estimated musical tempo in beats per minute (BPM).

3. Experimental results

Table 1 summarizes the recording dataset of 60 performances, as previously described, of an average length of 18.7 seconds, summing up almost 23 minutes of recorded music.

Table 1. Description of the recording dataset.

Recordings		Recording length (seconds)		Tempo (BPM)	
# Clarinetists	10	Max	22.1	Max	128.6
# free expressive performances	3	Min	16.5	Min	85.6
# metronome restricted performance	3	Mean	18.7	Mean	113.9
Total # of recordings	60	Total	1,126.6		

Figure 5 shows the thresholding procedure. Cumulative dissimilarity vector is represented by a 50-CDIS ($r = 50$), which means that a window of length 100 was used. Since the motion capture sampling rate is 100 samples/second, window length corresponded to exactly 1 second. According to Patrocínio Jr. et al. (2010), r-CDIS is specially good in finding segments of length $2r$. Larger values of r flattens the CDIS curve, making it harder to segment and losing precision in gesture position estimation, since it might sum up more than one gesture altogether. The threshold value used in all 50-CDIS curves was 8%, which was empirically obtained. Larger threshold values reduce the average length of motion segments, eliminating small segments and increasing the number of medium sized segments. On the other hand, smaller threshold values produce several meaningless micro-movements, and extra long macro-movements.

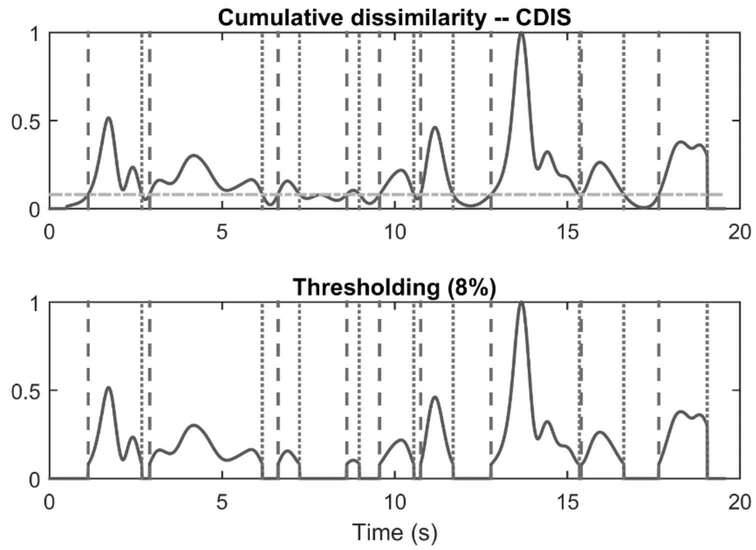


Figure 5. Thresholding procedure. Horizontal dashed line represents threshold (8%). Vertical lines represent onsets (dashed lines) and offsets (dotted lines). Continuous line represents normalized 50-CDIS as a function of time (s).

After the thresholding, CDIS vector was segmented and movement onsets and offsets were determined. First two rows of Table 2 show the results obtained from segmentation. The smallest detected movement has only 30 milliseconds, which is shorter than a sixty-fourth note in an *Allegro* tempo of 120 BPM. The largest segment was more than 15 seconds long. At first, we will consider that long movements occur when the musician keeps moving continuously and we will not force any further segmentation. However, we considered that too short segments do not correlate to music expressiveness and therefore should be removed by considering that ancillary gestures shorter than a sixteenth note are not long enough to convey expressive content.

Table 2. Segmentation and filtering results

Variables	Minimum value	Maximum value	Mean value
Duration all segments (s)	0.03	15.53	1.79
# Segments / performance	1	11	6.9
Duration long segments (s)	0.13	15.53	1.81
# Long segments / performance	1	11	6.8

The cut-off length was estimated by the duration of the sixteenth note for each performance tempo in BPM, estimated from note onsets. Figure 6 shows a histogram of the estimated tempi. One performer played the excerpt in tempo *Andante*, much slower than the expected *Allegro*.

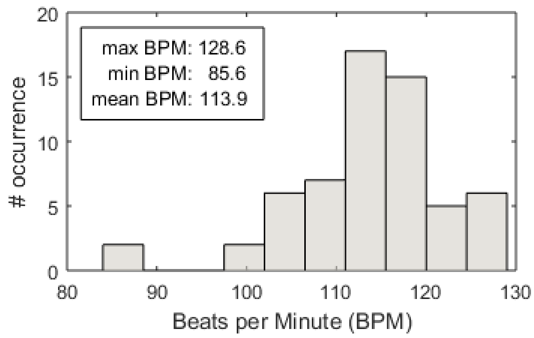


Figure 6. Histogram of recording tempos.

