Comparing Inertial and Optical MoCap Technologies for Synthesis Control

COMPARING INERTIAL AND OPTICAL MOCAP TECHNOLOGIES FOR SYNTHESIS CONTROL

Ståle A. Skogstad and Kristian Nymoen
fourMs - Music, Mind, Motion, Machines
Department of Informatics
University of Oslo
{savskogs,krisny}@ifi.uio.no

Mats Høvin
Robotics and Intelligent Systems group
Department of Informatics
University of Oslo
matsh@ifi.uio.no

ABSTRACT

This paper compares the use of two different technologies for controlling sound synthesis in real time: the infrared marker-based motion capture system *OptiTrack* and *Xsens MVN*, an inertial sensor-based motion capture suit. We present various quantitative comparisons between the data from the two systems and results from an experiment where a musician performed simple musical tasks with the two systems. Both systems are found to have their strengths and weaknesses, which we will present and discuss.

1. INTRODUCTION

Motion capture (MoCap) has become increasingly popular among music researchers, composers and performers [1]. There is a wide range of different MoCap technologies and manufacturers, and yet few comparative studies between the technologies have been published. Where one motion capture technology may outperform another in a sterilized laboratory setup, this may not be the case if the technologies are used in a different environment. Optical motion capture systems can suffer from optical occlusion, electromagnetic systems can suffer from magnetic disturbance, and so forth. Similarly, even though one motion capture system may be better than another at making accurate MoCap recordings and preparing the motion capture for offline analysis, the system may not be as good if the task is to do accurate motion capture in real time, to be used for example in controlling a sound synthesizer.

In this paper we compare the *real-time* performance of two motion capture systems (Figure 1) based on different technologies: Xsens MVN which is based on inertial sensors, and OptiTrack which is an infrared marker-based motion capture system (IrMoCap). Some of our remarks are also relevant to other motion capture systems than the ones discussed here, though the results and discussions are directed only toward OptiTrack and Xsens.

We will return to a description of these technologies in section 3. In the next section we will give a brief overview of related work. Section 4 will present results from comparisons between the two motion capture systems, which are then discussed in section 5.

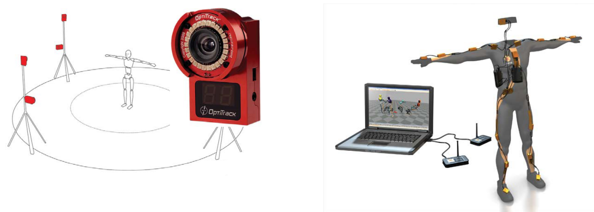

Figure 1. The NaturalPoint OptiTrack system (left) and the Xsens MVN system (right).

2. RELATED WORK AND BACKGROUND

Motion capture technologies have been used in musical contexts for a long time, and during the 00's we saw several examples of using various motion capture technologies for real-time control of sound. This includes electromagnetic motion capture [2], video-based motion capture [3], optical marker-based motion capture [4] and inertial motion capture [5], to mention a few.

Several researchers have reported on differences between motion capture technologies. Most of these reports, however, have been related to offline analysis for medical or animation purposes. Cloete et al. [6] have compared the kinematic reliability of the Xsens MVN suit with an IrMoCap system during routine gait studies. They conclude that the Xsens MVN system is comparable to IrMoCap systems but with shortcomings in some angle measurements. They also point out several practical advantages with the Xsens suit, like its wireless capabilities and quick set-up time. Another experiment by Thies et al. [7] found comparable acceleration values from two Xsens sensors and an IrMoCap system, and showed that calculating acceleration from the IrMoCap position data introduced noise. One of the conclusions from this experiment was that filtering methods need to be investigated further.

Miranda and Wanderley have pointed out some strengths and weaknesses with electromagnetic and optical motion capture systems [1]: Electromagnetic systems are able to track objects, even if it is not within the direct line of sight of external cameras. On the other hand, these systems need cables which may be obtrusive. Optical systems are superior to many other systems in terms of sampling rate, since they may track markers at sampling rates of more than 1000 Hz, and systems using passive markers have no need for obtrusive cables. Still, these systems need a direct line of sight between markers and cameras, and a passive marker system may not be able to uniquely identify each marker.

