An Artificial Intelligence Architecture for Musical Expressiveness that Learns by Imitation

Axel Tidemann
IDI, Norwegian University of Science and Technology
Sem Sælands vei 7-9
7491 Trondheim, Norway
axel.tidemann@gmail.com

ABSTRACT

Interacting with musical avatars have been increasingly popular over the years, with the introduction of games like Guitar Hero and Rock Band. These games provide MIDI-equipped controllers that look like their real-world counterparts (e.g. MIDI guitar, MIDI drumkit) that the users play to control their designated avatar in the game. The performance of the user is measured against a score that needs to be followed. However, the avatar does not move in response to how the user plays, it follows some predefined movement pattern. If the user plays badly, the game ends with the avatar ending the performance (i.e. throwing the guitar on the floor). The gaming experience would increase if the avatar would move in accordance with user input. This paper presents an architecture that couples musical input with body movement. Using imitation learning, a simulated human robot learns to play the drums like human drummers do, both visually and auditory. Learning data is recorded using MIDI and motion tracking. The system uses an artificial intelligence approach to implement imitation learning, employing artificial neural networks.

Keywords

Modeling Human Behaviour, Drumming, Artificial Intelligence

1. INTRODUCTION

The ubiquity of cheap processing power and new physical interfaces has led to the introduction of novel applications when it comes to expressive music performance in the digital realm. Although computers have been used for musical purposes for decades, they have become more prominent in popular culture with the introduction of games like Guitar Hero[1] and Rock Band[2]. In these games, the user plays along with a score displayed on the screen. The user performs with MIDI interfaces that look like real instruments, such as a guitar[3] or a drum kit. As part of the game, animated musicians play the different musical instruments in the song. However, these animated musicians (or avatars) do not move in accordance with the user input. If the user makes an error, it is not reflected in the behaviour of the corresponding avatar. The only way the avatar reacts to the input of the user is if the user performs poorly to the extent that the game is terminated before the song is over; the avatar subsequently throws the guitar on the floor.

These games would greatly benefit from some way to move the corresponding avatar in accordance with user input, with natural movement as a result. This would enhance the gaming experience. This paper presents an architecture that uses learning by imitation to move a simulated robot based on musical input. The system learns to play drums like human drummers do. The architecture is divided into two subsystems; a sound system that imitates the playing style (i.e. it sounds like a human drummer) and a motor system that generates the corresponding arm movements. Both systems use imitation as the learning principle. By seeing and hearing human drummers, the system is able to imitate their playing style. Why use imitation as the learning mechanism? First of all, this is a way that humans transfer motor knowledge between individuals. The ability to imitate is without a doubt a cornerstone of human society. Secondly, when trying to make a machine learn a human quality such as musical expressiveness, it makes sense to use the same mechanism as that of humans. Instead of trying to formulate human behaviour using mathematical formulas, it is more intuitive to simply *show* the machine what it should do. Furthermore, learning by imitation implies an internalization (i.e. a *model*) of the acquired knowledge. An artificial drummer that merely plays back a recording is not of great interest, neither expressively nor research-wise. The machine uses a learned model to generate new music, that will be *similar* to the original, but not *identical*. These are the main reasons imitation learning is employed in the architecture, which uses an artificial intelligence approach to implement imitation learning.

2. BACKGROUND: IMITATION LEARNING

Imitation learning has been extensively studied in psychology and is considered an important part of human society [17, 14]. The discovery of *mirror neurons* was considered as a possible “neural candidate” for the imitative capability in the human brain [19]. Mirror neurons were found to be active both during observation and production of the same movement. The mirror neurons were also hypothesized to be the neural mechanism behind empathy, allowing humans to transform their viewpoint into that of others [5]. However, recent studies have questioned the comparison between a mirror neuron system in monkeys and humans [12]; mirror neurons remain controversial.

