SoundGrasp: A Gestural Interface for the Performance of Live Music

Thomas Mitchell
University of West England
UK
tom.mitchell@uwe.ac.uk

Imogen Heap
Megaphonic Records Ltd.
UK
info@imogenheap.com

ABSTRACT

This paper documents the first developmental phase of an interface that enables the performance of live music using gestures and body movements. The work included focuses on the first step of this project: the composition and performance of live music using hand gestures captured using a single data glove. The paper provides a background to the field, the aim of the project and a technical description of the work completed so far. This includes the development of a robust posture vocabulary, an artificial neural network-based posture identification process and a state-based system to map identified postures onto a set of performance processes. The paper is closed with qualitative usage observations and a projection of future plans.

Keywords

Music Controller, Gestural Music, Data Glove, Neural Network, Live Music Composition, Looping, Imogen Heap

1. INTRODUCTION

This work began with a discussion between the authors of this paper regarding intuitive methods by which live musical performance processes can be controlled by simple gestures. The intention was to enable a performer to manipulate digital musical processes without having to defer audience engagement to undertake subtle interactions with machinery.

Since the earliest discussions and observations of computer-based electronic music performances, a recurring theme is the breakdown between the actions of the performer and the effect that these actions have on the sound which is produced. That is, the *transparency* of the mapping between the input to an instrument/device and its corresponding output [5]. Unlike traditional acoustic instruments, the control mapping for modern electronic music devices is often opaque and thus difficult for audiences to infer. Bahn *et al.* [2] argue that traditional notions of musicianship should be maintained in electronic music and consequently the connection between gesture and sound should be preserved. However, other authors contend that phlegmatic performances emanating from the glow of a laptop screen mark an inevitable evolution in contemporary, computer-mediated performance [15]; a change in culture to which audiences must adapt and in many instances already have. In either case, the incorporation of clear sound producing or ancillary gestures into a live performance can enhance both audience engagement [12] and communication between performer and listener [17].

In this work, a live sampler, looper and effects processor are controlled by hand gestures selected to convey the processes that they control. In doing so, the performer is extricated from machine interaction which could be perceived as ambiguous by an audience. The following sections will provide relevant background reading with an overview of the system divided into sections following the strategy proposed in [18] for the development of gestural music devices and mappings. These sections will include:

- the definition of a posture vocabulary,
- the methods by which gestures are acquired and identified,
- the mapping strategy for the assignment of these gestures to the control of audio processes.

2. BACKGROUND

There is a large body of research that examines human computer interaction with hand postures and gestures. A subset of this work is concerned with the use of these techniques for musical purposes. These works can be divided into two broad categories [16]: position tracking methods, using optical, magnetic or acoustic technology; and glove-based methods using electromechanical sensors that directly track fine motor activity. At this stage, SoundGrasp employs a single data glove to sense hand posture, consequently this background section is limited to glove-based input.

2.1 Data Gloves and Music

Since the development of the first data glove in the late 1970s, there have been numerous examples of their use within musical contexts. For example, the Cyber Composer system [10] has been developed to enable the composition and performance of live music using a vocabulary of hand gestures, which are mapped to construct chord and melody sequences. MusicGlove [7] enables a database of multimedia files to be searched and played back using simple hand gestures. Recent examples have seen the mapping of glove-captured gestures for the control of electronic percussion[4] and synthesis [18].

The work presented in this paper focuses on the acquisition of hand gestures and their mapping onto musical processes within a live performance environment. The system enables the realtime sampling and manipulation of sound using gestures that lend themselves to the processes that they control.

NIME'11, 30 May–1 June 2011, Oslo, Norway.

3. LIVE SAMPLING - SOUND GRASPING

Despite the wide musical application of glove-based gestural controllers, live sampling and looping is an area which has been relatively unexplored; although examples are beginning to emerge. One such system is the Vocal Augmentation and Manipulation Prosthesis (VAMP) [11]. Equipped with this device, a singer can 'freeze' a single note when the finger and thumb are pressed together, activating a pressure sensor located on the glove. This 'pinch' gesture captures a short frequency domain representation of the incoming signal which is resynthesised continuously until the pinch is released. Further harmony and amplitude modulation is facilitated through the use of flexion and acceleration sensors also attached to the glove. This mapping ascribes a widely understood gesture for the physical act of 'holding' to a process that 'holds' the incoming audio. Fels *et al.* [5] describe this appropriation of recognised gestures as *metaphor*, which can be used to increase the transparency of control mappings for both audiences and performers.