Possibilities, strengths and weaknesses for real-time motion capture in musical contexts are discussed individually for IrMoCap and full-body inertial sensor systems in [8] and [9]. In this paper we will compare the real-time abilities of the two technologies.

2.1 Initial remarks on requirements when using MoCap for real-time control of music

A musical instrument is normally controlled with excitation and modification actions [10]. We can further distinguish between two types of excitations: discrete (i.e. trigger), or continuous (like bowing a string instrument). Dobrian [11] identifies two types of control data: triggers and streams of discrete data representing a sampling of a continuous phenomenon. Following these remarks, we are looking for a system able to robustly trigger sound events with good temporal accuracy, and to continuously control a system with good spatial accuracy and little noise. Consequently, we have chosen to emphasize three properties: spatial accuracy, temporal accuracy and system robustness. We will come back to measurements and discussion of these properties in sections 4 and 5.

3. TECHNOLOGIES

3.1 NaturalPoint OptiTrack

NaturalPoint OptiTrack is an optical infrared marker-based motion capture system (IrMoCap). This technology uses several cameras, equipped with infrared light-emitting diodes. The infrared light from the cameras is reflected by reflective markers and captured by each camera as 2D point-display images. By combining several of these 2D images the system calculates the 3D position of all the markers within the capture space. A calibration process is needed beforehand to determine the position of the cameras in relationship to each other, and in relationship to a global coordinate system defined by the user.

By using a combination of several markers in a specific pattern, the software can identify rigid bodies or skeletons. A *rigid body* refers to an object that will not deform. By putting at least 3 markers on the rigid body in a unique and non-symmetric pattern, the motion capture system is able to recognize the object and determine its position and orientation. A *skeleton* is a combination of rigid bodies and/or markers, and rules for how they relate to each other. In a human skeleton model, such a rule may be that the bottom of the right thigh is connected to the top of the right calf, and that they can only rotate around a single axis. In the NaturalPoint motion capture software (Arena), there exist 2 predefined skeleton models for the human body. It is not possible to set up user-defined skeletons.

3.2 The Xsens MVN

The Xsens MVN technology can be divided into two parts: (1) the sensor and communication hardware that are responsible for collecting and transmitting the raw sensor data, and (2) the Xsens MVN software engine, which interprets and reconstructs the data to full body motion while trying to minimize positional drift.

The Xsens MVN suit [12] consists of 17 inertial MTx sensors, which are attached to key areas of the human body. Each sensor consists of 3D gyroscopes, accelerometers and magnetometers. The raw signals from the sensors are connected to a pair of Bluetooth 2.0-based wireless transmitters, which again transmit the raw motion capture data to a pair of wireless receivers.

The data from the Xsens MVN suit is fed to the MVN software engine that uses sensor fusion algorithms to produce absolute orientation values, which are used to transform the 3D linear accelerations to global coordinates. These in turn are translated to a human body model which implements joint constraints to minimize integration drift. The Xsens MVN system outputs information about body motion by expressing body postures sampled at a rate up to 120Hz. The postures are modeled by 23 body segments interconnected with 22 joints.

4. MEASUREMENTS

We carried out two recording sessions to compare the OptiTrack and Xsens systems. In the first session, a series of simple measurements were performed recording the data with both Xsens and OptiTrack simultaneously. These recordings were made to get an indication of the differences between the data from the systems. In the second session (Section 4.5), a musician was given some simple musical tasks, using the two MoCap systems separately to control a sound synthesizer.

4.1 Data comparison

Our focus is on comparing real-time data. Therefore, rather than using the built-in offline recording functionality in the two systems, data was streamed in real-time to a separate computer where it was time-stamped and recorded. This allows us to compare the quality of the data as it would appear to a synthesizer on a separate computer. Two terminal applications for translating the native motion capture data to Open Sound Control and sending it to the remote computer via UDP were used.

We have chosen to base our plots on the unfiltered data received from the motion capture systems. This might differ from how a MoCap system would be used in a real world application, where filtering would also be applied. Using unfiltered data rather than filtered data gives an indication of how much pre-processing is necessary before the data can be used for a musical application.

