Creating an Immersive Virtual Orchestra Conducting Experience

Mert Mermerci Hedvig Kjellström
KTH Royal Institute of Technology, Sweden, mermerci,hedvig@kth.se

Abstract

The role of a musical conductor is to coordinate and provide an interpretation to the music performance in an orchestra. Conducting a symphony orchestra is reserved for very few individuals. In the project described here, we would like to give more people this extraordinary experience. We are in the process of creating a virtual, immersive conducting experience in the visualization dome Wisdome Stockholm in Tekniska, the Swedish National Museum of Science and Technology. In the installation, the user (the museum visitor) will stand, wearing motion capture devices, in the middle of the dome theater, thus surrounded by the 180° projection of a recording of a symphony orchestra. The recording (made in February this year) depicts the Swedish Radio Symphony Orchestra performing the start of Beethoven's fifth symphony. The captured user motion will be fed into a gesture recognition module, which will regress a time signal indicating progress in the recording. The time signal will be fed to a video playback module that plays the recording at the pace controlled by the user motion. The gesture module will be trained with a dataset of motion from different conductors. The system will also feature some gamification in the form of orchestra reactions to success or failure. The installation will be on show in the museum in the beginning of 2025.

1. Introduction

Classical music sound production is structured by sheet music that specifies the music in some detail, but with room for variation. In larger ensembles, the interpretation of the sheet music is done by the conductor. Conducting is essentially a complex musical sign language to communicate the conductor's musical intentions and help coordinate the music production of an orchestra or ensemble, see Figure 1. There are large individual variations between the movement patterns of different conductors, but with an agreed-upon core of basic gestures [4, 5]. The role of the conductor is extremely important; even though the orchestra is an organism in itself, capable of autonomous music production, the conductor has a major responsibility for the common interpretation and coordination, and executes this by a combination of hand, arm, body motion and facial expression [6].

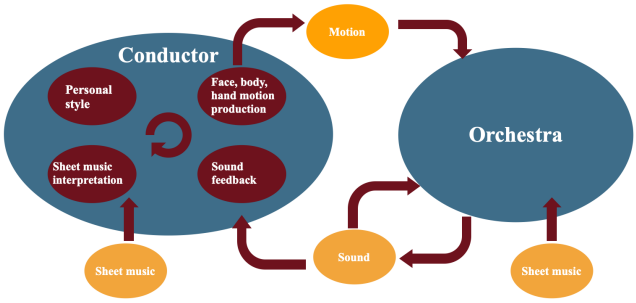

Figure 1. The conductor-orchestra interaction process, focusing on the conductor. Yellow ellipses represent observable variables.

Conducting a symphony orchestra is a quite mind-blowing experience, and reserved for very few individuals due to the resource demands; even when studying conducting at a conservatory, you get very little time with a full orchestra. In the project described here, we are creating a virtual, immersive conducting experience in the visualization dome Wisdome Stockholm in Tekniska, the Swedish National Museum of Science and Technology, which will allow a broader audience to get a first-hand experience of being a conductor.

As detailed in Section 3, the museum visitor will stand in the center of the dome theater, surrounded by a projection of recorded Swedish Radio Symphony Orchestra. The visitor's motion is tracked using wearable motion capture, and fed into a gesture recognition module which outputs a time signal that controls the pace of the recording playback. The production of the speed-controllable recording is described in Section 4 and the ongoing work of developing the gesture recognition module in Section 5.

2. Related Work

The installation here is inspired by earlier work. Nakra et al. [5] proposed the first applications in this realm; *Digital Baton* and *Conductor's Jacket*, which utilize position, acceleration, and pressure sensors to capture both physical (movement) and psychological (muscular activity, respira-

tion, and heart rate) aspects of a conductor's movements to influence the music production in real-time.

The *Personal Orchestra* by Borchers et al. [1], demonstrated at the House of Music in Vienna, provides an interactive platform for users, allowing them to conduct recordings of the Vienna Philharmonic Orchestra. This system, unlike the Conductor's Jacket designed for professional conductors, served as a public installation accessible to anyone who wishes to use it. Users can conduct recordings of the Vienna Philharmonic Orchestra, visualized on a screen in front of them. The baton emits infrared light, received and filtered by the system to extract features influencing the expressiveness of pre-recorded music. The system's emphasis on video response, coupled with real-time audio correspondence, distinguishes it from the previous ones.

Another development in the same direction is the *You're the Conductor* system [7], designed by Eric Lee in collaboration with Borchers and Nakra. This interactive conduction system, specifically tailored for children, presented at the Children's Museum of Boston, gamified conducting through hand movements, catering to diverse user groups, from professional conductors to children.

