The Rhythm In Anything

Michael Shell, John Doe, and Jane Doe

Index Terms—Music Information Retrieval, Audio Generation, Drum Synthesis, Artificial Intelligence

1 Introduction

The Rhythm In Anything, or TRIA, is an artificial intelligence model trained to generate drums with audio-prompted rhythm and timbre. The work from Patrick O'Reilly et al. fall within Music Information Retrieval discipline, which is the science of retrieving information from music.
The novelty here is the high fidelity drum recordings, generated from a rhythm and a timbre audio. The model takes simple rhythm patterns, an audio of some beatboxing or tapping sounds; and as drumkit timbres, it only takes an example recording. The strength of this model is not only the high fidelity of the audio generated, but also that it works even with timbres it was not trained on.
The article focuses on how the dualization of rhythm and timbre allows the model to give better synthetizations. The code is fully available on their github and webpage.

2 State of the Art - Before Publication

At the time of the publishing of their article,

3 Contribution

The authors partition their contribution into three separate parts. First, the model, then the dualization task, and finally how the evaluations carried out show the performances of the model.

3.1 Dataset

3.2 Tasks

3.3 Methodology

3.4 Experiments

The authors conducted subjective and objective evaluations to assess the performances of TRIA. They wanted to measure the quality of the synthetizations, as well as how close the audio generated was to its audio-prompted timbre and rhythm. For comparison purposes, in these evaluations the authors compare audio generated from the model TRIA 2-band with random audios extracted from the dataset MoisesDB, and also with audio generated from the model MelodyFlox 0.2.

The subjective evaluation was carried out in order to verify how musically pleasing the generated audios were, and also how the synthetizations from TRIA compared to MelodyFlow's generations and the random excerpts. Evaluators here are humans, recruited through the platfomr Prolific. The evaluations were done through ReSEval, which stands for Reproducible Subjective Evaluation. Which is a framework used for building subjective evaluations. In order to have an homogeneous testers base, the people recruited had to pass a listening test. Out of the 120 persons originally recruited, 116 passed the listening test and went on with the evaluation of the audios.
For the subjective evaluation, listeners rated audio from 80 sets. A set is composed of an audio rhythm prompt, the associated generation from TRIA 2-band, and from MelodyFlow 0.2, and a drum extract randomly taken from MoisesDB. The 80 generations were made with ten rhythm prompts: five with tapping sounds, and five with beatboxing audio, on 8 different audio timbre prompts. Each audio clip generated lasts for three to four seconds, which is the duration of the given rhythm prompt. Three pairwise comparisons were evaluated for each set by five persons, comparing TRIA to MelodyFlow, Tria to the random excerpt, and MelodyFlow to the random excerpt. Each listener was given ten pairwise comparisons to evaluate.

4 Discussion

5 State of the Art - After Publication

6 Conclusion

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.