Exploring the potential of interactive Machine Learning for Sound Generation: A preliminary study with sound artists

Gerardo Meza
Universidad Nacional Autónoma de México
3004 University Ave.
CDMX, México
gerardomeza@comunidad.unam.mx

ABSTRACT

Interactive Machine Learning (IML) is an approach previously explored in music discipline [10, 11, 13, 18]. However, its adaptation in sound synthesis as an algorithmic method of creation has not been examined. This article presents the prototype *ASCIML*, an Assistant for Sound Creation with Interactive Machine Learning, that allows musicians to use IML to create personalized datasets and generate new sounds. Additionally, a preliminary study is presented which aims to evaluate the potential of *ASCIML* as a tool for sound synthesis and to gather feedback and suggestions for future improvements. The prototype can be used in *Google Colaboratory* [2] and is divided into four main stages: Data Design, Training, Evaluation and Audio Creation. Results from the study, which involved 27 musicians with no prior knowledge of Machine Learning (ML), showed that most participants preferred using microphone recording and synthesis to design their dataset and that the Envelopegram visualization was found to be particularly meaningful to understand sound datasets. It was also found that the majority of participants preferred to implement a pre-trained model on their data and relied on hearing the audio reconstruction provided by the interface to evaluate the model performance. Overall, the study demonstrates the potential of *ASCIML* as a tool for hands-on neural audio sound synthesis and provides valuable insights for future developments in the field.

Author Keywords

Interactive Machine Learning, Sound Synthesis, Generative Models, Personalized Datasets, Audio Generation, Machine Learning in Music.

CCS Concepts

•**Applied computing** → *Sound and music computing;* •**Human-centered computing** → **Interaction design; Interface design prototyping;**

NIME'23, 31 May2 June, 2023, Mexico City, Mexico.

1. INTRODUCTION

Interactive Machine Learning (IML) is a revolutionary approach that empowers users to create Machine Learning-based systems tailored to their specific needs and goals [9, 7]. This paradigm provides accessibility to experts in the disciplines where these algorithms are inserted, allowing them to direct learning processes through trial and error [1]. This approach has been applied in various projects in music, including generic tools such as the *Teachable Machines* [4], which allows users to classify sounds without coding, and more specialized tools for the intersection of sound creation, parameter mapping and body gesture such as: *Wekinator* [10], *InteractML* [16], *Learner.js* [13] and*G-IMLeT* [18]. Although the implementation of IML models in music has been found since its inception, its application in neural audio sound synthesis has not been fully explored. This is likely due to the challenge that the high dimensionality and complexity of sound data poses for common ML architectures. However, recent studies utilizing generative model have yield promising results in the field of sound synthesis [15, 8, 3]. These models have shown that it is possible to create new sounds by learning the underlying structure of existing audio samples.

This article presents with *ASCIML* an integration of IML and a generative model to address this challenge. This user-friendly prototype allows to interactively create personalized datasets and conduct model training to generate short audios with no prior ML knowledge. The user navigates four principal sections: Data design, Training, Evaluation and Audio Creation. In these sections, the interface provides a variety of data visualization tools and allows the user to controls various parameters such as the number of training epochs, batch size and learning rate. Additionally, the prototype includes the option to use a pre-trained model, which can improve the audio reconstruction process and save time for the user.

Secondly to presenting the architecture, this article shows preliminary results of the application of the prototype with a group of undergraduate musicians. The study aims to evaluate the use of IML as a tool and gather insights into the preferences of musicians in creating personalized datasets, the effectiveness of different data visualizations, the impact of a pre-trained model on the audio reconstruction process, and the strategies used by musicians to generate new sounds. The prototype and results and further information can be consulted at `https://asciml.github.io/ASCIML/`.

