FROM TAPS TO DRUMS: AUDIO-TO-AUDIO PERCUSSION STYLE TRANSFER

André C. Santos **Amílcar Cardoso**
CISUC, DEI, University of Coimbra, Portugal
{andresantos, amilcar}@dei.uc.pt

ABSTRACT

A common, and arguably innate, human response when listening to music is to tap one's foot to mark the regular pulse of the beat. A more complex form of interactive synchronization occurs when listeners tap out rhythmic patterns using their fingers, hands, or even some form of improvised drumsticks.

In this late-breaking demo, we explore this interaction by leveraging the style transfer capabilities of a neural audio synthesis model by training it on a drum dataset and feeding it tapped rhythm recordings at inference time. We also provide a concise and high-level overview of the results, which, in our assessment, not only justify further research but also establish an intriguing baseline for future investigations. Finally, we point out several future research directions.

1. INTRODUCTION

The exploration of tapped rhythms is arguably a fairly common impulse among people, and perhaps even more so amongst musicians, particularly drummers and percussionists. It is not infrequent to see a drummer tapping on a desk while listening to a song, practising it and trying to replicate the song's drum part or simply improvising over it; or sometimes a percussionist will simply improvise a rhythmic pattern on a surface using only his/her imagination. Yet, this experience is, to the best of our knowledge, greatly under-studied, making it a relevant topic for further research. Moreover, recent studies have hypothesized that multi-voiced drum patterns can be represented in a more natural and compressed representation such as a dual-voice reduction [1].

Considering the current opportunities presented by emerging generative AI technologies, we aim to create a tool that harnesses this interaction's potential and translates it into tangible results. A key principle guiding our efforts is to ensure that it remains intuitive and accessible to a wide audience, much like Piano Genie for example [2]. To achieve this, we experiment with audio-to-audio style transfer and we show how it can be a viable option for this tool by briefly discussing the results.

Two works are crucial for the development of this application: the RAVE model by A. Caillon and P. Esling [3] and the GMD dataset by J. Gillick et al. [4]. We delve further into each of these aspects in the next section.

2. BACKGROUND

RAVE is a "variational autoencoder for fast and high-quality neural audio synthesis" [3]. It was first developed in 2021 and its main appeal is apparent in the subtitle: it can work in real-time applications and is capable of synthesizing high-quality audio, i.e. up to 48kHz. Until this point, the authors contend that no model had demonstrated this capability. Their success can be attributed to the relatively straightforward model architecture, primarily consisting of Convolutional Neural Networks (CNNs). However, the primary factor contributing to their achievement is the novel training procedure proposed by the authors.

The training is divided into two different stages: Firstly, the representation learning stage; and secondly the adversarial fine-tuning stage. In the representation learning stage, the authors train both the encoder and decoder with the following loss function

$$\mathcal{L}_{\text{vae}}\left(\mathbf{x}\right) = \mathbb{E}_{\hat{\mathbf{x}}\sim p(\mathbf{x}|\mathbf{z})}[S(\mathbf{x},\hat{\mathbf{x}})] + \beta \times \mathcal{D}_{\text{KL}}\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})\right],$$

where $S(\mathbf{x},\hat{\mathbf{x}})$ is the multiscale spectral distance proposed in [5], $\mathcal{D}_{\text{KL}}$ is the Kullback-Leibler (KL) divergence, and β is simply a parameter to weight the KL divergence as proposed in [6]. The multiscale spectral distance is an amplitude spectrum-based distance which ensures that the model is able to capture significant perceptual characteristics of the signal irrespective of a poorly reconstructed phase, which in this case is inconsequential. In this way, the encoder is able to learn a good representation of the input signals.

As for the adversarial fine-tuning stage, its goal is to make the reconstruction of the original signal as accurate as possible. In this stage, the encoder is fixed and only the decoder is trained. The authors take inspiration from GANs by confronting a generation loss function with a discrimination loss function. They also add a feature matching loss to the total loss [7]. For the sake of brevity, the formulas are not presented here, but more details can be found in the paper [3].

The other work central to our own is "Learning to Groove with Inverse Sequence Transformations" [4]. The authors set out to explore how to turn "abstract musical ideas" - like rhythmic patterns - into expressive performances given the high importance of the performative aspect in music. They propose a series of new tasks, among which the "Tap2Drum" task is particularly pertinent to our research wherein they seek to convert a monophonic drum pattern into a complete drum "groove". [1] Furthermore, the authors built both a dataset, the Groove MIDI Dataset (GMD), and a variational autoencoder model, GrooVAE, to solve the tasks they proposed. They recorded over 13 hours of drum recordings played by professional drummers containing dynamics information as well as precise time alignment. As the dataset's name suggests, one crucial difference between this work and RAVE is that the authors worked only on a symbolic level instead of on raw audio waveforms.

