ELECTROACOUSTIC PERFORMANCE INTERFACES THAT LEARN FROM THEIR USERS

Andrés Melo**, John Drever*† *and Geraint Wiggins**
Departments of Computing* and Music†, Goldsmiths' College, University of London, UK

ABSTRACT

We present preliminary outcomes of a feasibility study of a novel application of machine learning technology to the sound diffusion work of an electroacoustic composer/performer. We propose a simple but effective visualisation method for diffusion data, and present evidence that simple learning technology can learn the necessary structure to facilitate diffusion performance.

1. INTRODUCTION

Diffusion is the performance practice of real-time spatial deployment of fixed acousmatic music compositions of two source audio channels around a multi-channel sound system. It is distinct from *multi-channel composition*, which works with more than two channels, using distinct, pre-determined speaker outputs that are not manipulated in real time. Concern for real-time spatial movement in electroacoustic music dates back to Pierre Schaeffer and Pierre Henry's inaugural concert of *musique concrète* at the *Ecole Normale de Musique* on 18th March, 1950 [3]. Since then, a number of unique, purpose-built diffusion systems have been built, such as the *Groupe de Recherches Musicales' Acousmonium* in 1974 and University of Birmingham's *BEAST* (Birmingham ElectroAcoustic Sound Theatre) in 1982. Ranging from 4 to the extreme 48 outputs of the Acousmonium, diffusion has become common performance practice in the UK, France and Canada and is now increasingly popular in the USA.

In performance, the diffuser (who is not always the composer) provides a live interpretation of an otherwise fixed work. A work can thus be greatly enhanced. Often a diffuser will create a diffusion score to help prepare for approaching musical events. Ideally, there will be significant rehearsal the performance with a given diffusion system in a given performance space (since systems and performance spaces vary greatly), but, in reality, rehearsal time is often limited. In consequence, in performance, there is inevitably a dimension of improvisation, contingent upon the requirements of the work and its interpretation. Composers of music with diffusion intended as its definitive presentation consider the eventual diffusion in the making of the piece. So, in mere stereo playback of such works, an important dimension is missing.

Our project is investigating whether machine learning technology—in this case, artificial neural networks (ANNs) [8]—can be deployed to assist the diffuser. The simplest case would be to reduce the number of parameters required, allowing the performer to use just one handful of sliders instead of two. In order to do this, we had first to establish that there were patterns in our diffuser-subject's performance, and second to select and train appropriate ANN technology to learn and reproduce them.

2. RELATED WORK

Machine learning [9] has been used frequently in music-related work, in many ways. For example: Widmer and Tobudic [11] studied the extraction of symbolic performance expression rules from performance data; Ponsford et al. [10] modelled harmonic movement using Markovian statistical models; Arcos and López de Mántaras [1] used case-based reasoning to control expressive performance of synthesised saxophone solos [1]; and many music-related applications of ANNs have been reported over a long period [2, 6, 7, for example].

However, no work on learning systems in diffusion seems to have been published. Perhaps this is either because researchers are unwilling to dilute the remaining live element of electroacoustic performance, or because diffusion is less clearly understood in the AI/music world than activities which lead obviously to the production of sound. Neither of these is a reason for not studying the topic: even if we keep diffusion live, appropriate enquiry can yield new insight into performance practice, and need not necessarily lead to further automation. Performance practice is a major focus of the current work.

3. RETHINKING DIFFUSION AS PERFORMANCE PRACTICE

Diffusion usually constitutes the live element of tape- or CD-based electroacoustic performance. Multiple loudspeakers are arranged around the audience, usually, horizontally and symmetrically about the audience's axis of vision. Speakers are then associated with output channels of a multichannel sound mixer, and the diffuser uses this device to do their work: the source signal is usually in two channels only. The diffuser's job is to manipulate the recorded stereo signal into the multichannel performance by allocating the two source channels to different combinations of output channels, fading between them (to produce the illusion of movement) and so on. Sound projection rigs have as many as thirty channels, and since a diffuser

[1] We include CBR-based work under the heading of "machine learning" because it relies on the analogical application of prior knowledge, and can build ("learn") a library of cases as it goes along.

will want very detailed individual control over each, independently, they will need to manipulate at least as many sliders as there are channels.