The Xsens suit was put on in full-body configuration. For OptiTrack, a 34-marker skeleton was used. This skeleton model is one of the predefined ones in the Arena software. Markers were placed outside the Xsens suit, which made it necessary to adjust the position of some of the markers slightly, but this did not alter the stability of the OptiTrack system.

Both systems were carefully calibrated, but it was difficult to align their global coordinate systems perfectly. This is because OptiTrack uses a so-called L-frame on the floor to determine the global coordinate system, whereas Xsens uses the position of the person wearing the suit during the calibration to determine the origin of the global coordinate system. For this reason, we get a bias in the data from one system compared to the other. To compensate for this, the data has been adjusted so that the mean value of the data from the two systems more or less coincide. This allows us to observe general tendencies in the data.

4.2 Positional accuracy and drift

When comparing the Xsens and the OptiTrack systems there is one immediately evident difference. OptiTrack measures absolute position, while the sensors in the Xsens MVN suit can only observe relative motion. With Xsens, we are bound to experience some positional drift even though the system has several methods to keep it to a minimum [9].

4.2.1 Positional accuracy - still study

Figure 2 shows the position of the left foot of a person sitting in a chair without moving for 80 seconds. The upper plot shows the horizontal (XY) position and the lower plot shows vertical position (Z) over time. In the plot it is evident that Xsens suffers from positional drift, even though the person is sitting with the feet stationary on the floor. Xsens reports a continuous change of data, with a total drift of more than 0.2 m during the 80 seconds capture session. Equivalent plots of other limbs show similar drift, hence there is little relative drift between body limbs.

This measurement shows that OptiTrack is better at providing accurate and precise position data in this type of clinical setup. However, for the vertical axis, we do not observe any major drift, but the Xsens data is still noisier than the OptiTrack data.

4.2.2 Positional accuracy - walking path

The left plot in Figure 3 displays the horizontal (XY) position of the head of a person walking along a rectangular path in a large motion capture area recorded with Xsens. The plot shows a horizontal positional drift of about 2 meters during the 90 seconds capture session. Xsens shows

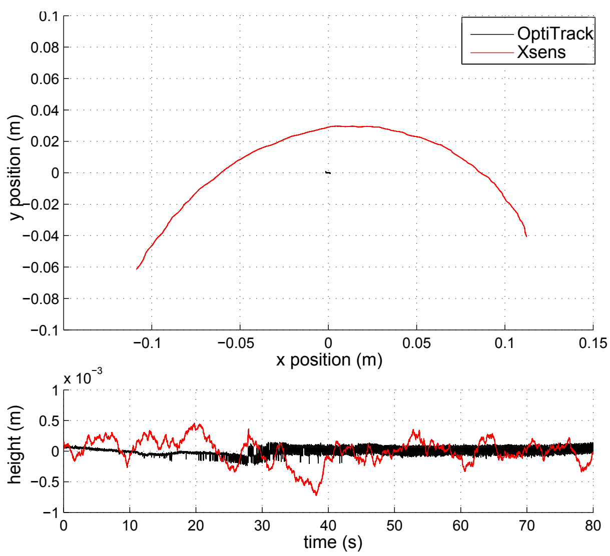

Figure 2. Horizontal and vertical plots of a stationary foot.

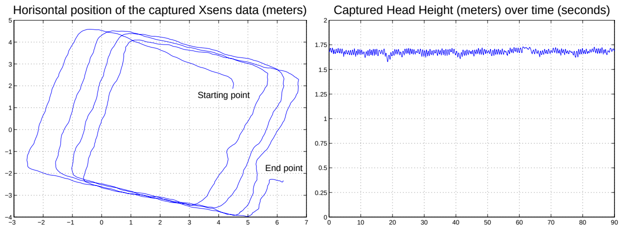

Figure 3. Recording of the horizontal (left) and vertical (right) position of the head.

no drift in the vertical direction (Z), as can be seen in the right plot. This is expected since the MVN engine maps the data to a human body model and assumes a fixed floor level. Because of the major horizontal drift we can conclude that Xsens MVN is not an ideal MoCap system if absolute horizontal position is needed.