3. STYLE TRANSFER

We propose a novel approach to the Tap2Drum task. Given the expressive nature of the exploration of tapped rhythms on arbitrary surfaces, we argue it is fundamental to base the modelling of tapped rhythms on the raw audio waveforms of the input signals. This way the model is able to capture the subtle nuances and intricacies of the tapped performance.

With this in mind, and given the fact that the RAVE model was found to be very apt at doing style transfer, we used the GMD dataset - which also contains the respective audio files for each MIDI drum pattern - to train a RAVE model and examine how it would fare at doing style transfer between tapped rhythms and drum sounds.

Specifically, we trained RAVE's v3 [2] for a total of three million steps, with Wassertwein regularization and causal convolutions. [3] We have yet to conduct a formal evaluation of the results, but listening to them and evaluating them purely from a qualitative point of view, there are some things we can point out:

- The model is proficient at capturing subtleties, i.e., it can discern that it is not just one instrument but comprises several, yet it is somewhat delicate and can be prone to inaccuracies. It lacks strong conviction, meaning that when the inputs are low in volume, it may not fully translate them into more standard and expectable drum sounds (for example, it tends to replicate the sound of striking the rim of a snare drum with a drumstick rather than striking the drum itself).
- The reconstructed drum sounds have a *lo-fi* quality to them. There is potential for improvement, possibly by incorporating more data, refining the training steps, or just using some more standard pre-processing and post-processing steps like normalization and noise reduction. On the other hand, this also opens the possibility for novel sound exploration.
- It is interesting to note that at times, the model translates the input into low tom-toms, while on other occasions, it interprets it as hi-hats. The reasons for these variations require further investigation to better understand the model's controllability.
- The model struggles to consistently capture what should be hi-hats, but it demonstrates a better ability to identify the kick drum and snare.
- It is also worth noting the experimental aspect of providing a drum kit as input and observing the model's reconstructed output. It captures the original drum sounds pretty reliably and reconstructs them accurately but the sound quality dropout is evident.

This is only an initial very high-level and qualitative evaluation of the results and further tests and examination are required to compile a full and formal evaluation of the style transfer task.

4. CONCLUSION

In this Late-Breaking Demo, we share preliminary results related to an innovative approach for the Tap2Drum task, initially introduced in [4]. Our approach involves training a RAVE model [3] using the Groove MIDI Dataset (GMD), a drum dataset. During inference, we directly apply this trained model to tapped rhythm recordings, examining its ability to do style transfer from taps to drum sounds.

In our opinion, the results are intriguing and warrant further attention in future research. Namely, as future avenues of research, we intend to evaluate the style transfer in a more formal and standard way; investigate why RAVE effectively performs this type of style transfer, identifying the specific features or patterns it recognizes and utilizing this understanding to create control instructions; explore the possibility of introducing voice *ad libs* at inference time to better control the output; enhance the RAVE model by incorporating a deeper understanding of musical structure and rhythmic patterns, potentially through the integration of Recurrent Neural Networks (RNNs); consider additional training possibilities, such as dataset augmentation or different training parametrization; explore the need for supervised training techniques for better representation learning and output reconstruction; and systematically vary the recording conditions within the test dataset, pre-processing, and post-processing steps to assess how the model generalizes across various contexts and conditions.

5. ACKNOWLEDGMENTS

We would like to express our gratitude to Mathew Davies for his guidance and support. This work is funded by national funds through the FCT - Foundation for Science and

[1] In this context, we use the term "groove" instead of "pattern" to emphasize the significance of expressive elements such as velocity, rather than merely focusing on playing the correct note at the correct time.

[2] v3 contains some improvements over the original version presented in the paper, namely snake activation, descript discriminator and Adaptive Instance Normalization for real style transfer. For more information visit the GitHub page: `https://github.com/acids-ircam/RAVE`.

[3] One million steps for the representation learning stage, and two million steps for the adversarial fine-tuning.

Technology, I.P., within the scope of the project CISUC - UID/CEC/00326/2020 and by European Social Fund, through the Regional Operational Program Centro 2020. The main author is also individually funded by Foundation for Science and Technology (FCT), Portugal, under the PhD fellowship grant 2021.05653.BD

6. REFERENCES

[1] B. Haki, B. Kotowski, C. Lee, and S. Jorda, "TapTamDrum: A Dataset for Dualized Drum Patterns," in *Proceedings of the 24th International Society for Music Information Retrieval Conference*. ISMIR, Nov. 2023, to appear.

[2] C. Donahue, I. Simon, and S. Dieleman, "Piano genie," in *Proceedings of the 24th ACM Conference on Intelligent User Interfaces*. ACM IUI, 2019.

[3] A. Caillon and P. Esling, "RAVE: A variational autoencoder for fast and high-quality neural audio synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2111.05011

[4] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman, "Learning to groove with inverse sequence transformations," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2269–2279. [Online]. Available: https://proceedings.mlr.press/v97/gillick19a.html

[5] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "Ddsp: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1x1ma4tDr

[6] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2016.

[7] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.