2. THE PROTOTYPE

ASCIML is a tool that enables musicians to use ML to synthesize short audio clips within a domain specified by examples. It can be used in *Google Colaboratory*, and the overall cells are grouped in four sections (Figure 1):

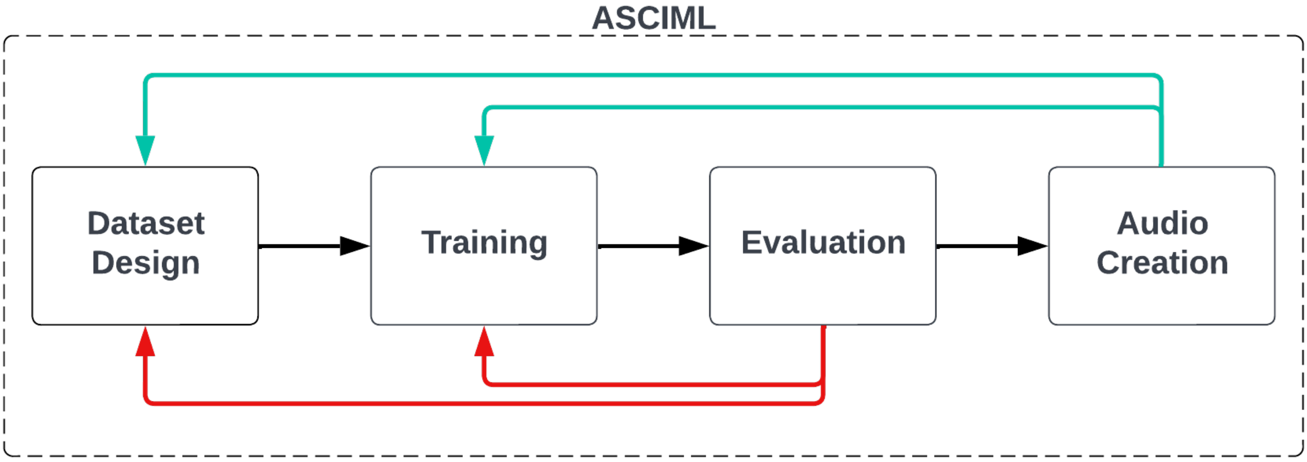

Figure 1: Interactive Machine Learning workflow. Arrows indicate possible directions that can be taken by the user.

Dataset Design: In this section, the user can create group IDs and generate personalized audio datasets choosing from the following techniques: microphone recording, audio synthesis (Amplitude Modulation [6], Frequency Modulation [5], Subtractive synthesis, Additive synthesis [17]) and upload pre-existing audio files. Audio features that describe fundamental frequency, amplitude envelope and timbre are extracted to summarize and visualize the groups to the user through six types of visualizations: Tables, Histograms, F0 dispersion, Spectral Centroid dispersion, Envelopegrams [1], and PCA.

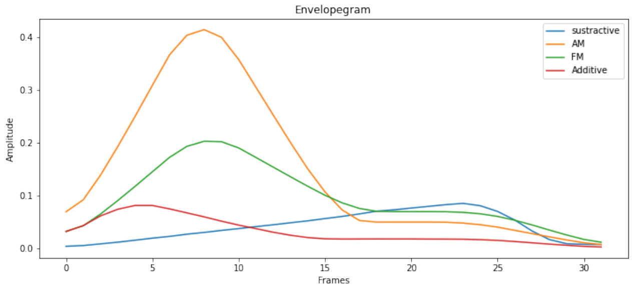

Figure 2: Envelopegram visualization example with four different synthesis techniques (Sustractive synthesis, Amplitude modulation, Frequency modulation and Additive synthesis).

Training: This cell involves training a Variational Autoencoder (VAE) model [12, 14] on the signals of the dataset created in the first stage. The VAE architecture takes as input 1 second of raw audio signal downsampled to 16Khz to reduce computational cost without significantly affecting the overall envelope and spectral changes. The encoder is composed of 4 convolutional layers with ReLU activation, 64, 128, 256, and 512 filters respectively, and a stride of 4,4,4,3. The filter size is fixed at 66, 1. These layers are followed by a Flatten layer and two Dense layers of 256 and 128 units. Through the interface, the user can select a new model to train or a model pre-trained for 3000 epochs with 640 synthesized sounds. Additionally, the number of training epochs, batch size, and learning rate can be controlled and set with sliders.