4.2.3 Camera occlusion noise

The spatial resolution of an IrMoCap system mainly relies on the quality of the cameras and the calibration. The cameras have a certain resolution and field of view, which means that the spatial resolution of a marker is higher close to the camera than far away from the camera. The calibration quality determines how well the motion capture system copes with the transitions that happen when a marker becomes visible to a different combination of cameras. With a "perfect" calibration, there might not be a visible effect, but in a real situation we experience a clearly visible change in the data whenever one or more cameras fail to see the marker, as shown in Figure 4. When a marker is occluded from a camera, the 3D calculation will be based on a different set of 2D images.

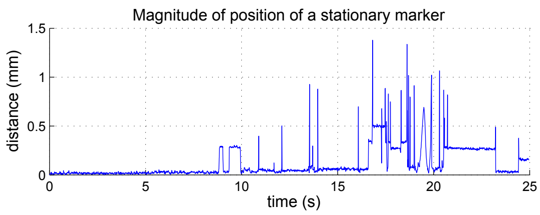

Figure 4. OptiTrack: Magnitude of the distance from the mean position of a stationary marker. The disturbances in the last part of the measurement is caused when a person moves around the marker, and thus blocks the marker in one or more cameras at a time. FrameRate 100 Hz

4.2.4 Xsens floor level change

If the motion capture area consists of different floor levels, like small elevated areas, the Xsens MVN engine will match the sensed raw data from the suit against the floor height where the suit was calibrated. This can be adjusted in post-processing, but real-time data will suffer from artifacts during floor level changes, as shown in Figure 5.

4.3 Acceleration and velocity data

In our experience, velocity and acceleration are highly usable motion features for controlling sound. High peaks in absolute acceleration can be used for triggering events, while velocity can be used for continuous excitation.

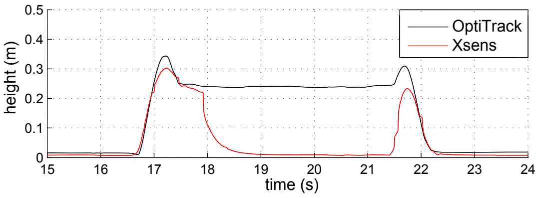

Figure 5. Recording of the vertical position of the left foot of a person, stepping onto an elevated area (around 0.25 m high). When the user plants his left foot on the object, the Xsens MVN engine will eventually map the stationary foot to floor level (18 to 19 s).

A difference between the two MoCap systems is that the Xsens system can offer velocity and acceleration data directly from the MVN engine [9]. When using the OptiTrack system we need to differentiate position data to estimate velocity and acceleration. If the positional data is noisy, the noise will be increased by differentiation (act as an high-pass filter), as we can see from Figure 6. The noise resulting from optical occlusion (see Section 4.2.3) is probably the cause for some of OptiTrack's positional noise.

Even though the Xsens position data is less accurate, it does offer smoother velocity and, in particular, acceleration data directly. We can use filters to smooth the data from the OptiTrack system; however, this will introduce a system delay, and hence increased latency.

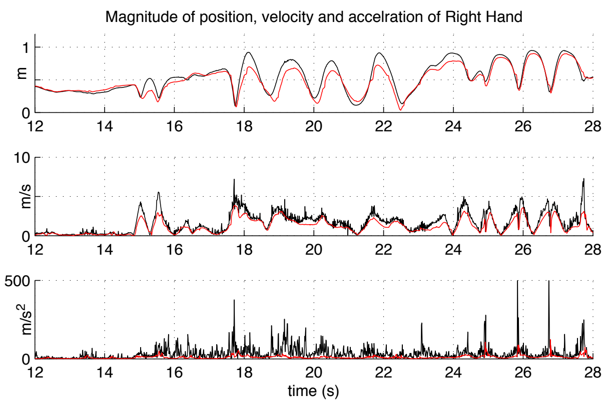

Figure 6. Velocity and acceleration data quality comparison (OptiTrack in black and Xsens in red).

4.4 Action-to-sound: latency and jitter

Low and stable latency is an important concern for *real-time* musical control [13], particularly if we want to use the system for triggering temporally accurate musical events. By *action-to-sound latency* we mean the time between the sound-producing action and the sonic reaction from the synthesizer.