The problem with this is that manual control of a high-dimensional controller is difficult, and in this case rendered more so because the controller being used (a mixing desk) was not originally designed for this use, and is not optimised for it. One solution would be to explore alternative controllers, but this would mean that the performer has to learn to use them, and that the intuition behind live performance usage (having the diffusion "under one's fingers", just like a pianist) has to be relearned afresh. For this reason, we explore alternatives which maintain that well-established intuition, mapping movements of sliders on to perceived movements in sound in more complex ways than is possible with an ordinary mixing desk, but leaving the actual user-interface unaltered.

The bottleneck for the conventional diffuser is, simply, the number of volume sliders that they can manage with just two hands. This is not as simple as being restricted to ten movements at once: the physiology of the human hand and the physical layout of the mixer constrain the space of possibilities much more than the number of digits alone.

The existence of these strong constraints means that that, notwithstanding the high dimensionality of the space being explored by the diffuser, they cannot actually explore much of it. Therefore it should be possible to reduce that dimensionality so that the diffuser can achieve the same effects (for a given piece) with a smaller number of sliders, possibly even with just one hand. This would have the advantage of leaving the diffuser with one hand free to manipulate other parameters of the performance.

Our thinking, then, was to build some intelligence into the diffusion system, to allow diffusers to use a familiar interface, but also to reduce the complexity of direct control required while maintaining expressivity. Since the actual usage of the space of diffusion possibilities is tightly constrained (and so only a small section of it is actually explored), we hypothesised that it should be possible to construct a function from some relatively small number of control inputs to a larger number of outputs, which would model a diffuser's interpretation of a given piece on a given diffusion system. The justification for this is that since only a subset of the points in the higher-dimensional space can be explored by the diffuser, those points should in principle be mappable onto the points in a smaller space. If that space has lower dimensionality (and so can be directly explored with fewer sliders), then the important question is whether trajectories in the higher-dimensional space can be modelled accurately with fewer dimensions—otherwise, for example, sounds which move smoothly in the unintelligent system might jump unacceptably around in the intelligent one.

To explore our hypothesis, we chose a simple, relatively well-understood learning system, a feed-forward perceptron network trained by back-propagation [8]. This kind of network learns a continuous function mapping its inputs to its outputs, and is capable of interpolation (generalisation) between the datapoints on which it is trained. This is crucial for our work, for two reasons: first, we need to be able to maintain smooth trajectories in the space of control parameters of our system; and, second, we wish to use the interpolated points to suggest new performance gestures to the performer, in future work.

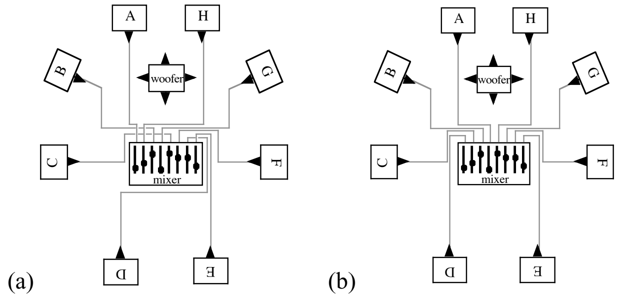

Figure 1. (a) The control setup for the output training data gathering phase. (b) The control setup for the input training data gathering phase.

4. A PRELIMINARY EMPIRICAL STUDY

We worked with a simple performance setup, so as not to obscure the data in our study. It had nine channels, one multidirectional woofer and eight unidirectional mid-range/tweeters. The diffuser sat centrally facing speakers A and H. The setup is shown in Figure 1a, with the eight physical mid/high-frequency channels labelled A to H.

