Diffy

JORDIE SHIER, Centre for Digital Music, Queen Mary University of London, UK
XIAOWAN YI, Centre for Digital Music, Queen Mary University of London, UK

Fig. 1. Photo of live setup with drumset, computer, Ableton controller, and a modular synthesizer.

Additional Key Words and Phrases: Hybrid electronic set, Percussion timbre re-mapping, Music human-machine-human interaction

1 Program Notes

Diffy is a duo music project comprising a drummer and a sound designer, connected by a set of machine learning-based sound design agents. In this project, we explore and juxtapose a set of three machine learning-based techniques for manipulating the timbral qualities of percussion instruments in real-time with low-latency. These techniques include a neural audio synthesizer trained on non-percussive material, a timbre remapping 808 drum synthesizer, and a modular synthesizer controlled by a neural network. Each sound design agent operates on different modes of timbral understanding—reacting to the drum performance based on this understanding, and suggesting sonic transformations. Sonic negotiations between the human sound designer and the sound-design agent are relayed back to the drummer, creating a feedback loop that shapes a structured improvisation.

Authors' Contact Information: Jordie Shier, Centre for Digital Music, Queen Mary University of London, London, UK, j.m.shier@qmul.ac.uk; Xiaowan Yi, Centre for Digital Music, Queen Mary University of London, London, UK.

NIME '25, June 24–27, 2025, Canberra, Australia

2 Project Description

Audio-driven synthesis involves the translation of audio from instruments onto controls for a synthesizer, expanding the timbral capabilities of the input instrument [9]. Several methods for achieving audio-driven synthesis have been explored in related technical work, each focusing on a different synthesis modality and operationalization of machine learning. For instance, Fasciani et al. [5] used timbral features and unsupervised learning to create vocal-to-synthesizer mappings with VST synthesizers. Recently, differentiable digital signal processing (DDSP) [4], which incorporates digital signal processing techniques, such as sinusoidal modelling with neural networks, has been proposed and enables training directly on instrumental audio examples. DDSP has enabled novel opportunities for musical expression including real-time timbre transfer [2]—the process of transforming the timbre of one instrument into another while maintaining key performance attributes such as pitch and rhythm. In another line of research, neural audio synthesis leverages the representational power of neural networks to *learn* to generate audio based on training data. The RAVE model has been a particularly popular model, owing to it's real-time capabilities and ability to perform timbre transfer on difference audio material [1]. Recent NIME-related work has explored the musical affordances through novel interfaces connecting to the latent parameter space of RAVE [6].

The first author involved in this musical proposal has explored a variety of machine learning-based methods for audio-driven synthesis, specifically focusing on real-time control from percussion instruments. This musical project—Diffy—was established during a research-oriented user-study that was being conducted by the first author with the second author performing drums. They were evaluating an audio plugin timbre remapping system developed by the first author [8]. A screenshot of this plugin (which is used in the proposed performance) is shown in Figure 2. This plugin accepts audio features and maps them onto parameters for an 808 drum synthesizer. Midway through the evaluation session, the first author started interacting with the controls of the synthesizer that was also being controlled, via audio, by the drumming of the second author. This changed the direction of the session from a science-based interaction into a musical collaboration, mediated by the neural network controlling the synthesizers.