To be able to measure the typical expected latency in a setup like that in Figure 7 we performed a simple experiment with an audio recorder. One computer was running one of the MoCap systems and sent OSC messages containing the MoCap information about the user's hands. A patch in Max/MSP was made that registered hand claps based on MoCap data and triggered a *click* sound for each clap. The time difference between the acoustic hand clap and the triggered sound should indicate the typical expected latency for the setup.

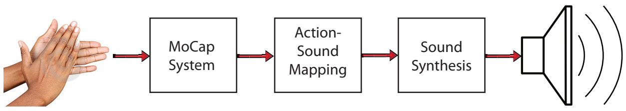

Figure 7. The acoustic hand clap and the triggered sound were recorded to measure latency of the systems.

Both MoCap systems were run on the same PC [1]. The sound-producing Max/MSP patch was run on a separate Mac laptop [2] and received OSC messages from the MoCap systems through a direct Gbit Ethernet link. All experiments used the same firewire connected sound card, *Edirol FA-101*, as output source. The hand claps and the click output from the Max patch was recorded with a microphone. Statistical results from the time delays between hand claps and corresponding click sound in the recorded audio files are given in Table 1. The values are based on 30 claps each. In this experiment, OptiTrack had a faster sound output response and a lower standard deviation than Xsens. The standard deviation is included as an indication of the jitter performance of the MoCap systems, since lower standard deviation indicates higher temporal precision.

Higher Xsens latency and jitter values are probably partly due to its use of Bluetooth wireless links. The Xsens MVN system also offers a direct USB connection option. We performed the same latency test with this option; and the results indicate that the connection is around 10-15 milliseconds faster, and has a lower jitter performance, than the Bluetooth link.

The upper bounds for "intimate control" have been suggested to be 10ms for latency and 1ms for its variations (jitter) [13]. If we compare the bounds with our results, we see that both systems have relatively large latencies. However, in our experience, a latency of 50ms is still usable in many cases. The high jitter properties of the Xsens system are probably the most problematic, especially when one wants high temporal accuracy.

	min	mean	max	std. dev.
OptiTrack	34	42.5	56	5.0
Xsens Bluetooth	41	52.2	83	8.4
Xsens USB	28	37.2	56	6.9

Table 1. Statistical results of the measured action-to-sound latency, in milliseconds.

4.5 Synthesizer control

In a second experiment, a musician was asked to perform simple music-related tasks with the two motion capture

[1] Intel 2.93 GHz i7 with 8GB RAM running Win 7
[2] MacBook Pro 10.6.6, 2.66 GHz Duo with 8GB RAM

systems. Three different control mappings to a sound synthesizer were prepared:

- Controlling pitch with the distance between the hands
- Triggering an impulsive sound based on high acceleration values
- Exciting a sustained sound based on the velocity of the hand

For the pitch mapping, the task was to match the pitch of one synthesizer to the pitch of another synthesizer moving in the simple melodic pattern displayed in Figure 8, which was repeated several times. This task was used to evaluate the use of position data from the two systems as the control data.

For the triggering mapping, the task was to follow a pulse by clapping the hands together. This task was given to evaluate acceleration data from the two systems as the control data, and to see if the action-to-sound latency and jitter would make it difficult to trigger events on time.

The excitation mapping was used to follow the loudness of a synthesizer, which alternated between ”on” and ”off” with a period of 1 second. This task was used to evaluate velocity data as control data.

The *reference sound* (the sound that the musician was supposed to follow) and the *controlled sound* (the sound that was controlled by the musician) were played through two different loudspeakers. The two sounds were also made with different timbral qualities so that it would be easy to distinguish them from each other. The musician was given some time to practice before each session. To get the best possible accuracy, both systems were used at their highest sampling rates for this experiment: Xsens at 120 Hz, and OptiTrack at 100 Hz.

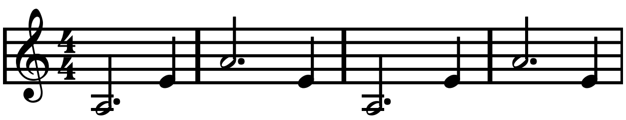

Figure 8. The simple melody in the pitch-following task. This was repeated for several iterations.

