

CAN A ROBOT GET SMARTER BY LISTENING TO ITSELF? MUSICAL MEMORY AS AN EXTENDED AUDITORY-NEURAL-MOTOR LOOP

ABSTRACT

We present a novel approach to robotic musical memory, based on the idea of extended auditory cognition, as well as on recent evidence for links between motor and auditory cortexes in the human brain during musical performance and recall. Given a recording containing the motor control signal of a drumming robot playing a rhythmic pattern, and the resulting acoustic signal, we show that a feed-forward artificial neural network can be trained to accurately predict future motor movements from the history of the audio signal. In addition, we show that such a network outperforms a similar network trained to predict the motor signal from the history of the motor signal itself; the robot can more easily learn a sequence of movements given its resulting sound. Such a network was implemented to control the real-time behavior of a drumming robot. We demonstrate that the robot is able to accurately reproduce a given set of rhythmic patterns, solely by feeding back information from the acoustic signal produced by it. The robot can be triggered to complete the pattern by exposure to parts of the pattern, either in the form of sound or in the form of movement. Moreover, the robot is able to recover from disturbances to the pattern. This makes the system especially appropriate for real-world musical interaction.

1. INTRODUCTION

For both humans and robots, performing a musical pattern requires the completion of a complex sequence of physical movements. An accurate performance of a movement sequence results in the generation of a specific sound sequence. Thus, teaching a robot to play a specific musical pattern is equivalent to teaching the robot to execute a specific sequence of actions in time, thereby generating a sound pattern. One can easily teach a robot to play a sequence, for example by storing the action sequence in a computer file and use the file to drive the robot, one action at a time. Such an approach is what we call a dualistic approach, meaning that sound is not playing a part in the process of musical recollection, but is merely the result of recollection. However, common knowledge, and a growing body of scientific evidence suggests that such a

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

© 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems

dualistic approach does not capture the way humans recall musical patterns [1, 11].

Imagine someone asking you to sing the 10th note of the tune Mary Had a Little Lamb. Most likely you would begin humming the first few notes of the tune, which would bring to mind the rest of the song, in the correct temporal order. The tactile and auditory feedback you would get from your own actions would create a memory retrieving loop that would help you accomplish the task. Such a loop is what we term an extended auditory-motor loop.

2. BACKGROUND

The motivation for the work is to work towards creating a flexible computational model of musical memory that allows artificial agents, to perform and interact musically with humans. From a theoretical perspective we are interested in exploring how the integration of multiple information streams of different modalities can lead to a more accurate prediction in musical contexts. From a broader perspective, musical memory and performance are critically intertwined when humans make music. This paper is a step towards modeling musical memory and musical performance as a single unified system.

2.1 Extended Cognition

Extended cognition is a term coined by Clark and Chalmers [2] as a part of their extended mind theory, which addresses the dividing point between the mind and the environment. The key idea of this theory is that the mind, through the movements of the body, utilizes objects in the external environment so as to aid, augment, and even create cognitive processes. According to this view, the mind, the body and external objects can act as a coupled system which can be seen as a complete cognitive system of its own. Clark and Chalmers do not focus on sound production in particular. However, we are of the opinion that sound production offers a unique example of fast, reliable and resonant external feedback for human movement which helps humans to recall complex action patterns. This idea, that motor and auditory information are used in parallel to bring about musical memory is largely supported by recent neurological evidence [1, 11].

2.2 Evidence from Neuroscience

Over the last few years there has been a growing evidence for the involvement of both auditory and motor brain areas in music performance and appreciation [6, 9, 10]. These neuroimaging experiments all suggest the existence of a

plastic, active, experience-dependent neural network between motor and auditory brain areas underlying musical activity (for an overview of this research see [7]). In a fascinating experiment, Lahav and colleagues [4] trained non-musicians to play a simple melody on the piano and showed that later exposure to a recording of the melody created an activation of the frontoparietal motor-related network. This may indicate that learning to perform a sound producing action sequence results in a specific functional neural link between motor and auditory representations of the sequence.

Building the interdependency of motor and auditory brain areas during music perception and interaction, and in particular animal model evidence for auditory "mirror neurons", Overy and Molnar Szakcs [7] [8] have proposed the Shared Affective Motion Experience (SAME) model of musical affect. Overy and Monlar-Szakcs stress the importance of viewing music not as an auditory signal but rather as an intentional, hierarchically organized sequence of expressive motor acts. Emotional response to music, according to this model, is the result of synchronizing motor movements, real or imaginary, to auditory inputs.

