



Audio Engineering Society Convention e-Brief 129

Presented at the 149th Convention
2020 October 21 – 24, New York, NY

This Engineering Brief was selected on the basis of a submitted synopsis. The author is solely responsible for its presentation, and the AES takes no responsibility for the contents. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Audio Engineering Society.

Multimodal Immersive Motion Capture (MIMiC): A workflow for the musical performance

Cindy Bui, Andrea Genovese, Trey Bradley, and Agnieszka Roginska

New York University - Music and Audio Research Lab

Correspondence should be addressed to (roginska@nyu.edu)

ABSTRACT

This paper introduces a workflow to record paired motion capture and sound of musical performances using an infrared light-based tracking system. This paper discusses example production cases ranging from short loops to full performances, addressing the production challenges and compromises, substantially different from a regular audio recording process. Motion capture is very versatile and can be applied in many disciplines including musical pedagogy, distributed performances, machine learning, and as creative assets. Paired audio-mocap data can be rendered into similar visual content for interactive virtual environments, live performances or fixed media. This extensible workflow can be adapted and extended for different purposes in musical performances to promote further research and creative content development.

1 Introduction

Motion capture (mocap) data records the three dimensional coordinates of tracked objects in space over time. In an optical, infrared light-based tracking system, physical markers capture rigid bodies, human motions, and gestures. The collection of markers of a human body is called a skeleton. Depending on the mocap software, these markers can be specialized for overall body motions or nuanced finger motion [1]. These data points are later turned into animations through skeleton rigging.

There is currently no widespread distribution of practical guidelines to capture precise mocap content for instrumental music performers. The process of collecting mocap data from musicians is affected by a number of factors and limitations of professional-grade motion tracking equipment, acoustic considerations, and the artistic nature of the content in question.

The purpose of this paper is to discuss and propose a mocap production workflow for musical performances and loops. Example case scenarios are presented to illustrate the compromises and adjustments needed to address the engineering and musical challenges. Specifically, we employed a modern optical infrared tracking system, making some of our discussion points only applicable to this choice of technology. This paper discusses prototype implementations and usage scenarios as a resource to performers and arts-research communities alike.

1.1 Literature and Applications

The use of mocap data is a relatively novel endeavour in the performing arts. Because of its flexibility, this kind of content can be utilized for a wide variety of creative applications. Compared to video, mocap is a more extensible and flexible alternative for an artist

since the data can be manipulated and customized for visualization over various kinds of displays. Real-time mocap and pre-captured performer data can be combined for augmented concerts and distributed performances, where live performers can musically interact with a virtual ensemble of avatars.

Previous related work by the authors in [2] looked into the design challenge of creating a collaborative virtual environment cohabited by pre-rendered audio-visual avatars of a drumming ensemble and a live-rendered performer. In that fixed installation, the live musician would perform along to the avatars' ensemble while being seen in real-time through a VR headset by an audience member. The biggest challenge faced in this scenario was the acoustical integration of the two performing elements, necessary to create a cohesive sound blend. In this case, given the controlled nature of the reproduction environment, the concert space could be treated to maximise timbral integrity with the acoustical character of the avatar recordings (which were non-anechoic). With the caveat that only the assigned space would provide an optimal acoustic blend. This kind of multimodal settings can be of interests in the field of augmented distributed music [3], in order to study the effect of body avatars in terms of telepresence, virtual co-presence [4], and musical output while under latency conditions.

Performers' audio and mocap paired data can find several applications in research across different fields. For example, applications in music education, pedagogy and physical therapy can benefit from audio and mocap paired recordings in order to correct for posture or analyze students' progress [5, 6]. The fields of musicology and ethnomusicology also may find interest in this sort of data in order to quantitatively analyze gestural expression across cultures and genres [7, 8]. On a larger scale, an audio-mocap repository dedicated to instrumental performers and related dance performances can serve as an important machine learning dataset for learning-based studies over several topics. Some examples are gestural mapping [9], genre classification [10], and procedural model animation [11]. Similar datasets based on motion capture for general movements and sports actions have been compiled [12, 13], however they do not cover musical activities or the presence of related audio material. Real-performer mocap recordings can thus be used as ground-truth data for assessing synthesized procedural animations and character models.