4.5.1 *Pitch-following results*

We found no significant difference between the performances with the two systems in the pitch-following task. Figure 9 displays an excerpt of the experiment, which shows how the participant performed with both Xsens and OptiTrack. The participant found this task to be difficult, but not more difficult for one system than the other. Also, the data shows no significant difference in the performances with the two systems. This indicates that the quality of relative position values (between markers/limbs) is equally good in the two systems for this kind of task.

4.5.2 *Triggering results*

Table 2 shows the results of the latency between the reference sound and the controlled sound for the triggering test. They are based on 40 hand claps for each of the two

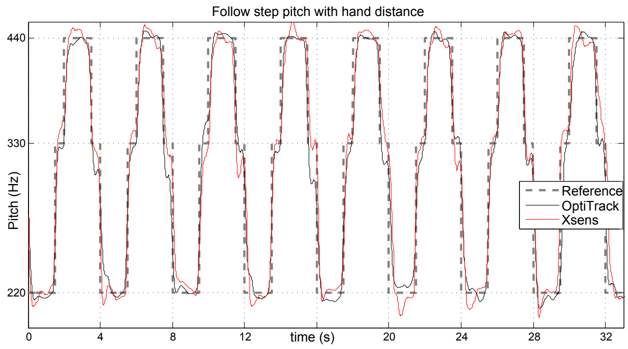

Figure 9. There was no significant difference between the two systems for the pitch-following task.

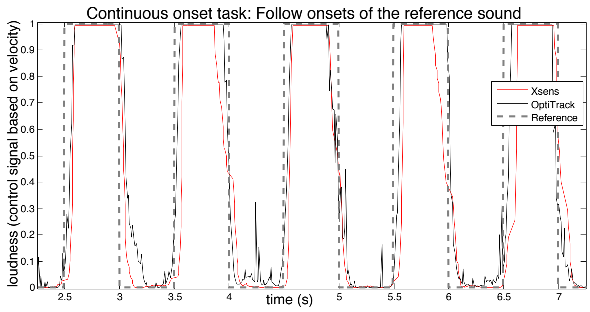

Figure 10. The major difference between the two systems in the continuous onset task was the noisy data from the OptiTrack system, which made it difficult to be quiet between the onsets. Apart from this, there was no big difference between the two systems.

MoCap systems. As we can see, the *mean* latency value is almost equal for Xsens and OptiTrack. Xsens has a higher standard deviation, which may indicate that the Xsens jitter shown in Table 1 makes it difficult for the user to make a steady trigger pulse.

	min	mean	max	std. dev.
OptiTrack	18.5	45.2	77.1	13.8
Xsens	2.6	44,7	96.3	28.3

Table 2. Statistical results, in milliseconds, of the measured time differences between reference signal and control signal.

4.5.3 *Continuous onset results*

For the continuous onset task, where the loudness of the sound was controlled by the absolute velocity of the right hand, we also observed a time delay between the onset of the reference tone and the onset of the sound played by our performer. This delay was present for both systems. In this task, the OptiTrack system suffered from noise, which was introduced when calculating the absolute velocity of the unfiltered OptiTrack data, as described in Section 4.3 (see Figure 10). The musician said that this made it more difficult to be quiet between the reference tones, and that this task was easier to perform with the Xsens system.

5. DISCUSSION

We have seen several positive and negative aspects with the quantitative measurements of the two technologies. In this section we will summarize our experiences of working with the two systems in a music-related context.

The main assets of the Xsens suit is its portability and wireless capabilities. The total weight of the suit is approximately 1.9 kg and the whole system comes in a suitcase with the total weight of 11 kg. Comparably, one could argue that a 8-camera OptiTrack setup could be portable, but this system requires tripods, which makes it more troublesome to transport and set up. OptiTrack is also wireless, in the sense that the user only wears reflective markers with no cables, but the capture area is restricted to the volume that is covered by the cameras, whereas Xsens can easily cover an area with a radius of more than 50 meters. When designing a system for real-time musical interaction based on OptiTrack, possible marker dropouts due to optical occlusion or a marker being moved out of the capture area must be taken into account. For Xsens, we have not experienced complete dropouts like this, but the Bluetooth link is vulnerable in areas with heavy wireless radio traffic, which may lead to data loss. Nevertheless, we consider Xsens to be the more robust system for on-stage performances.