We used a MIDI slider box as our mixing console, playing back the sound from Logic Audio, running under Mac OS X, *via* a MoTU 828 Mk2 multi-channel sound output module. Logic Audio software control units were used to apply the MIDI volume control signals generated by the faders controlling the output levels of the appropriate channels. Thus, we could record the diffuser's performance as a MIDI file, synchronised with the stereo source recording and subsequently interpose our software between the diffuser and the level controls. The resulting MIDI files contained our data, conveniently labelled by MIDI channel and time-stamped. We did not use the woofer control signal in learning, because, since bass frequencies are less directional than higher ones, it did not contribute significantly to the diffusion.

In this data-gathering phase, the controller was set up as in Figure 1a, which may be surprising: a more obvious layout is where the sliders correspond spacially with the physical speakers. In order to achieve certain common musical gestures (for example, smooth transitions of sound from front to back on both sides simultaneously), human hand physiology requires that certain sliders be placed close together: otherwise the gesture cannot be adequately controlled. Therefore, our diffuser used this setup when composing and performing his music.

In a feasibility study such as this, the first step is to examine one's data to ensure that there really is structure for a learning system to learn—otherwise, results could be misleading. We invited our diffuser to record several takes of each of two of his pieces, [4, 5]. This multiplicity of data would enable us to verify, first, that there was structure in each take, and, second, that there was correlation between the different performances, confirming that it was reasonable to generalise across these data. Finally, we used two pieces so that we could look at both inter- and intra-opus learned generalisations. To train a network, it is necessary to have lots of data, so that generalisations can be statistically reasonable. Aside from the multiple takes mentioned above, the pieces used were long (in excess of 20 minutes) and so generated large amounts of data from each performance, given that variation in the diffusion was continuous in both mathematical and vernacular senses.

Having captured the data, examining it was a non-trivial problem, because it was expressed as impenetrably large numbers of integers stored in text files. We devised a simple but effective visualisation method: on a time line, we laid out the eight channels, in two groups of four, for left and right, using a heat colour scale. Between the two groups of four is a monochrome scale, indicating overall energy in the sound, and the channels are graphed in terms of their relative contribution to this overall energy. Examples are shown (in monochrome) in Figure 2, with the timeline running down the page. As an example of how to read the graphs: the very first data of all performances shown has very low energy (the central line is very dark), and that energy is all in channels A and H, shown by the maxima in the two left-most columns of each group of four and minima in all other channels.

Study of the three visualisations shows that: there are areas of consistency and other areas where there is less consistency between performances; there is a strong tendency for the diffusion to be symmetrical about the visual axis of the diffuser (this is not surprising, as the original signal is in stereo and so diffusion in this spatial dimension is pre-definined); and (on the rather smaller scale) there are diffusion gestures which recur within the individual performances. So we concluded that it was reasonable to proceed with training a network from this data, to reduce the dimensionality of the data as proposed above.

To train a network to do our dimensionality reduction, we needed not just the control data for the physical channels, but also some desired "virtual" control data for performance. Methodology for recording this data was a significant issue, because data generated in a situation which was unnatural for the performer would probably be artificial, and our approach would be invalidated. As a first step to overcoming this, we used a very similar setup to re-record our diffuser's control signals, but with pre-recorded diffusion, and the diffuser "miming" with a configuration of sliders different from his usual: the same number of sliders, but mapped to different physical channels, as in Figure 1b. We gave our diffuser adequate opportunity to get used to the new setup, to encourage him to think about how to achieve the gestures he wanted with the new setup, but to liberate him from implicit habits based on the previous one. On the basis of the symmetry in our initial data, which suggested significant amounts of redundancy between the left and right hand gestures (see Figure 2), we took the data from the diffuser's left hand only, and used it, in time series (see below), for our input data.

Because diffusion is about changes in spatial projection, we needed a temporal context in our system. For this, we used time-series training: each set of four data (one per slider at each time step) was presented to the network in the training phase along with the data from 40ms previously, 80ms previously, 160ms previously and so on in powers of 2, to 40.96s. Thus, each set of data is related by the network with its context in the piece.