This paper presents a modest first attempt at real time robotic implementation of these ideas. We are not in a position to propose a full model of musical memory based on extended cognition, but we are able to show that auditory feedback clearly enhances the memory capacity of a drumming robot controlled by a feed-forward artificial neural network. In the next sections we describe how a robot is trained on a set of auditory-motor examples, and how it regenerates the patterns using an auditory-motor-neural loop.

3. METHOD

The first problem we seek to address was to show the benefit of auditory-motor action. To do this we recorded control and audio data of a robot playing a rhythmic pattern. Using the training data, we perform a simulation that compares the prediction of the motor control of the robot from its past motor control to the prediction of the robot's control from the audio that the robot generates as a result of playing the rhythm on a drum. Then the second problem is whether or not the robot could, by listening to itself, predict the control signal from a live audio stream and actually complete the pattern; could the robot, by listening to itself, know what to play next?

3.1 Haile The Drumming Robot

Haile is a drumming robot built by Driscoll and Weinberg [5]. It has two arms, the right arm controlled by a solenoid with a simple pad to strike against the drum and the left arm whose velocity is regulated by a microcontroller and whose striker is more round to provide greater dynamic range. We used the right arm because it operated with negligible lag time, where as the other arm has a fixed 200ms delay hardwired into its functionality. The drum that Haile uses is a Native American Pow Wow.

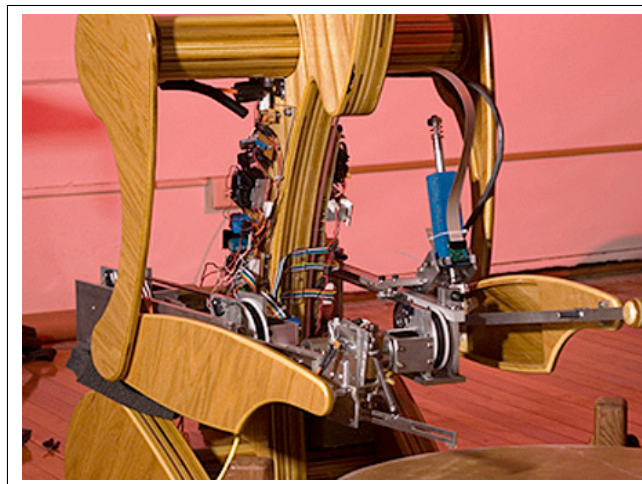


Figure 1. A Picture of Haile

Pattern Representation
1:...xxxxx..xx.x.x: 2:...xxx.x.xxx..xx: 3:...x.xxx.xxx..xx:
4:...x.x.xxxxx..xx: 5:...x..xx.x.xxxxx: 6:...xxx.xxx.xx..x:
7:...x.xxxx.xx..xx: 8:...xx..xxxxx.x.x: 9:...xx..x.xxx.xxx:
10:...x.xxx.xxxx..x: 11:...xxx.xx..xx.xx: 12:...xx.xxxx.x..xx:
13:...xx.xx.xxxx..x: 14:...xx..xx.xx.xxx: 15:...x..xxx.xxx.xx:
16:...xx.xxxx.xx..x: 17:...xx.xxx.xxx..x: 18:...xx.xxx..xx.xx:
19:...xx..xx.xxxx.x: 20:...xx..xx.xxx.xx: 21:...xxxxx.xx.x..x:
22:...xxxx.x..xxx.x: 23:...xxx..xx.xxx.x: 24:...x.xxx.x.xxxx:
25:...x.x..xxx.xxx: 26:...xxxx.x.x..xxx: 27:...xx.xxx.x..xxx:
28:...xx.x..xxx.xxx: 29:...x.xxxx.x..xxx: 30:...x..xxxxx.xx.x:
31:...xxxx.xxx..x.x: 32:...xxxx..xx.xx.x: 33:...xx.xxxx..xx.x:
34:...xx.x..xxxxx.x: 35:...x.x..xxx.xxxx:

Table 1. Povel-Essens Patterns

A message for Haile to strike includes the delay with which to strike and then also to retract back up to the ready position. Changing these delay parameters can cause some changes in amplitude, though we keep them at fixed values such that arm can consistently hit the drum without missing and with a consistent amplitude. The message is sent via ethernet UDP through a Max/MSP patch to the robot. Haile operates in a discretized fashion, striking and retracting only when a complete message is sent.