The second author of this paper regularly performs music incorporating the live sampling of vocals and acoustic instruments. The proposed system has been designed around the requirements of this situation:

1. The musical processes should be controlled without having to defer performativity to engage in machine interaction.

2. There should be a transparent mapping between the input to the gestural controller and the outgoing musical events.

3. Instrumental virtuosity should be compromised as little as possible.

The wearable components of the presented work are shown in Figure 1, comprising a fingerless data glove with a wrist-mounted microphone. This arrangement has minimal constraints on dexterity and unites the gestural controller with the sound capture device. This enables proximal sound sources to be sampled using a grasping metaphor: recording commences when the hand is opened and concludes when the hand is closed. Thus the sound appears to be 'caught' by hand.

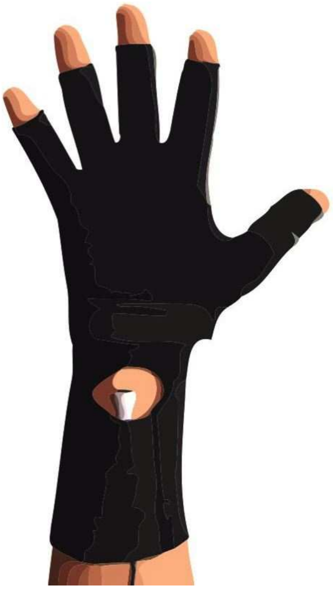

Figure 1: SoundGrasp glove with wrist mic

4. SYSTEM OVERVIEW

Gestural music devices are widely represented as a three part system: the gestural controller, the audio processing unit and the mapping that exists between the two [18]. For this work, the mapping and audio processing are both incorporated into a cross-platform C++ application which was developed using the library Juce [14].

Gestural Controller

Figure 1 shows the gestural controller which includes a single 5DT 14 Ultra glove [1] measuring finger flexion and abduction with 14 fibre optic bend sensors. Also connected to the glove is a lavaliere microphone to enable the recording of live input. Both the glove and microphone connect wirelessly to a computer managing the gestural mapping and audio processing.

Gestural Mapping

Raw serial data transmitted by the glove is decoded and routed to the inputs of an artificial neural network to identify discrete and static hand postures. Identified postures are subsequently used to control the state of the audio processing unit.

Audio Processing Unit

The audio processing unit is a software application which currently enables the recording, overdubbing, looping and modification of audio data.

5. POSTURE VOCABULARY

Previous efforts have been made to formalise universal sets of gestures, see for example Henze [8] for gestures associated with media playback. Many of these studies indicate a lack of consensus amongst participants. Consequently, the vocabulary of hand postures adopted for this work has been chosen pragmatically to be identifiably distinct and to enable the use of metaphor in the control mapping. The posture set is shown in Figure 2.

6. GESTURAL MAPPING

The mapping layer of the SoundGrasp system, mediating between the glove and the audio processing unit, consists of three parts: data processing, posture identification and audio control (Figure 3). Data processing involves the unpacking and normalisation of the serial data from the glove into floating-point sensor values in the range 0.0 to 1.0. The details of the posture identification and audio control process are provided below.

6.1 Posture Identification

Posture identification serves to process the calibrated sensor data to identify when the glove has formed a shape approximating a registered posture. This process forms a pattern recognition problem for which artificial neural networks have been demonstrated to be particularly well suited [6].

Artificial Neural Networks

Artificial neural networks provide a biologically inspired machine learning technique which is loosely modelled on the architecture of the brain. The type of neural network employed here is a multilayer perceptron, which is a fully connected feedforward neural network trained with the backpropogation supervised learning technique. This network architecture has been widely used for the non-linear control of audio and visual systems [13]. This section will only