In the artificial intelligence community, imitation learning has gained momentum as a way to program desired behaviours in robots. Schaal [21] suggests model-based ap-

[1] hub.guitarhero.com
[2] www.rockband.com
[3] Fender released a real guitar on March 1st, 2011 that can be played as a controller for Rock Band.

proaches as the best way to implement imitative behaviour; this consists of pairing an inverse model (controller) with a forward model (predictor), an approach that stems from control literature [11]. Wolpert et al. argue that such inverse/forward couplings are present in the cerebellum [28], leading to an architecture based on those principles. Demiris et al. have also investigated an imitative architecture based on such inverse/forward pairings [4]; there are some fMRI studies suggesting such an ordering is present in the brain [9].

There are other modular architectures for imitation learning that take a slightly different route by defining modules for different stages of sensorimotor processing, such as perception, recognition and action selection [6, 13]. Some researchers focus solely on neural network architectures designed for imitation learning [22, 2, 1].

In music, it is evident how humans imitate others when learning to play instruments. In the cross section between music technology, machine learning and music performance are systems that focus on capturing human expressiveness. Saunders et al. [20] use string kernels as a classification method for pianists. The string kernels are used to modify changes in tempo and velocity when playing a classical piece of music. Tobudic and Widmer use first-order logic to describe the same changes [26], the system can subsequently be used to classify pianists based on their playing style. Case-Based Reasoning (an artificial intelligence method where known solutions to old problems are re-used to find solutions to new problems) have been used to model human expressiveness, such as mood [3] and how the tempo can change, but still maintain the original sentiment [7]. Pachet [16] has a system called "The Continuator" that employs Hidden Markov Models to predict the next note; this is a real-time system that can be used to interact with other musicians. Raphael [18] has a system that allows a soloist to practice along with a computer playing a score; the system learns how the soloist varies the tempo over time, and plays along with the tempo drift. In the music software industry, sophisticated drum sample software (e.g. FXpansion BFD, Toontrack EZdrummer, DigiDesign Strike, Reason Drum Kits, Native Instruments Battery) contain gigabytes of samples, but no *intelligent* way of creating human-like drum tracks, apart from adding random noise that is to be perceived as human. The research in this paper addresses this issue.

3. ARCHITECTURE

The architecture that implements the artificial drummer is called "Software for Hierarchical Extraction and Imitation of Drum Patterns in a Learning Agent" (SHEILA). It is comprised of two subsystems, a sound system that imitates the playing style (i.e. what you can *hear*) and a motor system that imitates the corresponding motor actions (i.e. what you can *see*). How the two subsystems interact can be seen in figure 1. This separation reveals a simplification: the sound system can be used as a groovy drum machine by itself, since it outputs the imitated sound. The motor system generates the corresponding arm movements of the drummer. This separation was made for two reasons: development-wise, it was easier to make a clear division between sound and motor actions. Secondly, this frees up the necessity of simulating physical drums as well. The artificial drummer will move its arms in accordance with the sound that is produced, however the movement of the arms does not generate sound by hitting a drum. If the sound were to be generated by the arm movement, the problem would be vastly more complex, requiring a model of physical drums. For an on-screen avatar this is not necessary - for the end user of the system, the movement and sound from the artificial drummer will be realistic. The different subsystems will now be presented.

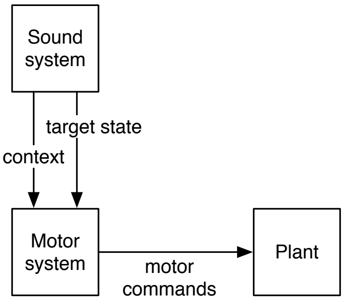

Figure 1: A simplified overview of the architecture: the sound system produces sound signals, as well as driving the motor system. The motor system issues motor commands to achieve the movements implied by the sound signals.

3.1 The Sound System

The sound system learns user-specific variations from human drummers. An important aspect of human drumming is the introduction of *variations*. The drummer can play small-scale variations, e.g. varying the velocity (how hard a note is played) and timing (how much the note is before or after the metronome). The drummer can also add large-scale variations, such as altering the pattern played altogether. This is often referred to as a *break*, something the drummer does for rhythmic and dramatic effect, adding dynamics to a song. The small- and large-scale variations add up to the *groove* of the drummer, which is what the sound system imitates.