Evaluation: This cell provides the user with loss visualization, signal reconstruction juxtaposed with the original signal, and the audio result to guide the training process (Figure 3).

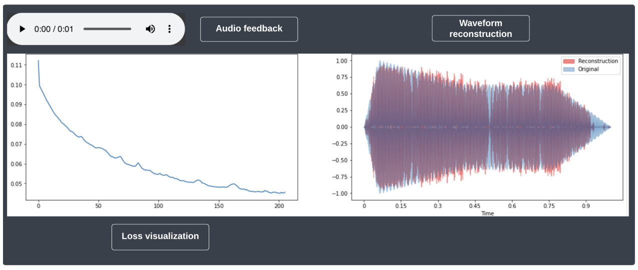

Figure 3: ASCIML's cell for Audiovisual Model Evaluation.

Audio Creation: This section allows the user to create new audio samples by linear interpolating between different sounds, represented by vectors, in the latent space generated by the model. The user has access to a 2D scatter plot representation of this information, and can interactively provide audio IDs of two targets, control the percentage of change, and see a visualization of the waveform, listen, save, and download the audio generated.

3. MUSICIANS AND ASCIML

This study aimed to evaluate the use of *ASCIML* as a tool for musicians. To accomplish this, two groups of undergraduate musicians studying Music and Artistic Technology at the UNAM, ENES Morelia were recruited, one with 16 participants just starting their studies and the other with 11 participants who have been studying for over a year. The activity consisted of a brief introduction to key concepts of ML, datasets, model training and evaluation. Later, each participant had 2 hours to generate a dataset with $\geq$ 60 audio files, created with the three techniques available, train a model from scratch and evaluate its performance, and generate new sound. Later, the participants were asked to use the pre-trained model option with the same dataset and repeat the last steps. The objectives of the study were to: a)Assess the preferences of musicians in creating personalized datasets, b) Evaluate the effectiveness of different data visualizations in understanding the dataset, c)Investigate the impact of a pre-trained model, d)Assess the strategies used by musicians in generating new sounds and the factors that influence these decisions, e)Understand the overall experience of the musicians in using IML for sound synthesis

[1] This visualization plots an RMS average per audio group created by the user (Figure 2).

and gather suggestions for future improvements.

3.1 Preliminary results

The study found that timbre was the most utilized criteria to generate their datasets. Also, that the participants favored using microphone recording and synthesis to create their dataset, with 76% rating these methods as efficient or very efficient (Figure 4). The Envelopegram visualization and tables were also found to be particularly useful, with the majority of participants stating that they provided relevant information. In contrast, histograms were voted as the least useful of all the visualizations provided.

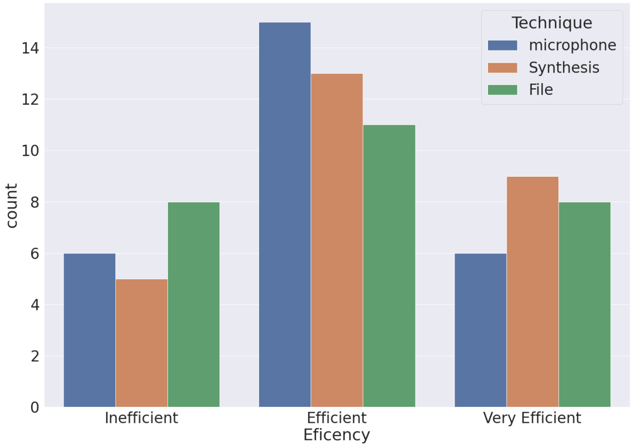

Figure 4: Efficiency results reported by users utilizing different audio dataset generation techniques (microphone, Synthesis, file upload).