All the above mentioned conducting recognition methods are designed with a rule-based methodology, not fully capturing the motion variability in a performance and also the variability between different conductors; conducting is learnt very much by practice, and each conductor develops a personal style. Moreover, a difficult aspect to capture with a rule-based system is that timing precision is more crucial during certain moments than others in a performance.

In contrast to the previously mentioned methods, the *Home Conducting* system, introduced by Friberg et al. [2], shifts the emphasis to fluidity. By utilizing a webcam and specialized software, users can interact with the system without the need for specific devices like a baton or gloves. Musical expression and speed is derived from changes in the overall quantity of user motion.

Instead, we propose to use a data-driven learning-based method for capturing the user's gestures and transforming them into commands for music production timing and speed, see Section 5. In this way, our approach constitutes a middle ground, allowing for strict timing control when appropriate but also giving room for individual variations in motion. Moreover, the presented installation will give a more immersive experience than its predecessors, thanks to the 180° projection in the dome, see Section 3.

3. The Installation

In the installation, see Figure 2, the user (the museum visitor) will stand in the middle of the dome theater, thus surrounded by the 180° projection of a symphony orchestra.

The orchestra projection will be based on a single recording of the Swedish Radio Symphony Orchestra, recorded

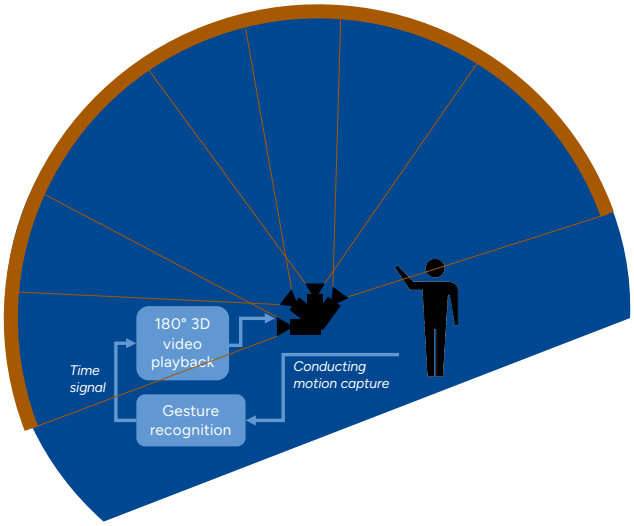

Figure 2. Museum installation set up: The motion of the visitor will be fed into a gesture recognition module, which produces a time signal that is used to control the progression of the orchestra recording.

with a wide angle camera capturing the entire hall, so that it can be played back in the dome at variable speed, see Section 4. The music in the recording is the first 90 seconds of Ludvig van Beethoven's fifth symphony, famous for its dynamic start which puts high demands on both the conductor's precision and the coordination among the orchestra musicians.

The user will wear smaller or larger parts of a Rokoko Smartsuit Pro V2 and Rokoko Smart Gloves™. (There is a trade-off between ease of use, e.g., speed of putting on the mocap devices, and gesture recognition accuracy. This will be investigated with user studies later in the project.)

Although in real life, the conductor controls many aspect of the orchestral sound, control in this installation will be limited to timing. The captured user motion will be fed into a *gesture recognition module*, see Section 5, which will regress a time signal indicating progress in the recording. The time signal will fed into the video-playback module to project the video and play the audio in the specified pace.

Music notation is built around bars, each with a specific number of beats, see Figure 3. The recording will be indexed by the musical score (using speciality software), giving a frame number for the start of each bar. In this way, the time signal can be expressed as the bar number and a measure $\in [0, 1]$ indicating progression through the bar. This can then be translated into a specific frame number in the recording. Figure 3 illustrates the time progression in a 4-beat music with constant speed. In our recording however, there are frequent stops (fermatas) when the progression is zero for some time, and then starts again by the command of the conductor. With their movements, the user will have

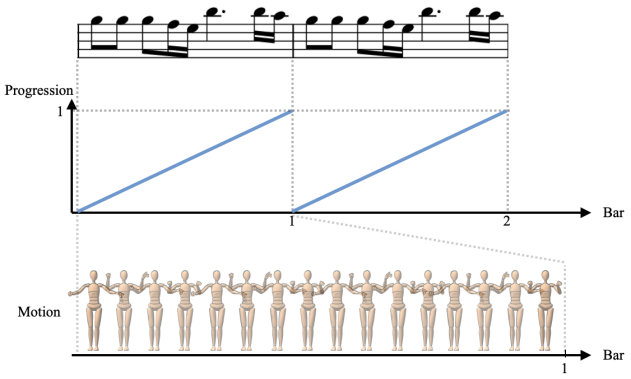

Figure 3. Progression in a 4-beat bar with fixed tempo.