3.2 Training Data

A training database was constructed containing the motor control signal for the robot and corresponding audio for a set of rhythmic patterns. Thirty-five patterns taken from Povel-Essens (PE) [3] were used and are shown in Table 1.

These were chosen because they contain patterns of varying complexity are widely known in the music cognition community. Each pattern was encoded in a text file using a simple format which was parsed using Max. A period indicates a silence and an x indicates a strike as shown in Table 1, A metronome object is used to cycle through the instructions of the pattern at a rate of 100ms. This results in the robot completing one iteration of the pattern in ap-

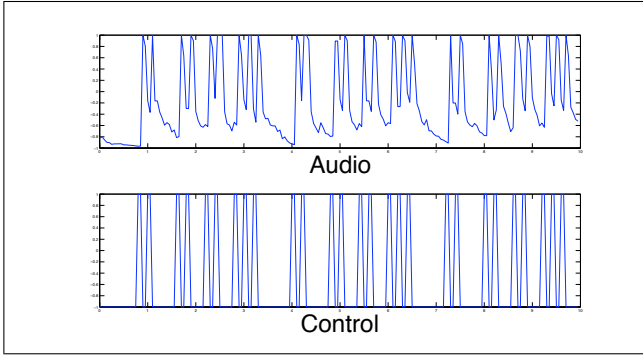


Figure 2. Example of Training Data

proximately 3 seconds. The robot plays each pattern for a fixed number of repetitions before iterating to the next in the text file. A monophonic audio recording was made of the robot drumming. As mentioned earlier, the robot is only triggered in a discrete fashion, meaning it only hits the drum and retracts when given a complete message. To deal with this, the control signal is recorded continuously and at the same rate as the audio so that both will be the same length and correspond in time to one another. The control signal itself is binary with values of -1 and 1. Then, when processing the continuous control signal we include logic such that the message for a complete hit will be sent to the robot only when the control signal transitions from -1 to 1. Thus we can have a control signal at the same sampling rate as the audio file. The audio and control signal are both sampled using the "peak" operation in Max which is a max filter that samples a 50ms window (every 2205 samples) of the 44.1kHz audio, and likewise for the corresponding control signal. Both the control signal and audio signal are then stored in the left and right channel of a single stereo wave file as shown in Figure 2.

3.3 Training Simulation

We wish to predict future values of the motor signal given past values of the corresponding audio signal. We can put this experiment in a standard machine learning framework for time-series prediction. First a window size (N) that determines the amount of history to use in the prediction is specified. The first input vector consists of the first N samples of the audio signal. Next we hop by one sample and form another length N vector.

$$\begin{aligned}
 a(1)\dots a(N) &: c(N + 1) \\
 a(2)\dots a(N + 1) &: c(N + 2) \\
 \dots & \\
 a(n)\dots a(N + n - 1) &: c(N + n) \\
 a(n - N)\dots a(n - 1) &: c(n) \\
 a(n - N)\dots a(n) &: c(n + 1)
 \end{aligned}$$

Thus, for each vector, we would like to predict the control signal at the next time step. For example, if we let

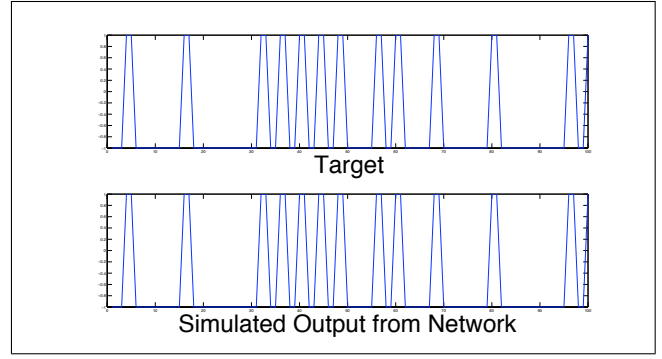


Figure 3. Network Training Perfectly

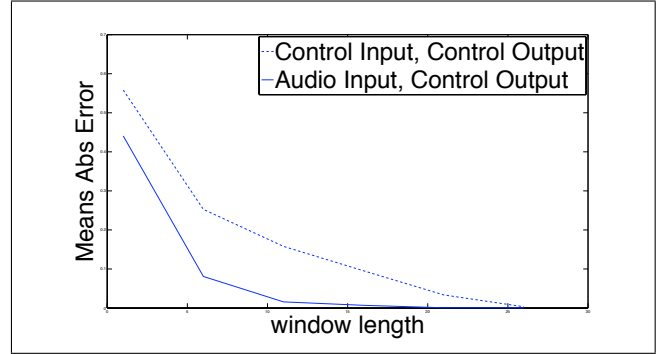


Figure 4. Comparison of Two Networks

n th sample of the corresponding control signal be $c(n)$ then we would like to predict $c(N + 1)$ given $a(1)\dots a(N)$.