2 Workflow goals

The driving principle behind planning a recording session is to obtain high quality paired audio and mocap data. The simultaneous capturing of the two components is crucial in order to effectively synchronize the musical output with the gestural motions. Regarding the audio data, it is suggested to follow design principles that relate to object-based audio recording. The sound should be as dry and anechoic as possible in order to allow custom effects in the post processing to blend with any virtual room. The best way to capture dry audio is through close-miking techniques.

Capturing high-quality "clean" motion data reduces the amount of necessary post-processing. It is customary to edit the data during post-processing to remove artifacts and errors, however it can be incredibly tedious to edit and fix faulty data into a smooth animation before the animation can be used. The goal of the capture stage is to reduce the amount of cleanup that needs to happen.

2.1 Challenges

There are several challenges when it comes to recording high quality mocap and audio simultaneously. These stem from the fact that they are two different platforms using capturing equipment that may mutually interfere with each other. To achieve the goals of dry sound and clean mocap, some technical and musical adjustments need to take place.

Constraints

The most complicated and difficult challenge when recording mocap and audio is handling infrared (IR) reflections. A typical optical system tracks motion by emitting and capturing IR light reflected by markers. However, shiny equipment and surfaces can also reflect light and create artifacts, leading to either a calibration failure or the rendering of extraneous *dots* (software representation of a marker). Nearly all audio recording equipment is made out of metal or other shiny materials and the vast majority of instruments are also shiny or have a glossy finish. People extensively using mocap may use paints and coatings that can remove the glossiness of their equipment, but it is not practical to coat expensive audio and musical equipment in paint or tape. Compromises are therefore necessary for microphone positioning and instrumentation.

On the other hand, markers that need to be tracked may be temporarily lost by the system. Musicians' body motions and instrument placement will need to be constrained since their natural movements may cause data occlusions. Temporal occlusions result in data gaps which are corrected by interpolating between clean neighbour points or by hand editing the marker coordinates. The cleaning process requires considerable time and effort, leading to discarded data in the most severe cases. This becomes more complicated if more than one marker was occluded at any given point in time. Depending on the type of motion during the dropout, it may sometimes be more time efficient to recapture the performance than spend hours trying to clean it.

Besides the constraints on the instrumentation choice, there are other musical implications. Wearing mocap suits is often a cumbersome experience. Performers wear a black velcro suit with the markers are placed in positions specified by the software. It is important for the suit to not move and adjust during the recording, so for instance the suit jacket should be over the pants and velcroed together. The nuance of natural performance movements may be hampered by the positioning of markers on certain body parts, making it necessary to choose between a compromised musical execution or a reduced tracking resolution (for example by excluding fingers).

The standard positioning for an ensemble may not be the best positioning to capture mocap. Most ensembles in a rehearsal space arrange themselves in a circle, but this arrangement will directly conflict with the mocap requirements if the space is too small. The musicians could end up on the edge of the tracking area and out of the field of view of most cameras. Another consideration regards music stands, which musician may demand to read scores. The stand will both occlude markers since it's in front of the musician and cause IR reflections since they are made out of metal. These problems need to be addressed in the planning phase, and the musicians should be aware of these constraints before the recording session to ensure a quality performance.

Synchronization

The two platforms are also separate asynchronous systems. The mocap sample rates are significantly slower than audio sample rates, and the equipment runs on two different hardware clocks. They need to be synchronized manually in the post processing stage. This

is overcome easily with a reference slate – similar to the slates that were once used in the film industry. One way to do it is to ask the performers to loudly clap at the start of the recording. The clap waveform transient can then be used for visual alignment with the motion data.

Space considerations

The choice and characteristics of a tracking space are important for both the sound and mocap. Mainly, the room should be as dry as possible, devoid of audible room modes and strong reverberation. This is especially important in situations where close-miking techniques interfere with the marker tracking and more distanced microphone placement is necessary. Small studio booths provide the best acoustic conditions but don't have much space for a performer to move.

For a single seated musician, a small tracking area usually suffices. For an ensemble, the tracked musicians may need to be strategically placed to avoid being near the edges of the mocap area. This can potentially conflict with audio recording techniques. For instance, a non-tracked musician may get placed next to a wall, which is not ideal since that recording will be colored by that particular wall's timbre.