Figure 7 shows histograms of gesture duration. Upper plot presents unfiltered gesture segments and lower plot segments after removal of short segments. It can be clearly seen that there are 5 large segments above 9 seconds long, in a total of 413 detected gestures. 127 segments (30%) have duration between 1 and 2 seconds. Lower plot shows results after filtering. Only 5 segments were removed and length distribution remained almost the same.

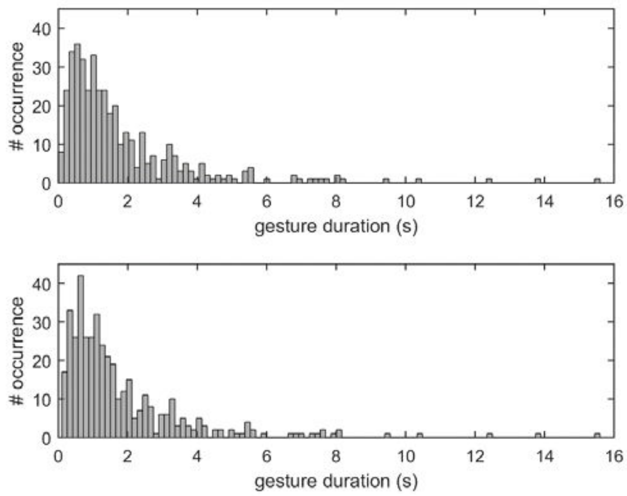


Figure 7. Histogram of gesture durations. Upper plot shows 413 unfiltered segments. Lower plot shows 408 BPM-filtered segments.

Figure 8 shows a sample segmentation of three expressive performances played by the same clarinetist. This clarinetist produced the largest (upper plot) and the smallest (middle plot) segments compared to all recordings. It shows how sensitive to the thresholding the segmentation procedure is. However, Figure 9 show other three expressive performances played by another clarinetist. It can be seen that this clarinetist has a steady behavior that was perfectly captured by the segmentation. In this case, all performances were segmented in four parts (the fifth part in the middle plot is actually silence, after the musician has finished playing). The four segments were also proportionally alike.

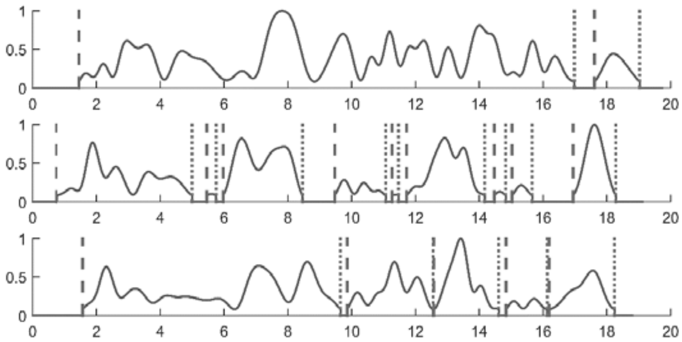


Figure 8. Sample segmentation of three free expressive performances from the same performer showing limitations of the model. Vertical lines represent movement onsets (dashed lines) and offsets (dotted lines). Continuous line represents normalized 50-CDIS as a function of time (s).

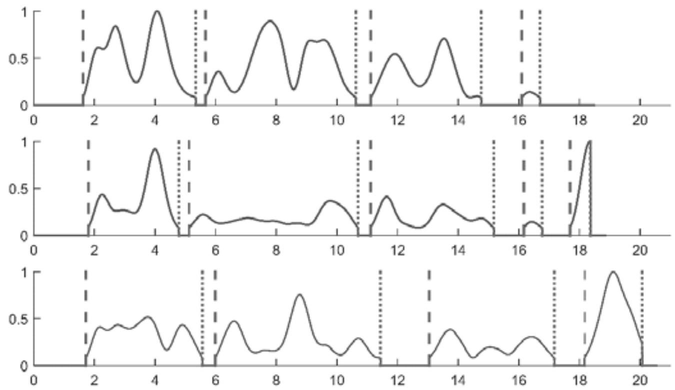
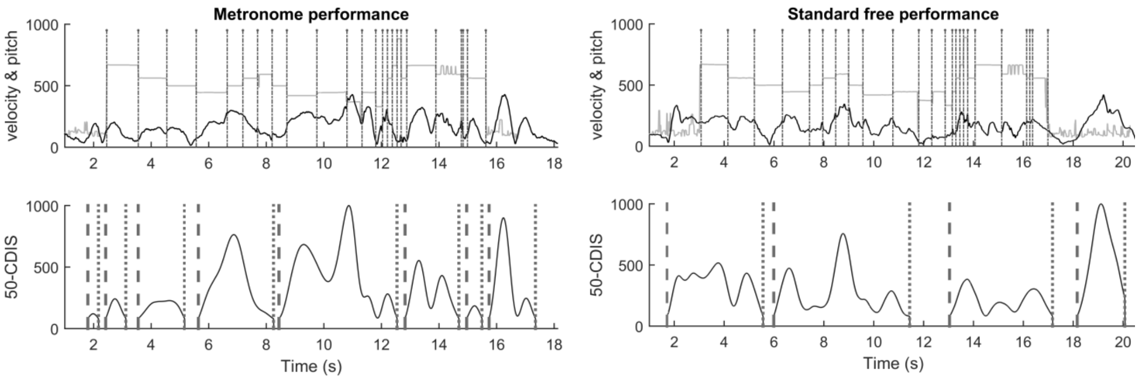