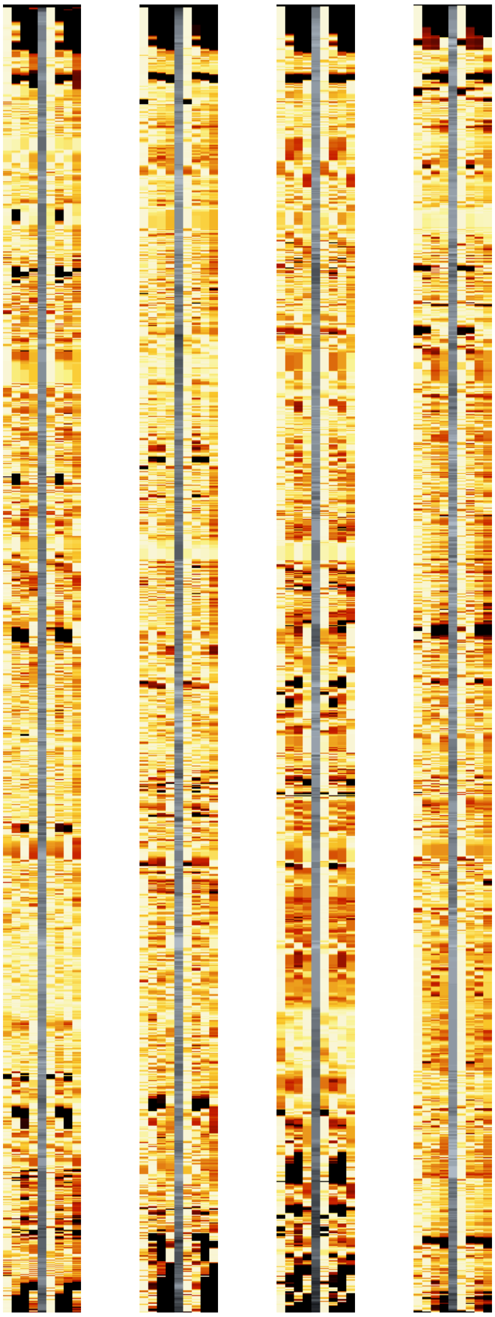

Figure 2. Visualisations of the diffusion data from four different performances of *Hippocampus* [5]. Time runs down the page; the visualisations are synchronised in time, to within 2 or 3 seconds; overall time is 23m 45s. Physical channels are in the order ABCDHGFE.

We chose a well-understood, simple network architecture, to facilitate comprehension of the outcomes. We used the public-domain SNNS system [12] to build feed-forward perceptron networks, trained by back-propagation, with one hidden layer, an input layer and an output layer; the input layer is fully connected to the hidden layer, and the hidden layer to the output layer, but the input and output layers are not directly connected. The input layer used the geometric time-series representation, above, giving 12 sets of 4 data in 48 nodes; the input layer training values were taken from the four left-hand virtual channel data, above. The output layer's eight nodes were assigned one each of the physical channel levels A-H; their training data was the normalised physical channel volume data as described above. The actual output values must then be reconstructed from the overall sound volume described by the diffuser with the four input sliders. The size of the hidden layer was a parameter of the study, and is discussed below. All data were synchronised in time, and the different takes of each piece were used together to train two separate piece-specific networks. Thus, we expected intra-opus generalisations to be reinforced; inter-opus generalisations are left for future study.

We ran several experiments with different networks, having between 6 and 30 hidden nodes, and with no hidden layer. The data were divided into training and test sets as usual: we divided the pieces into 150 second sections, using the first 30 seconds of each for testing and the rest for training. This sampling approach ensures that a good spread of each piece's gestural content was both learned and tested: it guarantees that the test set is fundamentally different from the training set.