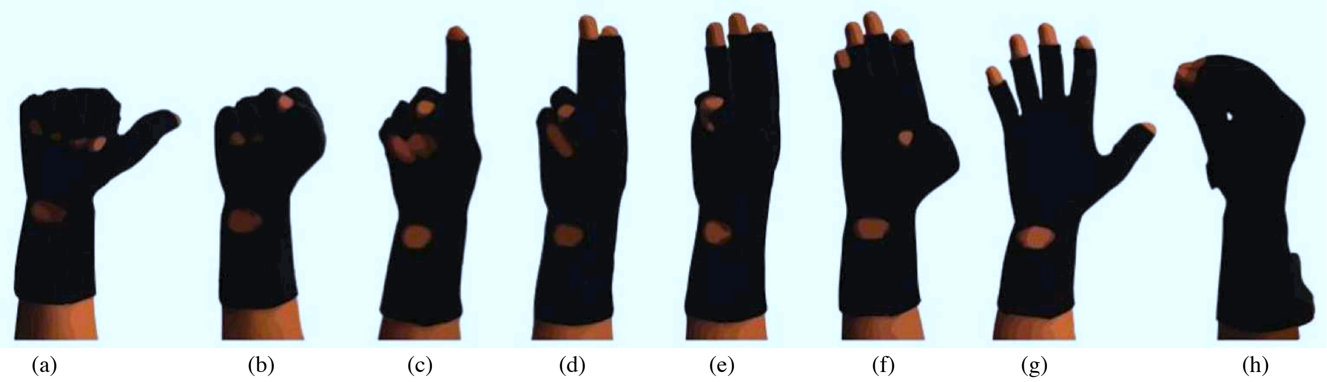

Figure 2: Current posture vocabulary for SoundGrasp

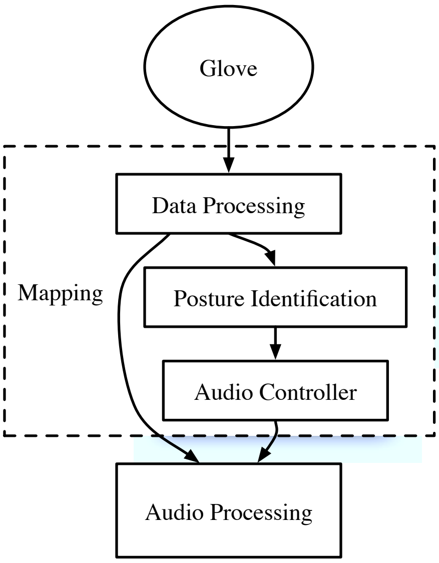

Figure 3: SoundGrasp system architecture

provide a brief summary of the relevant neural network architecture, for fuller treatment and implementation details the reader is referred to [3].

The multilayer perceptron is constructed from layers of interconnected computational units called neurons. Each neuron has one or more inputs and a single output, both of which can be connected either externally or to other neurons. Frequently, the network is configured with three layers: an input layer, a hidden layer and an output layer. The intention is to configure the network such that a known pattern of input values (finger positions) results in a target pattern of output values (identified hand positions). This is achieved with a supervised learning process using sets of training data. A single training set includes a pattern of input and target output values. Subsequent to a successful training procedure, the network should produce a mapping represented by the training set. That is, when the network inputs are set to match an input pattern from the training set, the output of the network should closely match the corresponding training set output pattern.

For the identification of hand postures in this work, the neural network was configured with 14 inputs, matching the quantity of normalised sensor values from the glove. The number of outputs was set to match the number of gestures in the gesture set, currently eight. Subsequent to training the gesture identification was found to be robust with 12 hidden neurons following recommendations set out in [3]. The configuration of the network with one output per posture enabled confidence testing to be performed, preventing the unintentional triggering of postures, while permitting subtle idiosyncrasies that occur when assuming the same hand position.

6.2 Audio Control

Recognised postures are mapped through a further layer, facilitating the selection of audio processes to be controlled using only one glove. This audio control layer manages a simple state based system which enables the performer to switch between modes with sequences of hand postures that form simple gestures [9]. This results in the distinction between two types of gesture:

1. Audio control gestures

2. State/mode control gestures

State control gestures switch the system between different modes which enable the performer to activate different types of audio control processes. This forms a one-to-many mapping between gestures and audio control where a single gesture can be mapped to multiple audio processes through different modes. In establishing the control mapping, audio control gestures, which directly affect the produced sound, use metaphor to increase transparency. In contrast, state control gestures, producing no audible effect, were chosen for performer usability.

Audio Control Processes

The audio control processes were divided into modes which are summarised in Table 1. The principle gesture for audio control is grasping, represented by transitions between postures (g) and (h) in Figure 2. Posture (g) is an open hand, while (h) forms a grasping posture with the tips of each finger in contact with the thumb. Recording is achieved as described earlier and the audio track is cleared with posture (c); raised to the lips, this forms a familiar gesture for silence. In play mode the grasping gesture is reused, playback is paused with (h) and resumed with (g). Reverse playback is initiated with (d) and forwards playback resumed with (g). The filter and effects modes access the sensor data directly with continuous control of the corresponding parameter with the average flexion reading for all four fingers. Lock mode deactivates the glove to enable hand movements without the risk of erroneous audio control, while playing an instrument, for example.

Mode	Audio Processes	Posture
Record	Record/overdub, clear	(h)
Play	Play, stop, reverse	(a)
Filter	Low-pass cutoff	(c)
Reverb	Reveb time	(d)
Delay	Delay time	(e)
Lock	None	(b)

Table 1: Audio Control Processes and Modes

State/Mode Switching

Mode switching is performed with a gesture consisting of two postures in sequence: the first posture (f) indicating the start of a mode switch and the second posture indicating which mode to select. Posture (f) was chosen to initiate the mode switch as full flexion of the lower and upper knuckles of the thumb rarely occurs incidentally. Subsequent mode switch postures are provided in the third column of Table 1.