Figure 9. Sample segmentation of three free expressive performances from a steady performer. Vertical lines represent movement onsets (dashed lines) and offsets (dotted lines). Continuous line represents normalized 50-CDIS as a function of time (s).

Figure 10 compares gesture segments to note onsets and tangential velocity. Pitch information is also presented to provide a better understanding of the performance. Figure 10 (a) shows the results obtained for a performance constrained by the metronome. It is important to note the correlation between tangential velocity and cumulative dissimilarity. However, CDIS curve eliminate local information about small movements and outstands the extent of the gesture as related to bell displacement. Visual analysis of that Figure also suggests correlation between gesture onsets and note onsets. Usually, gesture starts close to a note onset, but end earlier. An amplitude envelope of the note could possibly show some correlation to the decay of the gesture. It is not our goal yet to analyze the correlation between the gesture amplitude and musical content, but this must be further investigated.

Figure 10 (b) shows the results obtained for an expressive performance. It is interesting to note that the gestures obtained by this segmentation procedure show correlation to some points of the music structure, especially to ascending and descending sequences. Usually, the gesture starts one note before the beginning of the sequence suggesting that such musical movements are being anticipated by the musician. The third gesture, between seconds 13 and 17 is almost synchronized to the last *legato* ascending arpeggio that leads to the perfect cadence that closes the sentence.



(a) (b)

Figure 10. Gesture segmentation using 50-CDIS and 8% threshold compared to note onsets and tangential velocity. Vertical lines represent movement onsets (dashed lines) and offsets (dotted lines).

All CDIS plots were normalized in order to give a better understanding of the segmentation process, including the thresholding procedure. However, gesture amplitude may play an important role in future gesture analyzes.

4. Conclusion and Further Works

The main goal of this work was to provide a new way for automatic segmentation of body movement of musical performances into musical gestures. A method based on cumulative dissimilarity (CDIS) was presented. Sixty clarinet recordings were analyzed using 50-CDIS and thresholding of 8%. BPM based filtering allowed removal of meaningless short movements. The main contribution of our work is the application of a simple distance measure as a descriptor of musical gestures. Despite CDIS being a 1-dimensional descriptor, it retained some information about movement direction and shape in its value. Experimental results showed some correlation between gestures and some musical passages, suggesting that these gestures might contribute to convey performers' expressive intention.

However, segmentation results can be highly dependent on CDIS radius and threshold values. So, as a future work, we plan to investigate thoroughly the effect of these parameters in gesture segmentation and devise an automatic approach to parameter estimation. We also intend to further investigate the importance of gesture amplitude and its correlation to expressive musical content.

5. Acknowledgements

The authors would like to thank the Brazilian research agencies CNPq, CAPES and FAPEMIG for funding this study.

6. References

Barbosa, A., Yehia, H., and Vatikiotis-Bateson, E. (2008). "Temporal characterization of auditory-visual coupling in speech". Proceedings of Meetings on Acoustics, 1, pp. 1–14.

Campolina, T., Loureiro, M., and Mota, D. (2009). "Expan: a tool for musical expressiveness analysis". In L. Nevada (Ed.), Proceedings of the 2nd International Conference of Students of Systematic Musicology. Ghent: IPEM – Institute for Psychoacoustics and Electronic Music. pp. 24–27.

Camurri, A., Mazzarino, B., Ricchetti, M., Timmers, R., and Volpe, G. (2004). "Multimodal analysis of expressive gesture in music and dance performances". In A. Camurri, & G. Volpe (Eds.), Gesture-based communication in human-computer interaction. Berlin: Springer. pp. 20–39.

Caramiaux, B., Wanderley, M., and Bevilacqua, F. (2012). "Segmenting and parsing instrumentalists' gestures". Journal of New Music Research, 41 (1), pp. 13–29.