MIDI recordings of human drummers provide data that the system is trained on. Drum patterns are analyzed in a hierarchical manner: the MIDI drum sequence is transformed into a string. Similar patterns are found in the string by searching for supermaximal repeats, a method used to search for sequences in genes [8]. This method allows similar patterns to be extracted from the MIDI stream. The patterns are used to train Echo State Networks (ESNs) [10], a neural network architecture characterized by its huge memory capacity and fast training algorithm. These ESNs are not driven by input, they are self-generating networks; the networks use feedback connections from the output layer to reverberate in the desired state. The ESNs can be thought of as having a pulse that generates the desired groove after learning. More details can be found in [23].

3.2 The Motor System

The motor system is responsible for the imitation of arm movements. The approach is to pair an inverse model (a controller) with a forward model (a predictor), an approach well known in robot control literature [11]. The motor system uses several such pairs of inverse and forward models. The motor system is in turn inspired by two other architectures for motor control and learning that use multiple paired inverse and forward models [28, 4]. See [25] for more details.

3.3 Combining the Motor and Sound System

To create an animated artificial drummer that both sounds and looks like a real drummer, the two subsystems are connected to provide sound and animation. The output of the sound system is used to create the sound, but also to *drive* the motor system. The actual sound output is used as the *desired state* for the motor system. The inverse model receives signals that describe what the *end result* of the movement should be. This sound signal is in a different coordinate system than that of the current state of the motor system, which makes it harder for the inverse model to learn the corresponding relationships.

4. EXPERIMENTAL SETUP

In order to train the system, five human drummers were told to play specific patterns along with a song written by the author. The drummers could then introduce large-scale variations as they saw fit. MIDI was recorded using a velocity sensitive electronic drumkit, the Roland TD-3. Motion tracking was done with a Pro Reflex tracking system. Pro Reflex makes use of infrared cameras to track position of fluorescent markers over time. Using motion tracking effectively solves the correspondence problem [15], since the recorded 3D coordinates could be mapped directly to the artificial drummer. The robot arm was implemented as a four degrees of freedom (DOF) model based on the human arm [27] (a 3DOF spherical shoulder joint, 1DOF revolute elbow joint). The entire robot was described by 8DOF.

5. DISCUSSION

After the recording and training of the system, SHEILA was used to imitate the playing style of the human drummers that served as teacher. By performing statistical analysis on the resulting drum patterns, it was revealed that the imitated drum patterns are similar, but not identical. Further detailed results of the sound system can be found in [23].

The performance of the motor system was also very good. When comparing recorded training data with performance data, the error was less than 0.05%. The motor system relies heavily on biological properties such as self-organization during learning; it is an AI architecture for motor control and learning in itself. The self-organizing properties have been thoroughly investigated elsewhere, see [25, 24]. An example of the imitative capabilities can be seen online[4].

However, the focus of this paper is how this combination of AI subsystems can be used for musical expressiveness, and in particular in games like Guitar Hero and Rock Band. Why use a computationally expensive artificial intelligence approach, instead of simply playing back a recording of the desired behaviour? First of all, such an approach would yield an identical result each time it is used. By employing imitation learning, the generated drum patterns will sound similar, but not identical. Furthermore, in order to truly imitate human movement, it is imperative that the underlying approach is biologically inspired. For this reason, the research in this paper is multi-disciplinary; it focuses on imitation of musical expressiveness using artificial intelligence mechanisms that can faithfully reproduce this human behaviour.

The sound system was designed around a more pragmatic, hierarchical approach. However, it was implemented using Echo State Networks, which are modeled on the neural networks present in our brains. In order to implement a human quality such as groove, it makes sense to implement this capability using a biologically inspired method.

The motor system was more directly inspired by existing neuroscientific models of how the brain operate [28]. Motor control and learning have been a focal point for AI research for decades; an architecture that is to implement this ability would benefit from an approach based on neuroscientific principles. The research in this paper was done on a simulated robot, since a real robot with the agility equal to humans is prohibitively expensive. However, one can envision that in the future robot technology will be cheaper and with greater dexterity. The architecture could then be employed on a real robot, since its design is based on robot control mechanisms [11]: the continuous outputs of the inverse models (i.e. Echo State Networks) could easily be converted to voltages used to drive a real robot.

