MASKED TOKEN MODELING FOR ZERO-SHOT ANYTHING-TO-DRUMS CONVERSION

Patrick O'Reilly[1] **Hugo Flores Garcia**[1] **Prem Seetharaman**[2] **Bryan Pardo**[1]
[1]Department of Computer Science, Northwestern University , USA
[2]Adobe Research, San Francisco, USA
patrick.oreilly2024@u.northwestern.edu

ABSTRACT

Musicians often represent drum beats through sound gestures such as vocal imitation and finger tapping. While these gestures can convey rich rhythmic information, realizing them as fully-produced drum beats requires time and skill. We propose a system for mapping arbitrary percussive sound gestures to high-fidelity drum recordings. Our system, dubbed TRIA (**T**he **R**hythm **I**n **A**nything), takes as input two audio prompts – one specifying the desired drum timbre, and one specifying the desired rhythm – and generates audio satisfying both prompts (i.e. playing the desired rhythm with the desired timbre). TRIA can synthesize realistic drum audio given rhythm prompts from a variety of non-drum sound sources (e.g. beatboxing, environmental sound) in a zero-shot manner, enabling novel creative interactions.

1. INTRODUCTION

Musicians often communicate rhythmic ideas through sound gestures such as beatboxing and finger tapping, leveraging accessible sound production mechanisms to convey complex percussive textures.

In many musical idioms, drum beats contain many carefully layered and arranged sound sources (e.g. kick, snare, toms, cymbals, shaker). As a result, even with modern recording and production technology, it can take considerable time and skill to turn a sound gesture into a fully-produced drum beat that embodies both the musician's desired rhythm and timbre. We seek a tool to lets a user quickly input a gestural rhythmic sketch (e.g. drumming a rhythm on the table) and a brief example of the desired sonic palette (e.g. a short clip of someone playing the darbuka) and do zero-shot style transfer, creating a drum pattern in the timbre of the sonic palette.

Approaches for realizing music from simplified or gestural representations have often focused on output in the symbolic domain [1, 2]. While deep neural network-based systems for audio music generation [3, 4] have shown promise in converting textual and stylistic prompts into high-fidelity audio recordings, existing systems lack the fine-grained rhythmic control required to realize the idea embodied in a sound gesture. Systems such as RAVE [5] and DDSP [6, 7] allow for real-time timbre transfer between sound sources, including drums, but must be retrained to mimic a new target timbre.

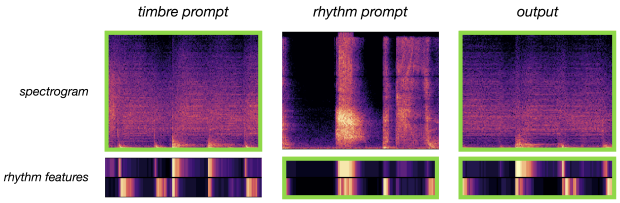


Figure 1. Given two prompt recordings specifying the desired drum timbre (left) and rhythm (center), TRIA synthesizes a new drum recording satisying both prompts (right). Here we visualize timbre via a spectrogram, and rhythm via the two-band spectral features described in Section 2.

To bridge this gap, we propose TRIA (**T**he **R**hythm **I**n **A**nything), a deep neural network-based system for realizing rhythmic sound gestures as high fidelity drum recordings. TRIA takes as input two audio prompts – one specifying the desired drum timbre, and one specifying the desired rhythm – and generates audio satisfying both prompts (i.e. playing the desired rhythm with the desired timbre). By facilitating zero-shot conversion of arbitrary sounds into drum beats, we hope to enable new creative interactions for novice and expert musicians alike. We provide audio examples on our supplementary webpage [1] .

2. TRIA

TRIA generates drum recordings using masked language modeling of tokens obtained from Descript Audio Codec (DAC) [8], a neural network-based audio tokenizer. The overall architecture is similar to those of SoundStorm [9] and VampNet [10] albeit much smaller, consisting 12 stacked standard transformer encoder [11] layers with a hidden dimension of 512, 8 attention heads each, and rotary positional encoding [12], for a total of 43m parameters. TRIA takes as input a buffer of partially-masked

[1] https://oreillyp.github.io/tria

© P. O'Reilly, H. F. Garcia, P. Seetharaman, and B. Pardo. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** P. O'Reilly, H. F. Garcia, P. Seetharaman, and B. Pardo, "Masked Token Modeling for Zero-Shot Anything-to-Drums Conversion", in *Extended Abstracts for the Late-Breaking Demo Session of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