3 Capture stage case studies

The recording sessions took place in the James L. Dolan Studio live room at NYU. The space is approximately 4.5 by 9 meters in dimension with a tracking area of approximately 4 by 6.5 meters with a relatively short reverberation time of 0.35 seconds. Ten motion capture cameras connected to a network switch, were setup on the ceiling and faced down into the area. Calibration of the cameras had to be performed for each session, or each time the cameras were reoriented¹. A PC with the motion capture tracking software (*Optitrack Motive*²) was connected to this switch. Tracked marker dots are used to create a digital skeleton entity, each dot labels a part of the skeleton (e.g. head front, left wrist out, chest front, etc.). Audio recordings were processed on a separate machine.

¹The exact form of the calibration process depends on the system, typically it consists on cross-comparing sample data points through a "wandering" process.

²<https://optitrack.com/software/>

3.1 Quartet

This pilot session involved an ensemble of four musicians: two violins, a tenor saxophone, and grand piano. The goal of this session looked into the feasibility of recording mocap (for the saxophone player and one violinist), spot microphones, and the spatial soundfield in the same session - without special arrangements in regards to the discussed constraints. Musicians were disposed in a circle around a cluster of microphone arrays (an HOA array, an FOA microphone, a Blumlein coincident pair, and an MSZ coincident trio). Each instrument was also close-miked, and recorded through an overhead array (Hamasaki square).

In this particular session, the calibration and skeleton imaging were attempted after the microphones were set up in the tracking area. However, the optical tracking system failed to accurately calibrate due to the high level of equipment clutter blocking the visual field. During capturing it appeared that the tracking area was too small for the size of the ensemble, pushing musicians to the edge of the area and reducing the number of cameras tracking each individual marker. By extension, the computed skeleton images were deemed faulty and imprecise, with the limbs and joints showing severely erroneous orientations.

The whole session was repeated to better fit the mocap limitations. The microphone arrays which were not strictly necessary to the session were removed and the musicians rearranged into an "L" configuration, with the mocapped players placed at the centre of the capture area. This arrangement was reportedly more difficult for the performers as they could not make easy visual contact, but the decision favoured the mocap quality, resulting in considerably cleaner and stabler data.

3.2 African drumming trio

This session consisted in the recording of a three-minute African percussion piece (*Doundounba*) composed of three voices, each one progressively recorded in layers by the same performer at a tempo of 150 bpm (Fig. ??). To capture different tones of the instrument, we used two sets of microphones - an XY coincident stereo pair, very close to the djembe head, and a large diaphragm dynamic microphone at the bottom of the drum. We also placed acoustic isolators and carpets around the musician to ensure a dry audio capture while reducing spurious infrared unwanted reflections. This

multitrack session was a re-capture of the piece included in previous experimental displays [2], this time with better controlled acoustic conditions.

The simplicity of the setup allowed for a reasonably clean capture, besides a mislabeling issue at one of the hands. Each time the drummer's left hand hit the drum, the outer wrist marker would hit the left thigh marker. This confused the tracking, and the software would temporarily mix up the two markers. This mislabelling issue accounted for the majority of the cleaning process. Being drumming is a relatively simple motion, no further complications due to instrument and finger tracking were deemed necessary.

We additionally captured a pair of dancers for this performance using the same system. With their wide range of motions, this exposed severe occlusion issues with our mocap camera setup which prompted us to rethink about the positioning of the cameras and optimize them for different recording sessions.

3.3 Snare drum loops

After recording these performances, we decided to do a recording session to try to capture mocap loops. These loops could be integrated in an XR experience to produce a virtual "drum pad". This session was a comprehensive recording of a classical snare drummer playing different rudiment patterns at different tempos. The snare drummer played excerpts from *Stick Control* by George Lawrence Stone [14]. In order to capture the nuance of how speed naturally affects motion, each chosen pattern was performed at tempos of 60, 80, 100, 120, 140, and 160 bpm. The effect of motion dynamics was also captured by repeating each condition at *f* (forte) and at *p* (piano) intensities. The loop patterns are shown in Table 1.

The snare drum was captured using a single microphone placed at the center of the capture area, pointing to the upper diaphragm. Because the snare drum itself causes calibration and reflection issues with the mocap system, all equipment was removed before calibrating the cameras and carpets were layered down on the floor. Once the mocap system was calibrated and the skeleton was created, we moved the snare drum, drum throne, and microphone back into the tracking area.