A key element is that the architecture is in principle independent of what kind of instrument it is supposed to imitate. Both the sound and motor system are independent of the drumming domain. As long as there is some repetitive melodic structure (e.g. guitar riffs and bass lines), the sound system can model it. Motion tracking can be used on various parts of the body. Why was drumming chosen as the application? There are two main reasons: 1) Playing drums is very repetitive, where the pattern is normally reproduced every bar. For melodic instruments, the repeated pattern (i.e. melody) can last longer. The makes it easier to learn models of a particular playing style, and made for a good starting point when exploring this research path. 2) Imitating the movement of the drummer could be limited to the arms only. Granted, the drummer invariably moves the entire body, however the arms will provide a sufficient subset of the body movement in order to imitate a playing style, since a drummer is stationary during playing. The movement of the arms is also easy to visualize. In the case of guitarists, the playing style to be imitated can sometimes involve more of the entire body. Extreme examples are the particular walk of AC/DC guitarist Angus Young, Jimi Hendrix playing the guitar behind his back, or The Who's guitarist Pete Townshend who plays the guitar with a "windmill" motion. These are prime examples of the possibility to imitate the playing style of guitarists.

Given the independence of SHEILA regarding which instrument to imitate, it could be employed in imitative settings in other applications. When it comes games like Guitar Hero and Rock band, two possible ways of implementing the architecture could be envisioned: first, musicians on screen that are *not* controlled by humans could be implemented using SHEILA. The whole point of these games is to give the illusion of playing in a live rock band. If all the other computer controlled characters were implemented using SHEILA, their performance would be slightly different each time, but still recognizable. No human musician plays a musical piece exactly the same way twice, so this would greatly add to the feeling of realism of playing along with other characters. Secondly, it could be envisioned that human players wanting to control the on screen musicians could take the place of the sound system. The input of the player would then drive the motor system, so the on screen musician would move in response to the player's input, but would still look like the original musician. For instance, if the player is controlling Lars Ulrich of Metallica, the sound of Ulrich playing would correspond to the performance of the player, but still *look* like how Ulrich would play it. The input from the user would most likely not be identical to that of Ulrich himself, but an advantage of employing neural networks is their ability to generalize and handle noisy situations, which would deal with these kinds of situations. An important aspect of employing the SHEILA architecture would be the cost: using motion capture is an expensive process. However, motion capture is already being used for the creation of such games[5], so the cost issue in this regard

[4] www.idi.ntnu.no/~tidemann/sheila/SHEILAweb.mov

[5] www.usatoday.com/tech/gaming/2008-12-14-metallica-game-qanda_N.htm, retrieved 2011-02-04

would not be prohibitive. To conclude, SHEILA has so far shown promising results regarding its ability to imitate human musical expressiveness, and would be a good approach to enhance games like Rock Band or Guitar Hero.

6. FUTURE WORK

Parts of this paper have been focusing on how this research can be applied in commercially available applications. An open source program called *Frets on Fire*[6] could serve as the starting point for developing SHEILA in a game similar to Rock Band or Guitar Hero.

Although the architecture has shown good results when it comes to imitation of known patterns, the next step will be to examine whether it can generalize and play *new* patterns that have not been part of the training data. This can be tested by recording different patterns from a human drummer, and training the system on selected patterns. The artificial drummer could then be told to play a novel pattern that the system has not been trained on. The output of the system could then be matched against how the teacher drummer would actually play this pattern.

7. REFERENCES

[1] A. Billard and G. Hayes. DRAMA, a connectionist architecture for control and learning in autonomous robots. *Adaptive Behavior*, 7(1):35–63, 1999.

[2] A. Cangelosi and T. Riga. An embodied model for sensorimotor grounding and grounding transfer: Experiments with epigenetic robots. *Cognitive Science*, 30(4):673–689, 2006.