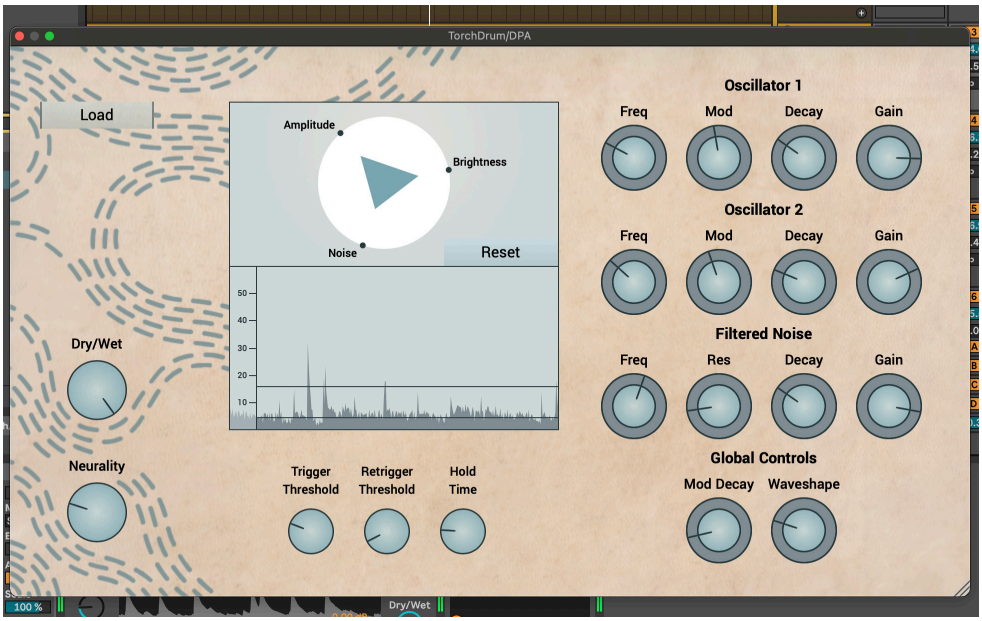

Fig. 2. DDSP-based timbre remapping plugin used in this music proposal

Since then the duo has built a musical practice revolving around the exploration of audio-driven methods for transferring the percussive performance onto synthesizers in various ways. We're particularly interested in the interplay between different modes of representing audio computationally, and how that impacts the process of real-time control of synthesizers. For instance, we utilize a low-latency version of a RAVE model (BRAVE) [3] trained on saxophone sounds which is then is asked to "listen" to a drummer. Interventions upon this neural "listening" are applied in real-time by the human sound designer to actively draw out pitched material that the RAVE "saxonphonist" can latch onto. Conversely, we also explore methods that have been explicitly trained on the drumset in practice, which provide different possibilities to mapping, such as mapping to a Eurorack modular synthesizer. This method for timbre remapping, which is presented alongside this performance at NIME 2025 [7], utilizes a genetic algorithm to generate a dataset of synthesizer parameter variations that

result in desired timbre trajectories[1]. This dataset becomes training material for neural network mapping from percussive audio features onto modular synthesizer parameters.

During the proposed NIME musical performance we will perform with three different audio-driven synthesizer technologies that have either been trained or developed by the musicians involved. A visual overview of the three audio-driven synthesizers, the inputs they receive from audio, and the control inputs from an APC40 controller is shown in Figure 3.

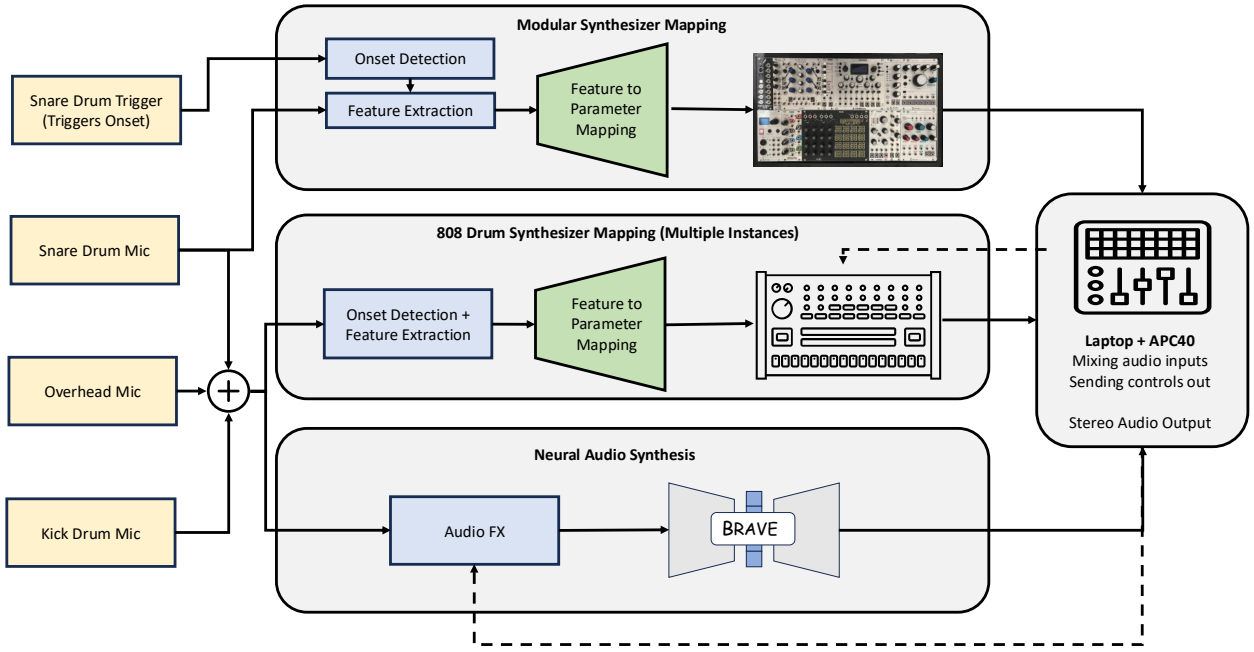

Fig. 3. Block diagram showing flow of audio from microphones to the three audio-driven synthesizers. The top network maps from the snare drum to a modular synthesizer. A drum trigger is sent to an onset detection network, which triggers feature extraction. Features are sent to a neural network parameter mapping system which generates modular synthesizer parameters. The middle network is a DDSP-based 808 drum mapping network. This network receives a combination of snare, overhead, and kick microphone input. These feed input to a combined onset detection and feature extraction block, which maps to synthesizer parameter modulations. Synthesizer controls are also fed in from the APC40 controller. Multiple 808 drum synthesizers run in parallel and their levels adjusted in real-time. Finally, audio from the snare, overhead, and kick mixture are passed into audio effect chain and then into a BRAVE model trained on saxophone. All audio is sent to an audio interface and levels mixed in Ableton Live with an APC40 controller. The output is stereo audio.