Loop pattern	Tempo (bpm)	Dynamics
	60, 80, 100, 120, 140, 160	<i>p, f</i>
	60, 80, 100, 120, 140, 160	<i>p, f</i>
	60, 80, 100, 120, 140, 160	<i>p, f</i>
	60, 80, 100, 120, 140, 160	<i>p, f</i>
	60, 80, 100, 120, 140, 160	<i>p, f</i>
	60, 80, 100, 120, 140, 160	<i>p, f</i>
	60, 80, 100, 120, 140, 160	<i>p, f</i>
	60, 80, 100, 120, 140, 160	<i>p, f</i>

Table 1: Set of loops for audio-mocap recordings with a snare drummer

4 Discussion

Throughout our case studies, there were several compromises and alterations that had to be made in order to ensure that both the mocap recordings and audio recordings were of sufficient quality. Some alterations were easy to implement and coincided with both mocap and audio goals, while others clashed and needed modifications to come up with a good compromise.

Each scenario presented its unique challenge and required an ad hoc solution, in order to get the cleanest data capture. Generally, the positioning of the cameras and choice of instruments should be planned with a flexibility-oriented mindset. This is notably contrary to the typical recording studio workflow. The usual workflow for recording studios are to setup all of the audio equipment first and then introduce the musicians a few

hours later when the setup is ready. The audio-mocap workflow requires more time and cooperation from the musicians.

4.1 Mocap Setup

The most important insight learned in the process was that the mocap setup and calibration needs to happen before the audio setup. The calibration results are used to help triangulate the marker positions among all the cameras. If the calibration result is poor, then the capture quality is expected to be equally poor.

For an optimal calibration, the tracking area should be cleared of obstacles and repeated at the beginning of every session. Small rumbles or vibrations, such as a door closing, may affect the camera position over time. Calibration may also need to happen in the middle of a

long session if the cameras drift too much. Cameras are usually on the ceiling facing down, but a few cameras can be moved to ground level or eye level to get a more complete field of view. There is a caveat that cameras shouldn't be facing other cameras. A mocap camera directly in the field of view of another mocap camera will be depicted as an interfering dot. Dots like these can sometimes be masked out in software, but preemptively rearranging the cameras will have better results than masking it in software. An alternative is a hardware filter that makes the camera emit a different frequency of infrared. If any camera is moved, then the entire system needs to be re-calibrated.

To maximize tracking stability, it is important to focus the calibration process on wherever the performers are planned to be, adjusting the floor level accordingly to the performer's feet. As a rule of thumb, if a single seated person is being recorded, the cameras should all be centered on that one location. If the goal is to capture performers with a large range of motion, the cameras should be oriented so that they cover the widest range of view. In the example of the snare drummer session, cameras were oriented to point to the center of the tracking area, then calibrated to collect sample data points where the musician was planned to stand. The calibration process should be followed by a skeleton imaging step, where a character "entity" is assigned to a set number of dots (the performer markers). Once the skeleton is created, the software can properly differentiate between any new dots that enter the scene and the pre-existing skeleton.

At this stage, the audio equipment can be moved in. The motion capture recording software should be actively monitored during this time. The arrangement of the cameras and instruments may need to be re-adjusted according to the amount of equipment and people. The instruments should also be adjusted if necessary and if possible. A drumset, for example, will occlude the kick drum foot of the drummer. Some solutions are to move the camera to be directly on top of the drumset, or to move the floor tom so at least one camera can see it. Guitarists will often have amplifiers, and sometimes those amplifiers will be shiny and reflective. A blanket can be wrapped around the outside cover of the amplifier to prevent reflections.

4.2 Audio Setup and Monitoring

Our primary audio goal was to record dry sound. For audio purposes, we placed dampening panels in our

space, and we close miked the instruments. The panels were beneficial for both audio and mocap since the tracking space had glass windows and doors could reflect infrared light. We also placed blankets on the floor to prevent any infrared reflections coming from our hardwood floors. The microphone stands and the music stands' legs can be hidden underneath the blankets if they caused extraneous dots in the scene. This was also inconsequential to the audio.