OptiTrack has the benefit of costing less than most other motion capture technologies with equivalent resolution in time and space. The full Xsens suit is not comfortable to wear for a longer time period, whereas OptiTrack markers impose no or little discomfort. On the other hand, OptiTrack markers can fall off when tape is used to attach them. Also, OptiTrack's own solution for hand markers, where a plastic structure is attached to the wrist with Velcro, tends to wobble a lot, causing very noisy data for high acceleration movement, something we experienced when we set up the hand clapping tests. Xsens has a similar problem with the foot attachments of its sensors, which seems to cause positional artifacts.

Sections 4.2 to 4.5 show a number of differences between Xsens and OptiTrack. In summary, OptiTrack offers a higher positional precision than Xsens without significant drift, and seemingly also lower latency and jitter. Xsens delivers smoother data, particularly for acceleration and velocity. Our musician subject performed equally well in most of the musical tasks. However, the noisy OptiTrack data introduced some difficulties in the continuous onset task, and also made it challenging to develop a robust algorithm for the triggering task. Furthermore, Xsens jitter made the triggering task more difficult for the musician.

6. CONCLUSIONS

Both OptiTrack and Xsens offer useful MoCap data for musical interaction. They have some shared and some individual weaknesses, and in the end it is not the clinical data that matters, but the intended usage. If high positional precision is required, OptiTrack is preferable over Xsens, but if acceleration values are more important, Xsens provide less noisy data without occlusion problems. Overall, we find Xsens to be the most robust and stage-friendly MoCap system for real-time synthesis control.

7. REFERENCES

[1] E. R. Miranda and M. Wanderley, *New Digital Musical Instruments: Control And Interaction Beyond the Keyboard*. A-R Editions, Inc., 2006.

[2] J. Michel Couturier and D. Arfib, "Pointing fingers: Using multiple direct interactions with visual objects to perform music," in *Proc . NIME*, 2003, pp. 184–188.

[3] G. Castellano, R. Bresin, A. Camurri, and G. Volpe, "Expressive control of music and visual media by full-body movement," in *Proc. NIME*. New York, USA: ACM, 2007, pp. 390–391.

[4] F. Bevilacqua, J. Ridenour, and D. J. Cuccia, "3d motion capture data: motion analysis and mapping to music," in *Proc. Workshop/Symposium SIMS*, California, Santa Barbara, 2002.

[5] P.-J. Maes, M. Leman, M. Lesaffre, M. Demey, and D. Moelants, "From expressive gesture to sound," *Journal on Multimodal User Interfaces*, vol. 3, pp. 67–78, 2010.

[6] T. Cloete and C. Scheffer, "Benchmarking of a full-body inertial motion capture system for clinical gait analysis," in *EMBS*, 2008, pp. 4579 –4582.

[7] S. Thies, P. Tresadern, L. Kenney, D. Howard, J. Goulermas, C. Smith, and J. Rigby, "Comparison of linear accelerations from three measurement systems during reach & grasp," *Medical Engineering & Physics*, vol. 29, no. 9, pp. 967 – 972, 2007.

[8] S. A. Skogstad, A. R. Jensenius, and K. Nymoen, "Using IR optical marker based motion capture for exploring musical interaction," in *Proc. NIME*, Sydney, Australia, 2010, pp. 407–410.

[9] S. A. Skogstad, K. Nymoen, Y. de Quay, and A. R. Jensenius, "Osc implementation and evaluation of the xsens mvn suit," in *Proc of NIME*, Oslo, Norway, 2011.

[10] A. R. Jensenius, M. M. Wanderley, R. I. Godøy, and M. Leman, "Musical gestures: concepts and methods in research," in *Musical Gestures: Sound, Movement, and Meaning*, R. I. Godøy and M. Leman, Eds. New York: Routledge, 2010, pp. 12–35.

[11] C. Dobrian, "Aesthetic considerations in the use of 'virtual' music instruments," in *Proc. Workshop on Current Research Directions in Computer Music*, 2001.

[12] D. Rosenberg, H. Luinge, and P. Slycke, "Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors," *Xsens Technologies*, 2009.

[13] D. Wessel and M. Wright, "Problems and prospects for intimate musical control of computers," in *Proc. NIME*, Seattle, USA, 2001.