The participants of the first group suggested integrating more auditory information into the interface. This feature was implemented and tested with the second group. In terms of evaluating the model, over 50% of the participants relied on hearing the audio reconstruction provided by the interface, while 40% relied on the waveforms visualization. Furthermore, almost half of the participants considered that the pretrained model gave better results in reconstructing their data in the given time of the activity.

When it came to creating new sounds, a significant proportion of participants were able to obtain musically interesting sounds within a short timeframe, specifically, 37% of participants reported obtaining these sounds in under 20 minutes (Figure 5). An experimental approach to sound synthesis was common between the participants (44.4%), while over 30% focused on timbre contrast and affinities (Figure 5). Lastly, during both studies it was also observed that nearly all composers spent most of the activity creating the datasets, small collections of sounds at the time, followed by the exploration of their interpolation results.

4. DISCUSSION

The results of this study indicate that IML has the potential to be a valuable tool for sound synthesis and composition. Specifically, this study found that the majority of participants preferred to create their datasets using microphone recording and synthesis over loading pre-existing audio files. This research's hypothesis is that the process of searching examples that meet one's creative needs on the web or in owned libraries can be time consuming and less engaging compared to directly creating and providing examples. Additionally, the Envelopegram visualization was found to be particularly useful for understanding the dataset, with most participants stating that it provided relevant information. This is likely because the Envelopegram summarize temporal information and represents it in a more synthesized way, with a profile representing the entirety of a set.

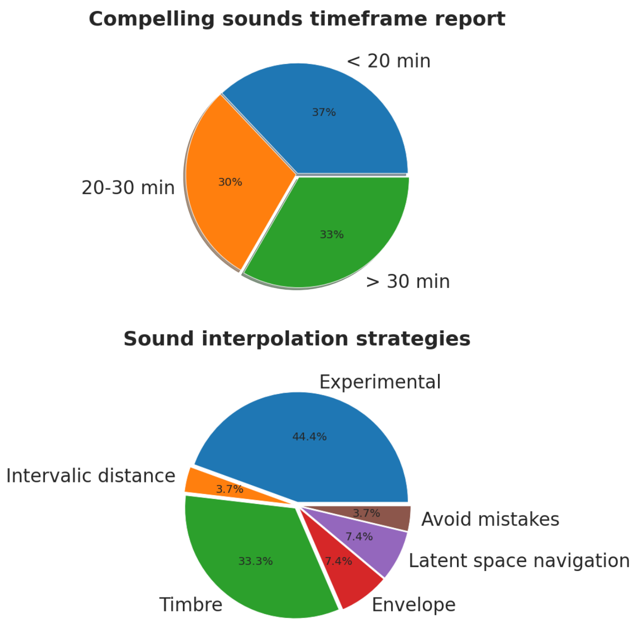

Figure 5: User's compelling sound creation by means of interpolation timeframe and strategie report".

Furthermore, the inclusion of auditory information in the second study was found to be beneficial, as it was used extensively by the participants. This suggests that providing more auditory cues in the interface can help musicians better understand and utilize the tool. This can be seen in the model evaluation process, where the majority of participants relied on hearing the audio reconstruction provided by the interface, while a significant proportion also leaned on waveform visualization. This may be because hearing the audio reconstruction is closely related to the results of direct evaluation conducted by [11] and resonates with familiar tools and concepts. When comparing the results of a pre-trained model versus a new model, almost half of the participants considered that the pre-trained model gave better results in reconstructing their data in the given time of the activity.

When generating new sounds, the majority of participants approached the task experimentally, with no preconceived plan, while a significant proportion also paid attention to the timbral contrast and similarities between sounds. These findings suggest that sound quality and experimentation were key considerations for participants when creating new sounds. This may have been influenced by the limited duration of the activity, as well as the participants' unfamiliarity with the prototype or their compositional method. Further research and data is needed to better understand these factors.