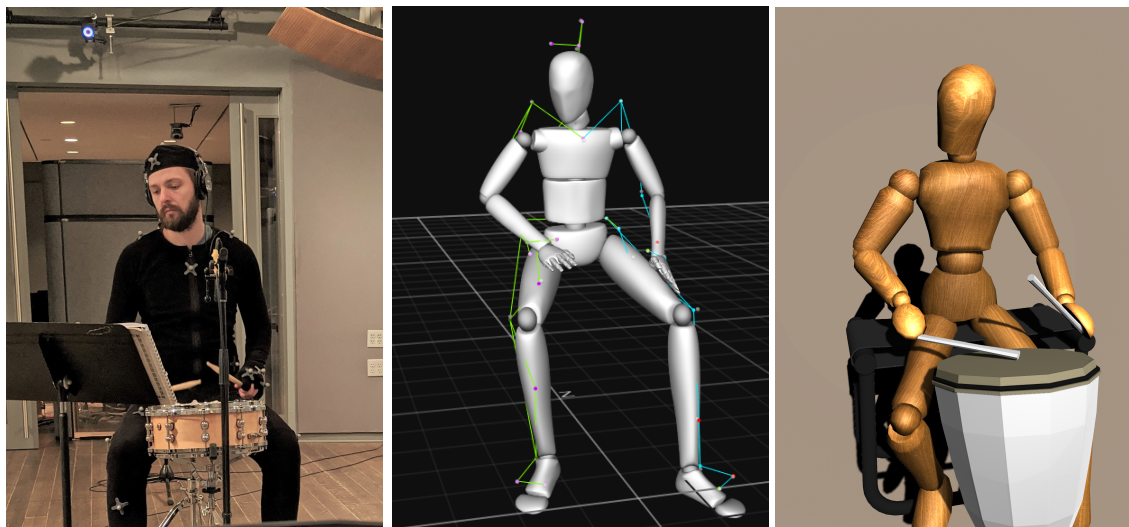
During the actual recording session, the mocap recording software needs to be actively monitored in case the skeleton becomes distorted. This can happen over time if the musician adjusts their suit, or if the cameras move slightly, from structural vibrations. Distortions can also happen if the software loses track of a marker for a long period of time, due to occlusions. If distortion happens, then the performer needs to do a calibration pose. In our software, this was a T-Pose where the performer puts their arms out sideways. It is thus safer to record short loops or phrases rather than long pieces so that way the mocap inconsistencies can be flagged immediately.

5 Post-processing

Once the data has been captured, the mocap data needs to go through a "cleaning" process. Regardless of how well the data was captured, the chance of errors or rough edges remains high, with the process usually ranging from 10 minutes up to 5 hours according to the quality of the capture. The mocap software has the capability to visualize the markers' coordinates over time as a time series. Occlusion dropouts, temporary mislabels, noisy signals, or sharp unnatural movements can be seen in this time series. Mislabelling can happen if two markers touch each other or come into close proximity of each other. The software has ways to swap labels on dot data, deleting dot data over specific amount of time, interpolating gaps in data, and smoothing coordinate motion curves. The cleaning software exports the mocap data into a *biovision hierarchical data* (bvh) file, which can be easily imported into a rigging software of choice.

As clean motion data is just a time-series of coordinates, it does not inherently include a visualizable mesh. The visualization can be implemented by associating the coordinates with a rigged 3D character model through a modeling software (e.g. *Blender*³). The association

³<https://blender.org>



(a) Mocapped performer

(b) Raw mocap data

(c) Rigged and rendered avatar

Fig. 1: Capturing, cleaning and rendering stages for the snare drum loops session

process attaches the marker coordinates to the corresponding "bone" texture of the avatar model. The process can be straightforward to implement provided the avatar model is of the "humanoid" type, otherwise needing a creative rigging process. The output of the rigging software is a standalone animation object which can be imported in a game engine or rendered into video (see Fig 1).

Finally, audio can be added to the scene and synchronized with the animation. Time-domain synchronization with audio and take-trimming can either be performed within the game-engine or at previous stages, using the slate as reference point for establishing the time marks of edit points. The slate, enacted by the mocapped performer, serves as a visual aid to align audio and mocap data, as audio transients and physical motions can be easily spotted and linked.

5.1 Game engine production

A possible use of this data, is to port the rigged mocap-animated model into a game engine (e.g. *Unity* or *Unreal*) as a multimedia asset. The main goal at this stage is to design the scene where the animation lives in, apply aesthetic design, and possibly synchronize the audio with the animation if the process was not yet performed. The advantage of synchronizing within a game engine is the possibility of computationally tweaking

the synch and the relationship between the performance and sound. For example, a group of performers can be simulated by introducing random coordinate noise and off-synch artificial imperfections on copies of the original avatar.

In addition, a game engine software provides tools and plugins to configure interactive user experience elements that can allow for user control over the animation and introduce other immersive elements, like spatial audio. Final deployment can equally occur for an XR device headset or screen display. However, the application and device choice may affect the need to treat the acoustical character of the recordings for achieving a cohesive blend with the virtual or mixed-reality environment or live-collaborating entities.