3 Artistic Reflections

We've both spent considerable time performing with drums and synthesizers in an electronic music context, but this setup led to a much tighter integration between electronic sounds and live percussion. All synthesized sound sources rely on input from the drums — there is no sound without them. This is most obvious rhythmically, as the synthesized sounds follow the drums and are layered on top of them. The extra layer of timbral control from the drums added a new dimension of modulation and led to more dynamic synthesizer sounds, which we found aesthetically pleasing, although this required careful tuning during setup. The most immediate impact of the rhythmic and timbral mapping is that it freed Jordie to focus on macro aspects of the improvisation, such as mixing different sound elements and adjusting the overall timbre of the synthesizers. We created different structural elements using the different sonic qualities of each synthesizer. Because these were so tied to the drums, listening to each other — such as Xiaowan moving to cymbals, and Jordie responding — was important for generating movement. A limitation of this tight integration is that the synthesizer is rhythmically bound to the drums. We began to break out of this by playing with the decay of the synthesized sounds. For example, adding long decays and filtering the drums before input to the RAVE model allowed the synthesized sources to "break

[1]Here, timbre trajectories refer to relative differences in timbral audio features computed between two sounds. We use these relative differences to *transpose* timbre sequences from the acoustic drum onto the synthesizer in a similar manner to how one might transpose a melodic melody between different keys

away" from the drums — if only slightly. Exploring how to effectively diverge from the drums rhythmically and temporally represents an interesting opportunity we'd like to pursue in future iterations of this project.

4 Technical Notes

Our duo live set has a duration of between 10 - 13 minutes. The second author plays the acoustic drum kit and interacts with the audio output from machine learning-based sound design agents. The first author plays and interacts using the computer, an Ableton APC40 controller, and a modular synthesizer. We will send stereo audio output from an audio interface from the laptop.

We will bring the following:

- Laptop + Ableton APC40 Controller
- Audio interface
- Modular Synthezier
- Drum triggers
- 1/4" TS Cables for Drum Triggers
- 2x 1/4" TS Cables output from audio interface

We will require the following from the venue:

- Drumset
- Drum microphones (kick, snare, overhead) for real-time processing
- Drum microphones for sound reinforcement
- XLRs for microphones
- Stereo DI (output from audio interface)
- Table for electronics
- 4x plugins

5 Media Links

Video: https://youtu.be/mwq-rSJN048?feature=shared

6 Ethical Standards

This musical proposal does not involve experiments with human participants beyond the two involved in the project, hence no institutional ethics board review was required. Machine learning methods are employed in this work where all training data was either recorded by the musicians' themselves or acquired from open access datasets.

Acknowledgments

This research is supported by the UKRI Centre for Doctoral Training in Artificial Intelligence and Music (EP/S022694/1).

References

[1] Antoine Caillon and Philippe Esling. 2021. RAVE: A variational autoencoder for fast and high-quality neural audio synthesis. http://arxiv.org/abs/2111.05011

[2] Michelle Carney, Chong Li, Edwin Toh, Ping Yu, and Jesse Engel. 2021. Tone Transfer: In-Browser Interactive Neural Audio Synthesis. In *Joint Proceedings of the ACM IUI 2021 Workshops*.

[3] Franco Caspe, Jordie Shier, Mark Sandler, Charalampos Saitis, and Andrew McPherson. 2025. Designing Neural Synthesizers for Low-Latency Interaction. *Journal of the Audio Engineering Society* (2025).

[4] Jesse Engel, Lamtharn (Hanoi) Hantrakul, Chenjie Gu, and Adam Roberts. 2020. DDSP: Differentiable Digital Signal Processing. In *8th International Conference on Learning Representations*.

[5] Stefano Fasciani and Lonce Wyse. 2018. Vocal Control of Sound Synthesis Personalized by Unsupervised Machine Listening and Learning. *Computer Music Journal* 42, 1 (2018), 37–59. https://doi.org/10.1162/comj_a_00450

[6] Nicola Privato, Victor Shepardson, Giacomo Lepri, and Thor Magnusson. 2024. Stacco: Exploring the Embodied Perception of Latent Representations in Neural Synthesis. In *International Conference on New Interfaces for Musical Expression*.

[7] Jordie Shier, Rodrigo Constanzo, Charalampos Saitis, Andrew Robertson, and Andrew McPherson. 2025. Designing Percussive Timbre Remappings: Negotiating Audio Representations and Evolving Parameter Spaces. In *International Conference on New Interfaces for Musical Expression*.

[8] Jordie Shier, Charalampos Saitis, Andrew Robertson, and Andrew McPherson. 2024. Real-time Timbre Remapping with Differentiable DSP. In *Proceedings of the International Conference on New Interfaces for Musical Expression*.

[9] Dan Stowell. 2010. *Making music through real-time voice timbre analysis: machine learning and timbral control.* PhD Thesis. Queen Mary University of London.