Testing and pruning of the networks suggested an optimal hidden layer size of about 16. In this case, the mean square error obtained on testing was 0.214 on average, spread across all eight channels. It is not yet clear what proportion of error was due to inconsistencies in the data and what was due to the limitations of the system. Also, more study is required of the effect of these inaccuracies on the diffuser in performance.

5. IMPLICATIONS: PERFORMANCE PRACTICE

This work has implications for diffusion practice. Some direct effects of the approach are beneficial, even before the system has been deployed in performance. Our diffuser reported that use of the system seems to lead to a deeper understanding of the music and the diffusion, and of what can be done in future performance and composition. This is because the diffuser is forced—and enabled—to think in more detail and with more preparation about what they are doing, as follows.

Because the network must be trained, there is no alternative to spending time on rehearsal, and listening to the music many times. Producing the input training data requires an unusual approach: diffusing a work while hearing a previous performance; this is a good training technique: it develops listening skills. Also because of the need for input training data, one must experiment with different fader configurations, and then reflect in new ways on how the configuration affects the performance.

Our diffuser reported that the visualisation graphs are valuable in the analysis of diffusion. We believe that no similar support tool exists for diffusers. They allow the diffuser to compare performances, to reflect on strategies used, and to match them directly with *post hoc* analysis. The visualisation tools imbue what is often an improvisatory process with informed reflective practice.

6. CONCLUSION & FUTURE WORK

We have presented a novel approach to the automatic assistance of electroacoustic music diffusion. We have carried out a preliminary feasibility study, whose positive results we have presented here. We believe that this work could potentially add a new dimension to the often sterile performance of electroacoustic music by allowing the diffuser to control more parameters of the performance and thus to add more live expression to the medium.

The interim results presented here suggest that the simple learning technology used will be able to bridge the gap between the diffuser's chosen physical interface and the level of control required for diffusion. The next step will be to record performances using the neural network to map between one-handed control and the full eight channels. This data will then be evaluated in terms of the ability of the diffuser to achieve the gestures he requires.

Acknowledgments

John Drever was supported by an ACE/AHRB Arts and Science Fellowship, 2003-2004.

References

[1] J. L. Arcos and R. López de Mántaras. An interactive case-based reasoning approach for generating expressive music. *Applied Intelligence*, 14(1):115–129, 2001.

[2] J. Barucha. Music cognition and perceptual facilitation: a connectionist framework. *Music Perception*, 5:1–30, 1987.

[3] J. Chadabe. *Electric Sound: the Past and Promise of Electronic Music*. Prentice Hall, New Jersey, 1997.

[4] J. Drever. Hippocampus, 1998. Electronic music for tape and diffusion.

[5] J. Drever. Peregrinations, 1999. Electronic music for tape and diffusion.

[6] D. Eck and J. Schmidhuber. Finding temporal structure in music: Blues improvisation with lstm recurrent networks. In H. Bourlard, editor, *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, pages 747–756, New York, 2002. IEEE.

[7] D. Gang, D. Lehmann, and N. Wagner. Harmonizing melodies in real time: the connectionist approach. In P. R. Cook, editor, *Proceedings of ICMC'97*, 1997.

[8] K. Mehotra, C. K. Mohan, and S. Ranka. *Elements of Artificial Neural Networks*. Bradford Books, 1996. ISBN 0-262-13328-8.

[9] R. S. Michalski et al., editors. *Machine Learning: An Artificial Intelligence Approach*. Springer-Verlag, Berlin, 1983.

[10] D. Ponsford, G. A. Wiggins, and C. S. Mellish. Statistical learning of harmonic movement. *Journal of New Music Research*, 28(2):150–177, 1999.

[11] G. Widmer and A. Tobudic. Playing Mozart by analogy: Learning multi-level timing and dynamics strategies. *Journal of New Music Research*, 32(3):259–268, 2003.

[12] A. Zell et al. Stuttgart neural network simulator, 2002. Available for free download from `http://www-ra.informatik.uni-tuebingen.de/SNNS/`.