A standard architecture for such a problem is a feed-forward neural network with N input nodes, corresponding to the dimension of the input audio vectors described above, and one output node, corresponding to the target control value. The number of hidden layers and nodes in each hidden layer is a parameter that is usually determined empirically, trading off between the greater expressive power of more complex networks, and the tendency of complex networks to overfit training data. After some experimentation, we used a network with 2 hidden layers, with 17 and 11 nodes respectively. Each node used a hyperbolic tangent sigmoid transfer function. The network was implemented using the MATLAB neural network toolbox. Training was accomplished with the Levenberg-Marquardt algorithm, `trainLM`, with an error threshold of 10^{-5} . Sometimes the training algorithm would get stuck in local minima and then would be restarted, but all 35 patterns were able to reach this error criterion; therefore the network was able to learn the input-output mapping. Figure 3 shows the output of the simulated network and the true target values for one pattern. In some cases the training algorithm became stuck in a local minimum and was restarted. Thus for each pattern a network whose weights encoded the audio-control input-output relationship was created.

3.4 Audio to Control Comparison

We then performed a comparison to establish whether there was any advantage in this method compared to simply pre-

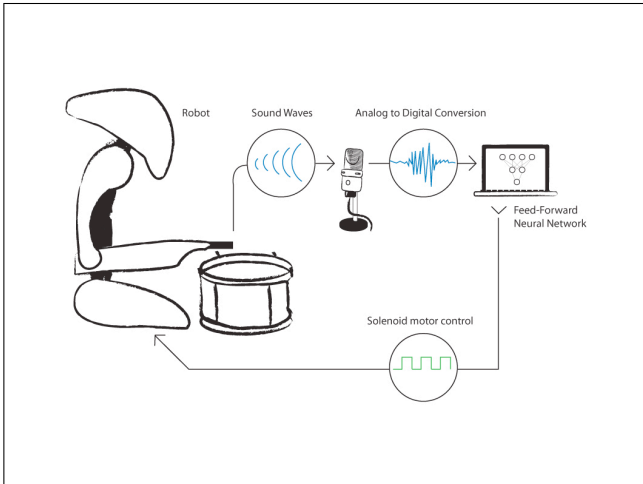


Figure 5. Feedback Loop

dicting future actions based on past actions (i.e. FFNN-based time-series prediction on the control signal alone). To test this the performance of a control-to-audio network and control-to-control network were compared for various window lengths. It can be seen in Figure 4 that audio-to-control outperforms control-to-control. This was true for all PE patterns when for window lengths typically between 15 and 30 samples (.75 and 1.5 seconds).

3.5 Application to the Robot

Using the same training data, we tested the performance of the network in a live scenario as opposed to just simulation. The hope is that the network will be able to generalize to the subtleties present in real world data; though the performer of the drum is a robot, the audio resulting from the drum will not be exactly the same each time it is hit. Because of the increased difficulty of this task, the window length was set to 50 samples for these experiments.

3.5.1 Passive Mode

The first step in this process was to verify that the control signal generated by the neural network, using the training audio as input, could be used to cause the robot to play the pattern. Since this is merely a verification of the training data we called this "passive mode". The training audio data is fed from the file into the neural network which generates a control signal and causes the robot to strike the drum, playing the pattern it has learned until the recorded audio data is finished. A video of the robot playing a pattern in passive mode can be seen at <http://ANONYMOUS.com/research/ismir10robot/>.