De Poli, G., Canazza, S., Drioli, C., Rodà, A., and Vidolin, A. (2004). "Modeling and control of expressiveness in music performance". Proceedings of the IEEE, 92 (4), pp. 686–701.

Desmet, F., Nijs, L., Demey, M., Lesaffre, M., Martens, J., and Leman, M. (2012). "Assessing a clarinet player's performer gestures in relation to locally intended musical targets". Journal of New Music Research, 41 (1), pp. 31–48.

Fenza, D., Mion, L., Canazza, S. and Rodà, A. (2005). "Physical movement and musical gestures: A multilevel mapping strategy". Proceedings of Sound and Music Computing. pp. 1–6.

Juslin, P. (2000). "Cue utilization in communication of emotion in music performance: Relating performance to perception". Journal of Experimental Psychology: Human Perception and Performance, 26, pp. 1797–1813.

Maestre, E. and Gómez, E. (2005). "Automatic characterization of dynamics and articulation of expressive monophonic recordings". Audio Engineering Society 118th Convention Paper 8. New York: Audio Engineering Society (AES).

Mitra, S., and Acharya, T. (2007). "Gesture recognition: A survey". IEEE Transactions on Systems, Man and Cybernetics, 37 (3), pp. 311–324.

Naveda, L., and Leman, M. (2009). "A cross-modal heuristic for periodic pattern analysis of samba music and dance". Journal of New Music Research, 38 (3), pp. 255–283.

Patrocínio Jr., Z. K. G., Guimarães, S. J. F. Silva, H. B. and Souza, K. J. F. (2010) "A unified transition detection based on bipartite graph matching approach". In: BLOCH, I.; CESAR, R. M. (Ed.). Iberoamerican Congress on Pattern Recognition. [S.l.]: Springer. (Lecture Notes in Computer Science, v. 6419). ISBN 978-3-642-16686-0. pp. 184–192.

Rasamimanana, N. (2012). "Towards a conceptual framework for exploring and modelling expressive musical gestures". Journal of New Music Research, 41 (1), pp. 3–12.

Rodger, M. W. M., O'Modhrain, S., and Craig, C. M. (2013). "Temporal guidance of musicians' performance movement is an acquired skill". Experimental brain research. 226 (2), pp. 221–30.

Teixeira, E., Loureiro, M. and Yehia, H. (2010). “Methodological aspects of the research in musical expressiveness based on corporal movement information”. Unpublished Report. Retrieved from http://hal.archives-ouvertes.fr/hal-00611660/en/

Teixeira, E. C.F., Loureiro, M. A., Wanderley, M. M. and Yehia, H. C. (2014). “Motion Analysis of Clarinet Performers”. Journal of New Music Research, 44 DOI: 10.1080/09298215.2014.925939. pp. 97–111.

Teixeira, E. C. F., Yehia, H. C. and Loureiro, M. A. (2015). “Relating movement recurrence and expressive timing patterns in music performances”. The Journal of the Acoustical Society of America, 138, EL212-EL216, DOI:10.1121/1.4929621.

Vatavu, R.-D., Grisoni, L., and Pentiuc, S.-G. (2009). “Gesture Recognition Based on Elastic Deformation Energies”. In Gesture-Based Human-Computer Interaction and Simulation, Miguel Sales Dias, Sylvie Gibet, Marcelo M. Wanderley, and Rafael Bastos (Eds.). Lecture Notes In Artificial Intelligence, Vol. 5085. Springer-Verlag, Berlin, Heidelberg DOI: 10.1007/978-3-540-92865-2_1. pp. 1–12.

Vines, B., Dalca, I. and Wanderley, M. (2006). “Variation in expressive physical gestures of clarinetists”. Baroni, M., Addessi, A.R., Caterina, R. & Costa, M. eds. Proceedings of the 9th International Conference on Music Perception and Cognition. New York: The Society for Music Perception & Cognition (SMPC) and Jyväskylä: European Society for the Cognitive Sciences of Music (ESCOM). pp. 1721–1722.

Wanderley, M. (1999). “Non-obvious performer gestures in instrumental music”. Gesture-based Communication in Human–Computer Interaction, 1. pp. 37–48.

Wanderley, M. (2002). “Quantitative analysis of non-obvious performer gestures”. In I. Wachsmuth, & T. Sowa (Eds.), Gesture and sign language in human-computer interaction. Berlin: Springer. pp. 241–253.

Wanderley, M., Vines, B., 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), pp. 97–113.

Winter, D. A. (1995). “Human balance and posture control during standing and walking”, Gait & Posture, 3 (4). ISSN 0966-6362, DOI: 10.1016/0966-6362(96)82849-9. pp. 193–214.