Figure 4. Wide angle camera view of the orchestra which will be projected onto the dome.

the freedom to change this progression rate.

There will be an instruction video giving a crash course in conducting while the visitors wait in line. There will also be some elements of gamification where the orchestra becomes unhappy if the user fails to communicate, and applauds if the user finishes successfully.

4. The Recording

For the purpose of this installation, a recording (February 21, 2024) was made of the Swedish Radio Symphony Orchestra led by its chief conductor Daniel Harding, performing the first 90 seconds of Ludvig van Beethoven's fifth symphony. The orchestra was recorded for 2×45 minutes; several takes of the same music were made, along with extra takes for the gamification features, see the previous section. One of the takes was selected for the installation based on musical quality, and will be used in its entirety.

The recording was made by the video production company IVAR Studios, with a 270° wide angle camera covering the entire concert hall. A screenshot from the recording is presented in Figure 4. This will be edited into a dome video file that can be played back in the dome theater at variable speed, controlled by the user as described in the previous section.

The orchestra's sound technicians made a professional sound recording of the performance, which will be used for the dome video soundtrack.

In order to gather training data for the gesture recognition module, see Section 5, the motions, hand gestures, and head/gaze direction of the conductor was recorded simultaneously, using a Rokoko Smartsuit Pro V2 and Rokoko Smart Gloves™. It should be noted that the conductor is the only one not captured in the recording, since they, so-to-speak, will be replaced by the user of the museum installation. However, the conductor of the original recording is extremely important, both to shape the recorded performance and make the orchestra perform at its top, and also as a source of data for the gesture recognition module.

We will hereafter proceed to gather a wider variety of conducting motion to train the gesture recognition module. In order to get additional conducting motion for the same exact recording, we will use a mock-up procedure: Firstly, the sound recording will be provided with clicks to indicate the original conductor's impulses (primarily, starting the orchestra after fermatas). Then, 10-20 additional conductors, from experienced professional conductors to Bachelor-level conducting students, will be engaged. Each conductor will be asked to pretend that they conducted this specific recording, and be recorded with the exact same motion capture suit as the original conductor. They will not get to see the original conductor's motion, but perform their own interpretation of how to produce the same impulses. Around 10 recordings will be made of each conductor. In this way, we will have 100-200 clips of conducting motion corresponding to the exact same recording. Each clip $\mathbf{c}_i$ will be around 90 seconds long with a framerate of 60 Hz, corresponding to a length of $T = 5400$ frames.

A frame t of the clip $\mathbf{c}_i(t)$ is a representation of the conductor pose at t. The recording is synchronized in time with the pose sequence, and also with the music score. This means that for the time frame t, there is a corresponding label (b, ρ) where b is the bar number in the music score at time t, and ρ is the progression in the current bar, see Figure 3. We thus have the choice of training a bar-specific recognition model that is specific to the current bar, or a more generic model that just recognizes the progression in the bar, regardless of bar number.

5. The Gesture Recognition Module

In this section we outline the gesture recognition module under development, which transforms the user's motion into a timing signal, see Section 2.

We will train a neural network model that, incrementally over time, takes a new conducting motion and outputs the progression in the music score. The new conducting motion will not be indexed in the same way as the training clips, since the museum visitor will choose the progression speed freely. Let us call the new mocap frame $t_{\mathbf{new}}$.

The task is then to use a learning-based approach and train a neural network model that, incrementally over time $t_{\mathbf{new}}$, takes a new conducting motion $\mathbf{c}(1 : t_{\mathbf{new}})$ and outputs a time stamp in the original parameterization (frame number in the recording): $t = G(\mathbf{c}(1 : t_{\mathbf{new}}))$. If a generic model of progression in the bar is used, the model might output only the bar progression: $\rho = G(\mathbf{c}(1 : t_{\mathbf{new}}))$. The bar number b is estimated by counting from the recording start, and the recording time frame t can then be computed from (b, ρ).

The model is trained with the data set of conducting motions $[\mathbf{c}_i]$, all in the same time frame t as the recording (since they were collected in synchrony with the recording). Different implementations of the model $t = G(\mathbf{c}(1 : t_{\mathbf{new}}))$ will be evaluated, such as LSTM [3] or Transformer [9]).

The museum environment provides challenges. One will be the variation of motions in users. Several user studies will be carried out to evaluate the robustness, accuracy and usability of the system to the range of variability. Another challenge is the mocap setup. Using a textile full-body suit is not feasible for practical reasons. The most viable approach is to limit the capture to only a few points on the body, and attach markers at these points. Another opportunity is markerless motion capture. We will record each conductor movement sequence in stereo video and consider markerless methods like SMPL-X [8].