3.5.2 Active Mode

The next step after passive mode verification is to complete the feedback loop as shown in Figure 5 and allow the robot to listen to itself while playing. We called this "active mode" since now the input is coming in live from a microphone source as opposed to a file; in short, active mode uses testing data (from the microphone), and hopefully the network performance will be similar as when we

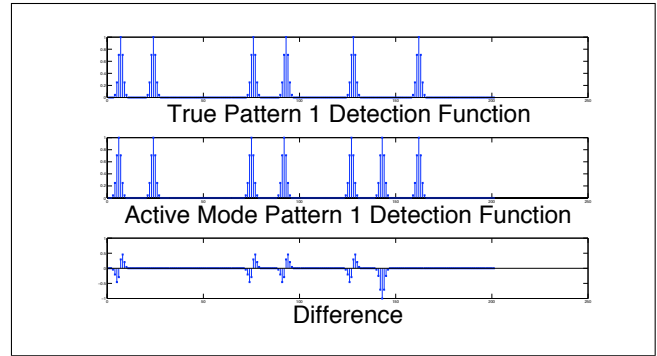


Figure 6. Example of a False Onset

used training data (data from the file). The robot is initially primed in passive mode, where the training audio from the file is fed into the network, and a control signal is produced. Upon switching to active mode, the input source to the network is changed to the microphone, the same one which was used previously to record the training data. Within the Max/MSP patch this live stream of audio is filtered with the peak operation, in the same manner that the audio was recorded for obtaining the training data, and then fed into the network. The network produces the control signal, which then causes the robot to strike the drum, producing more audio which then feeds back into the network. The crucial difference between passive and active mode is that passive mode is merely a simulation using data that the network has already seen. Active mode is the proof that the network can generalize. Active mode is much more difficult not only because of the natural variance of the audio compared with the training signal, but also because errors are compounded. For example, incorrectly predicting the control signal can lead to a false hit that will significantly alter the subsequent audio. This can easily lead to the robot losing the pattern.

3.5.3 Active Mode Measure

We found that the audio-to-control network was able to recreate all of the 35 PE patterns faithfully in active mode. Qualitatively, the switch from passive to active mode is seamless and the robot is able to continue a test pattern exactly as it had been trained to do. For completeness, we assessed how well active mode performed in comparison to a passive mode audio recording of the robot. A spectral difference detection function was used on audio frames captured from both recordings. The result was thresholded to obtain a form of a detection function that contains only impulses at the onsets. Then this result was convolved with a Gaussian window of length 7 and variance 22.5 (using the command `Gausswin(7)` in MATLAB). The two Gaussian-convolved detection functions were then subtracted from one another to obtain a difference. The reason for doing this is to be able to distinguish between very minor time offsets, which are not errors, and falsely added onsets or missing offsets, which are errors. If two of the Gaussian windows are at the same location in time, the difference is zero. If they are one sample off, a difference of the Gaus-

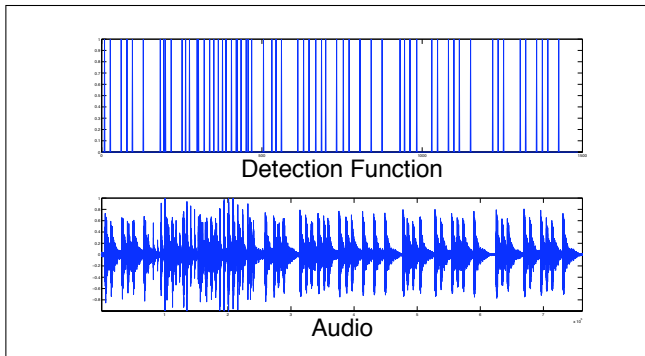


Figure 7. Disturbance and Recovery for Pattern 23

sians, resembling a sinusoid, will be visible that will have a max of .5 and -.5. This maximum will increase to -1 and 1 the more offset the two Gaussian onsets. An error, which occurs when an onset is missing or falsely added, will not have a corresponding twin and thus the detection function will result in just one half of a wave moving towards -1 if there is an extra onset or 1 if there is a deletion. Figure 6 shows the results of this process for PE. It can be seen that the added onset is clearly detectable while allowing minor shifts to be ignored.

Pattern	Precision	Recall	F-Measure
1	1	.9778	.9888
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
...
34	1	1	1
35	1	1	1

Table 2. F Measure

We then compared the detected onsets in each, calculating the precision and recall. The F measure,

$$F = 2 \frac{(Precision * Recall)}{(Precision + Recall)}, \quad (1)$$

for each pattern is given in Table 2. The F measure is 1 for all 35 patterns except for PE 1. This very high F measure is reasonable considering that audio generated from the robot’s audio to control loop and the original recording of the pattern sounded the same. It is important to note that this is not simply a comparison of similar test data. Had the neural network been trained poorly, say with only 1 or 2 epochs, the network performance would be poor, which would cause the robot to make errors in control signal prediction, which would then cause the audio generated to be unfaithful to the recorded audio in the training data.