7. RESULTS FUTURE WORK

Informal testing with a small number of subjects indicated that, after the neural network was trained for individual users, the system was intuitive and easy to lean. Response to hand postures was prompt and stable enabling users to record accurately timed loops consistently. Users were observed to develop their own metaphors adding ancillary gestures over and above those required. For example, several subjects issued audio control gestures with both hands, particularly in the control of playback: amplifying the 'releasing' and 'holding' gestures. The reverse mode, activated with a two fingered point was often accompanied with an additional swipe towards the body, and released with a converse swipe, as if playing an invisible turntable. Some problems were encountered when users wanted to switch modes from postures other than the open hand (g). For example, users wishing to switch modes with playback reversed, record mode disabled or playback halted frequently formed hybrid postures combining the mode switching posture (f) with postures (d) or (h). These issues were solved by adding these hybrid postures to the neural network training set, or by providing the user with further guidance instructions. Alternative solutions will be explored with different neural network architectures to enable thumb postures to be identified in isolation. The authors have many plans for future extensions to this work. Immediate development will incorporate an additional glove and the use of position, orientation and/or acceleration sensors. A second glove hugely increases the degrees of freedom and capacity for further audio and state control switching affording a much more comprehensive range of musical controls. Furthermore, a means of feedback will also be developed as there is currently no mechanism communicating the internal state of the system to the performer. Should a mode switch occur in error, the performer is unaware until the wrong audio processes are subsequently activated. Visual feedback from LEDs attached to the glove will be developed for this purpose.

8. ACKNOWLEGEMENTS

Thanks to Martin Robinson for proof reading this paper and developing the UGen++ library from which the neural network code for this work was ported. For providing the equipment and supporting the project, a big thank you to Professor Tony Pipe and the researchers in the Bristol Robotics Laboratory. Also thank you to the anonymous reviewers of this paper for their suggestions.

9. REFERENCES

[1] Fifth dimension technologies, www.5dt.com, 2011.

[2] C. Bahn, T. Hahn, and D. Trueman. Physicality and feedback: a focus on the body in the performance of electronic music. In *Proceedings of the International Computer Music Conference*, pages 44–51, 2001.

[3] A. Blum. *Neural networks in C++: an object-oriented framework for building connectionist systems.* John Wiley & Sons, Inc., New York, NY, USA, 1992.

[4] S. Chantasuban and S. Thiemjarus. Ubiband: A framework for music composition with bsns. In *Sixth International Workshop on Wearable and Implantable Body Sensor Networks*, 2009.

[5] S. Fels, A. Gadd, and A. Mulder. Mapping transparency through metaphor: towards more expressive musical instruments. *Organised Sound*, 7(2):109–126, 2002.

[6] S. Fels and G. Hinton. Glove-talk: a neural network interface between a data-glove and a speech synthesizer. *IEEE Transactions on Neural Networks*, 4(1):2 – 8, 1993.

[7] K. Hayafuchi and K. Suzuki. Musicglove: A wearable musical controller for massive media library. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2008.

[8] N. Henze, A. Löcken, S. Boll, T. Hesselmann, and M. Pielot. Free-hand gestures for music playback: deriving gestures with a user-centred process. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia*, 2010.

[9] P. Hong, T. S. Huang, and M. Turk. Gesture modeling and recognition using finite state machines. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

[10] H. Ip, K. Law, and B. Kwong. Cyber composer: Hand gesture-driven intelligent music composition and generation. In *Proceedings of the 11th International Multimedia Modelling Conference*, 2005.

[11] E. Jessop. The vocal augmentation and manipulation prosthesis (vamp): A conducting-based gestural controller for vocal performance. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, 2009.

[12] G. Paine. Interfacing for dynamic morphology in computer music performance. In *Proceedings of the International Conference on Music Communication Science*, Sydney, 2007.

[13] M. Robinson. Neural networks for audio-visual control. In *Proceedings of the CREAM Symposium, Cybersonica*, 2003.

[14] J. Storer. Juce, www.rawmaterialsoftware.com, 2011.

[15] C. Stuart. The object of performance: Aural performativity in contemporary laptop music. *Contemporary Music Review*, 22(4):59–65, 2003.

[16] D. Sturman and D. Zeltzer. A survey of glove-based input. *Computer Graphics and Applications, IEEE*, 14(1):30 –39, jan 1994.

[17] W. F. Thompson, P. Graham, and F. A. Russo. Seeing music performance: Visual influences on perception and experience. *Semiotica*, 156(1/4):203–227, 2005.

[18] M. Wanderley and P. Depalle. Gestural control of sound synthesis. *Proceedings of the IEEE*, 92(4):632 – 644, April 2004.