6 Conclusions and future work

This paper introduces a proposed workflow for the production of audio-mocap data of musicians. The engineering challenges and musical compromises faced in our capture scenarios are presented and discussed in order to provide an overview of the kind of considerations taken when capturing and implementing audio-mocap content. The key driving principles for simultaneous paired recordings regard the necessities of capturing dry audio and maximizing the quality of the motion

data, while preserving the creative intentions as much as possible.

All of the captured data, notably the snare drum loops, are interesting for the purposes of multimodal distributed music studies over XR headsets. Current cutting edge distributed music performances will stream audio and video data. Adding another modality of a mocap stream, whether live or pre-recorded, can add to the quality of these performances. More modalities can be tested such as artificial reverb and facial capture. Experiments can be designed to explore the performer and audience responses with different combinations of these modalities. For instance, future experiments can study the impact of tempo stability and synchronicity in the presence of latency. Ultimately, these experiments can build towards the creation of a virtual rehearsal environment which brings together musicians in a cohesive audiovisual space.

7 Acknowledgements

The authors would like to thank all the artists who participated and provided the performance data, in particular the percussionists Chris O’Leary and Robert Mieth. We also thank the sound engineers at the NYU Steinhardt Dolan Studio, the Immersive Audio Group, and the Holodeck Project Consortium who purchased the motion capture equipment.

References

- [1] Kitagawa, M. and Windsor, B., *MoCap for artists: workflow and techniques for motion capture*, CRC Press, 2012.
- [2] Genovese, A., Gospodarek, M., and Roginska, A., “Mixed Realities: a live collaborative musical performance,” in *Audio for Virtual, Augmented and Mixed Realities: Proceedings of ICSA 2019; 5th International Conference on Spatial Audio; September 26th to 28th, 2019, Ilmenau, Germany*, pp. 159–164, 2019.
- [3] Cooperstock, J. R., “Multimodal telepresence systems,” *IEEE Signal Processing Magazine*, 28(1), pp. 77–86, 2010.
- [4] Kraut, R. E., Gergle, D., and Fussell, S. R., “The use of visual information in shared visual spaces: Informing the development of virtual co-presence,” in *Proceedings of the 2002 ACM conference on Computer supported cooperative work*, pp. 31–40, 2002.
- [5] Metcalf, C. D., Irvine, T. A., Sims, J. L., Wang, Y. L., Su, A. W., and Norris, D. O., “Complex hand dexterity: a review of biomechanical methods for measuring musical performance,” *Frontiers in psychology*, 5, p. 414, 2014.
- [6] Cheng, M., “Introducing motion-capturing technology into the music practice room as a feedback tool for working towards the precision of rubato,” *Journal of Music, Technology & Education*, 11(2), pp. 149–170, 2018.
- [7] Bonini-Baraldi, F., Bigand, E., and Pozzo, T., “Measuring Aksak Rhythm and Synchronization in Transylvanian Village Music by Using Motion Capture,” *Empirical Musicology Review*, 10(4), pp. 265–291, 2016.
- [8] Jensenius, A. R., “Methods for studying music-related body motion,” in *Springer Handbook of Systematic Musicology*, pp. 805–818, Springer, 2018.
- [9] Visi, F., Schramm, R., and Miranda, E., “Gesture in performance with traditional musical instruments and electronics: Use of embodied music cognition and multimodal motion capture to design gestural mapping strategies,” in *Proceedings of the 2014 International Workshop on Movement and Computing*, pp. 100–105, 2014.
- [10] Carlson, E., Saari, P., Burger, B., and Toivainen, P., “Dance to your own drum: Identification of musical genre and individual dancer from motion capture using machine learning,” *Journal of New Music Research*, pp. 1–16, 2020.
- [11] Holden, D., Komura, T., and Saito, J., “Phase-functioned neural networks for character control,” *ACM Transactions on Graphics (TOG)*, 36(4), pp. 1–13, 2017.
- [12] Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., and Weber, A., “Documentation mocap database hdm05,” 2007.
- [13] Guerra-Filho, G. and Biswas, A., “The human motion database: A cognitive and parametric sampling of human motion,” *Image and Vision Computing*, 30(3), pp. 251–261, 2012.
- [14] Stone, G. L., *Stick control: for the snare drummer*, Alfred Music, 2013.