Despite anticipating the presence of reconstruction noise, a substantial number of participants were able to successfully generate musically compelling sounds within a short period of time. This highlights the effectiveness of the prototype as a tool for musicians in the field of sound synthesis, even an aesthetic embrace of error. Overall, the study highlights the potential of IML as a valuable tool in the field of sound synthesis, but also highlights the need for further research to fully understand and improve the user experience.

5. CONCLUSIONS AND FUTURE WORK

This article demonstrates the potential of ASCIML as a tool for musicians to create personalized datasets and generate new sounds. The majority of participants preferred using microphone recording and synthesis to create their datasets and focused on timbre when designing them. The data visualizations provided in the study were also found to be effective in understanding collections of sounds, with the Envelopgram proving to be particularly meaningful. Additionally, the use of a pre-trained model was found to speed the audio reconstruction process.

Overall, the study established that musicians were engaged in the activity and found it to be a valuable learning experience. However, there are still some areas that could be improved in future research, such as expanding the audio-length reconstruction capabilities of the model and provide more auditory information and interactive visualization of the data within the interface. Also, further insights will be obtained by conducting a study with a longer duration in which participants from different backgrounds incorporate the generated sounds in musical contexts.

6. ACKNOWLEDGMENTS

The author would like to extend his sincerest appreciation to the musicians who participated in the study for their valuable time and feedback. Thanks to Mariana Sepúlveda (UNAM) for her insights and comments on the manuscript. And finally, a special thanks to the Music and Artistic Technology Department at the ENES Morelia and the Mexican Center for Music and Sonic Arts (CMMAS) for their support and assistance throughout the research.

7. ETHICAL STANDARDS

This work is currently being carried out with financial support from the posgraduate program of Music Technology at the UNAM. It is important to note that all students participated in this study voluntarily and survey results were collected anonymously.

8. REFERENCES

[1] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

[2] E. Bisong. *Building machine learning and deep learning models on Google cloud platform.* Springer, 2019.

[3] A. Caillon and P. Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021.

[4] M. Carney, B. Webster, I. Alvarado, K. Phillips, N. Howell, J. Griffith, J. Jongejan, A. Pitaru, and A. Chen. Teachable machine: Approachable web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–8, 2020.

[5] J. M. Chowning. The synthesis of complex audio spectra by means of frequency modulation. *Journal of the audio engineering society*, 21(7):526–534, 1973.

[6] C. Dodge and T. A. Jerse. *Computer music: synthesis, composition, and performance.* Macmillan Library Reference, 1985.

[7] J. J. Dudley and P. O. Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37, 2018.

[8] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077. PMLR, 2017.

[9] J. A. Fails and D. R. Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45, 2003.

[10] R. Fiebrink and P. R. Cook. The wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR 2010)(Utrecht)*, volume 3, 2010.

[11] R. Fiebrink, P. R. Cook, and D. Trueman. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 147–156, 2011.

[12] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[13] L. McCallum and M. S. Grierson. Supporting interactive machine learning approaches to building musical instruments in the browser. In *Proceedings of the International Conference on New Interfaces for Musical Expression*, pages 271–272. Birmingham City University Birmingham, UK, 2020.

[14] S. Min, B. Lee, and S. Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

[15] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[16] N. Plant, C. Hilton, M. Gillies, R. Fiebrink, P. Perry, C. González Díaz, R. Gibson, B. Martelli, and M. Zbyszynski. Interactive machine learning for embodied interaction design: A tool and methodology. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 1–5, 2021.

[17] J. O. Smith. *Spectral Audio Signal Processing.* `http`:http://ccrma.stanford.edu/ jos/sasp///-`ccrma.stanford.edu/~jos/sasp/`, accessed 2023. online book, 2011 edition.

[18] F. G. Visi and A. Tanaka. Towards assisted interactive machine learning: exploring gesture-sound mappings using reinforcement learning. In *ICLI 2020the fifth international conference on live interfaces*, pages 9–11, 2020.