As an initial feasibility study, we recorded the motion of a professional orchestra conductor beating a 4-beat pattern with a metronome. The captured motion $\mathbf{c}$ was synchronized with the beats, and each frame was labeled with the progression ρ within the bar. An illustration is presented in Figure 3. We trained an individual specific LSTM model to regress the progression rate ρ from motion $\mathbf{c}$. 70 5-bar sequences were used for training and 7 for validation.

The LSTM network we employed consists of an LSTM layer followed by a linear layer for regression. We set the dropout rate to 0.5 for better generalization of the movement. We used Adam optimizer and trained the network for 1500 epochs with a batch size of 64. Mean Absolute Error (MAE) was used to assess performance. Since the dataset is limited, we set early stopping criteria to prevent over-fitting. The average MAE we get from the test set is 0.02893.

Figure 6 shows an example result. We see that the prediction error increases around the beginning and end of each bar. Motion is continuous so it is expected to see some error around the transition periods between bars, due to the similarity and sequential repetition of the motion.

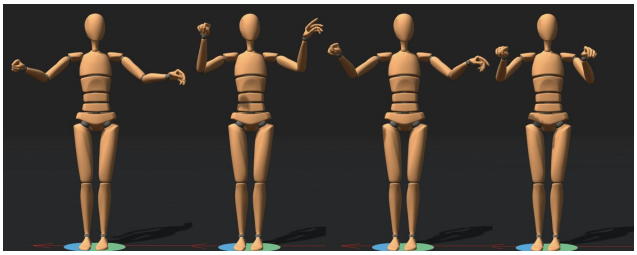

Figure 5. Sample poses from a sequence of 4-beat conducting motion.

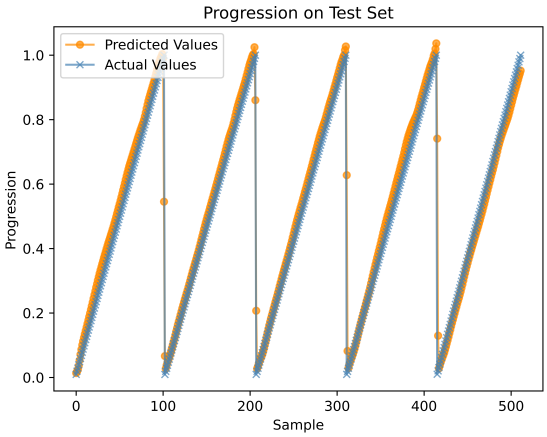

Figure 6. Estimation of progression through the bar.

6. Conclusion

This paper introduces our ongoing project where we create an immersive virtual orchestra conducting experience in the visualization dome Wisdome Stockholm in Tekniska, the Swedish National Museum of Science and Technology. Visitors will conduct a virtual orchestra, i.e., control the playback speed of a recording of an orchestra with their motion. We have currently performed the dome video recordings of the orchestra along with mocap recording of the conductor. More recordings will be performed with additional conductors to create a diverse motion dataset.

We will analyze the data to inform our choice of motion capture equipment for the installation, weighing practical usability factors against gesture recognition performance. Then, we will train an attention-based network to track the progression through the score (thus through the recording, synchronized with the score) based on user movements. A preliminary experiment with an LSTM model indicated the feasibility of this approach, accurately predicting progression in a 4-beat bar, although for a single individual.

The installation will open in 2025.

References

[1] J. Borchers, E. Lee, W. Samminger, and M. Mühlhäuser. Personal orchestra: A real-time audio/video system for interactive conducting. *Multimedia Systems*, 9, 2004. 2

[2] A. Friberg. Home conducting: Control the overall musical expression with gestures. In *International Computer Music Conference*, 2005. 2

[3] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 4

[4] Y.-F. Huang, S. Coleman, E. Barnhill, R. MacDonald, and N. Moran. How do conductors' movements communicate compositional features and interpretational intentions? *Psychomusicology: Music, Mind, and Brain*, 27(3), 2017. 1

[5] G. Johanssen and T. Marrin Nakra. Conductors' gestures and their mapping to sound synthesis. In *Musical Gestures: Sound, Movement, and Meaning*. 2009. 1

[6] K. Karipidou*, J. Ahnlund*, A. Friberg, S. Alexanderson, and H. Kjellström. Computer analysis of sentiment interpretation in musical conducting. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2017. (*joint first authors). 1

[7] E. Lee, T. Nakra, and J. Borchers. You're the conductor: A realistic interactive conducting system for children. In *New Interfaces for Musical Expression*, 2004. 2

[8] G. Pavlakos*, V. Choutas*, N. Ghorbani T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. (*joint first authors). 4

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 4