[3] R. L. de Mantaras and J. L. Arcos. AI and music from composition to expressive performance. *AI Mag.*, 23(3):43–57, 2002.

[4] Y. Demiris and B. Khadhouri. Hierarchical attentive multiple models for execution and recognition of actions. *Robotics and Autonomous Systems*, 54:361–369, 2006.

[5] V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12), 1998.

[6] P. Gaussier, S. Moga, J. P. Banquet, and M. Quoy. From perception-action loops to imitation processes: A bottom-up approach of learning by imitation. *Applied Artificial Intelligence*, 1(7):701–727, 1998.

[7] M. Grachten, J. Arcos, and R. de Mantaras. A case based approach to expressivity-aware tempo transformation. *Machine Learning*, 65(2):411–437, 2006.

[8] D. Gusfield. *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, New York, NY, USA, 1997.

[9] H. Imamizu, T. Kuroda, T. Yoshioka, and M. Kawato. Functional magnetic resonance imaging examination of two modular architectures for switching multiple internal models. *Journal of Neuroscience*, 24(5):1173–1181, 2004.

[10] H. Jaeger and H. Haas. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(5667):78–80, 2004.

[11] M. I. Jordan and D. E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, 16:307–354, 1992.

[12] A. Lingnau, B. Gesierich, and A. Caramazza. Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans. *Proceedings of the National Academy of Sciences*, 106(24):9925–9930, 2009.

[13] M. J. Matarić. *Imitation in animals and artifacts*, chapter Sensory-Motor Primitives as a Basis for Learning by Imitation: Linking Perception to Action and Biology to Robotics, pages 392–422. MIT Press, Cambridge, 2002.

[14] A. N. Meltzoff and M. K. Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198:75–78, October 1977.

[15] C. L. Nehaniv and K. Dautenhahn. *Imitation in Animals and Artifacts*, chapter The Correspondence Problem, pages 41–63. MIT Press, Cambridge, 2002.

[16] F. Pachet. Interacting with a musical learning system: The continuator. In *ICMAI '02: Proceedings of the Second International Conference on Music and Artificial Intelligence*, pages 119–132, London, UK, 2002. Springer-Verlag.

[17] J. Piaget. *Play, dreams and imitation in childhood.* W. W. Norton, New York, 1962.

[18] C. Raphael. Orchestra in a box: A system for real-time musical accompaniment. In *IJCAI workshop program APP-5*, pages 5–10, 2003.

[19] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141, 1996.

[20] C. Saunders, D. Hardoon, J. Shawe-Taylor, and G. Widmer. Using string kernels to identify famous performers from their playing style. *Intelligent Data Analysis*, 12(4):425–440, 2008.

[21] S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 3(6):233–242, 1999.

[22] J. Tani, M. Ito, and Y. Sugita. Self-organization of distributedly represented multiple behavior schemata in a mirror system: Reviews of robot experiments using RNNPB. *Neural Networks*, 17:1273–1289, 2004.

[23] A. Tidemann and Y. Demiris. Groovy neural networks. In *18th European Conference on Artificial Intelligence*, volume 178, pages 271–275. IOS press, July 2008.

[24] A. Tidemann and P. Öztürk. Using multiple models to imitate drumming. In *Robotics and Applications*, IASTED Technology Conferences, pages 443–452. ACTA Press, 2010.

[25] A. Tidemann, P. Öztürk, and Y. Demiris. A groovy virtual drumming agent. In *Intelligent Virtual Agents*, volume 5773 of *Lecture Notes in Computer Science*, pages 104–117. Springer Berlin / Heidelberg, 2009.

[26] A. Tobudic and G. Widmer. Learning to play like the great pianists. In L. P. Kaelbling and A. Saffiotti, editors, *IJCAI*, pages 871–876. Professional Book Center, 2005.

[27] D. Tolani and N. I. Badler. Real-time inverse kinematics of the human arm. *Presence*, 5(4):393–401, 1996.

[28] D. M. Wolpert, R. C. Miall, and M. Kawato. Internal models in the cerebellum. *Trends in Cognitive Sciences*, 2(9), 1998.

[6] fretsonfire.sourceforge.net