A final qualitative experiment was performed to determine if the network could recover from disturbances to the input stream. While in active mode, the experimenter would cause a disturbance, by either clapping into the microphone or playing on the drum itself. This resulted in

confusing data sent through the network which causes it to lose the pattern. Once the pattern is lost, the disturbance is stopped and we wait for the network to recover the pattern. After a variable amount of time, which is ultimately dependent upon the type of the disturbance and the time within the pattern where it was launched, the network is able to recover the pattern. Figure 7 shows an example of this where the pattern 23 is disturbed and then recovers after a few seconds. Additionally, a video example of this can be seen at <http://ANONYMOUS.com/ismir10robot/>.

What is also interesting to note is that for some patterns after a disturbance, the network would lock into a stable, repeatable pattern that was not actually the one it upon which it was trained. One audio example can be heard where only one network was trained, but 3 unique patterns were discovered. Therefore it is possible to train the network on multiple patterns and have the network be able to recall each of them separately when prompted. A video example showing Multiple patterns can be seen at <http://ANONYMOUS.com/ismir10robot/>, where the system has been trained both upon Povel Essens patterns 1 and 2. The ability to recover from a pattern and lock to other types of patterns implies potential for a very fast synchronization system.

4. CONCLUSIONS AND FUTURE WORK

This paper is a first step towards a full model of robotic musical memory that utilizes the advantages inherent in the combination of auditory and motor information. We have shown through simulation that an audio-control modality can outperform control-control for certain window sizes. We have demonstrated that auditory feedback creates a robust musical memory loop. Additionally, we have already made preliminary experiments that show that such a loop can be trained on multiple patterns. With such a network the robot can complete different patterns when exposed to sub-sections of them.

The next step in developing the model would be to bring back the motor information into the network and create a complex network which will generate auditory and motor information based on the history of both. We predict that such a hybrid motor-auditory network will prove to be very flexible and powerful for creating musical interaction between robots and humans. We are also currently working on applying the same architecture to human-generated data. This would mean training a network on a set of examples of hand movement and resulting audio signal, collected while a human subject is playing a drum.

Finally, we would like to posit that, in concert with previous search and the work conducted here, the informational advantage of the extended auditory cognition approach is not unique to the specific system, be it the robot or the artificial neural network, and may in fact be an essential part of human interaction with sound. Philosophically speaking, listening or playing music could be viewed as an extended auditory cognition cycle, from which musical computation, musical memory and musical communication emerge.

5. ACKNOWLEDGMENTS

6. REFERENCES

- [1] J.L. Chen, R.J. Zatorre, and V.B. Penhune. Moving on time: brain network for auditory-motor synchronization is modulated by rhythm complexity and musical training. *Journal of Cognitive Neuroscience*, pages 226–239, 2008.
- [2] A. Clark and Chalmers D. The extended mind. *Analysis*, 58:7–19, 1995.
- [3] Povel D. and Essens P. Perception of temporal pattern. *Music Perception*, 2(4):411–440, 1985.
- [4] Lahav A. Saltzman E. and Schlaug G. Action representation of sound: Audiomotor recognition network while listening to newly acquired actions. *Journal of Neuroscience*, 27:308–314, 2007.
- [5] Weinberg G. and Driscoll S. Toward robotic musicianship. *Computer Music Journal*, 30(4):28–45, 2006.
- [6] H. Haslinger B. Erhard P. Altenmuller E. Schroeder U. Boecker and Ceballos-Baumann A. O. Transmodal sensorimotor networks during action observation in professional pianists. *Journal of Cognitive Neuroscience*, 19:893–906, 2005.
- [7] Molnare-Szakacs I. and Overy K. Music and mirror neurons: From motion to 'e'motion. *Socialcognitive and Affective Neuroscience*, 1:235–241, 2006.
- [8] Molnare-Szakacs I. and Overy K. Being together in time: Musical experience and the mirror neuron system. *Music Perception*, 26:489–504, 2009.
- [9] Blood A. J. and Zatorre R. J. Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences*, pages 11818–11823, 2001.
- [10] T. Muller K. Koelsch, S. Fritz and Friederici A. D. Investigating emotion with music: An fmri study. *Human Brain Mapping*, 27:239–250, 2006.
- [11] R.J Zatorre, J.L. Chen, and V.B. Penhune. When the brain plays music. auditory-motor interactions in music perception and production. *Nature Reviews Neuroscience*, pages 547–558, 2007.