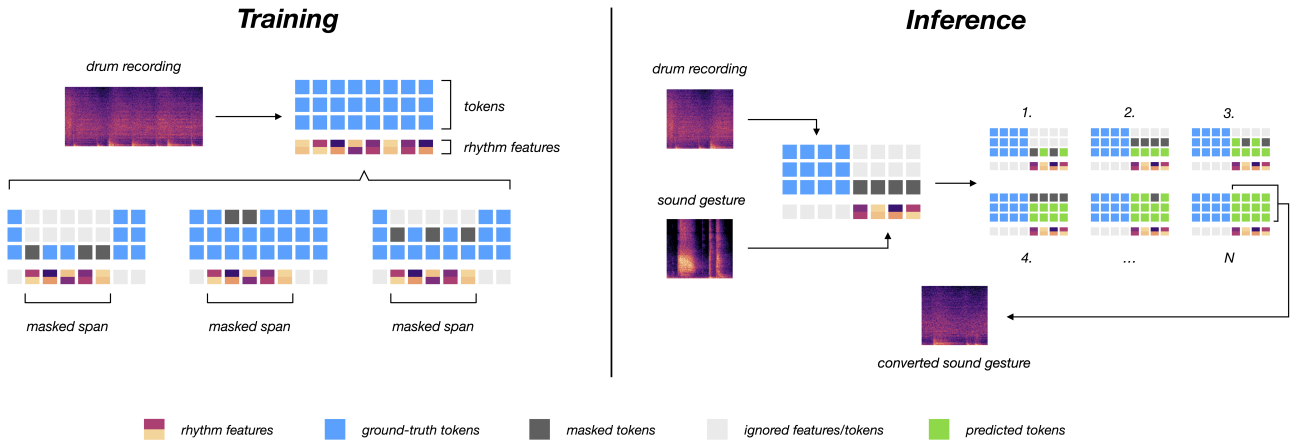


Figure 2. The training (left) and inference (right) procedures for the proposed TRIA system. Training involves predicting randomly-masked tokens within a selected frame span and codebook level, as illustrated in three examples. Inference requires the prediction of all tokens for a given rhythm prompt conditioned on aligned rhythm features and the prefixed timbre prompt. While DAC contains 9 codebooks, we picture only 3 for legibility.

token indices of shape $(C \times T)$ where C is the number of tokenizer codebooks and T is the number of frames, and outputs a predicted distribution over token indices at masked positions within a given codebook. This allows for MaskGIT-style training and sampling, in which masked indices are gradually replaced by predicted indices one codebook at a time [13].

To incorporate rhythm information, we augment our input token buffer with coarse-grained spectral features computed at the codec frame rate. Given an 80-bin mel-spectrogram of the input audio, we adaptively split the frequency dimension into two equal-energy bands. We then apply standardization followed by a sigmoid nonlinearity so that all features fall in the range $[0, 1]$. Finally, we quantize all values to 33 steps $(0, \frac{1}{32}, \frac{2}{32}, ..., 1)$ to prevent leakage of timbre information into the rhythm features. To allow for robust performance on rhythm prompts recorded in-the-wild, we apply a random combination of noise, band-pass filtering, pitch shift, phase shift, and equalization to audio before computing rhythm features.

At each training iteration, we select a random codebook and span of consecutive frames in which to mask a subset of tokens. Rhythm features outside the span and tokens from “higher" codebooks within the span are zeroed, forcing the model to leverage the aligned rhythm features and surrounding token context to reconstruct the missing information. We train on drum stems from the MUSDB-18 dataset [14], which contains roughly 10 hours of high-quality drum audio.

At inference time, we concatenate the timbre prompt to the rhythm prompt as a prefix, mask all tokens for the rhythm prompt, zero all rhythm features for the timbre prompt, and predict tokens for the rhythm prompt. Thus, the model must infer the “missing" tokens using timbral information from the timbre prompt and rhythmic information from the rhythm prompt. The final predicted tokens are decoded to 44.1kHz audio using the DAC decoder. Our approach draws from aspects of other recent masked audio language modeling works, including the temperature-based mask sampling from VampNet [10], the causal masking bias from StemGen [15], and the use of additional aligned conditioning features from GenHancer [16] (in our case, rhythm features). Training and inference procedures are illustrated in Figure 2.

3. INTERFACE

We develop a simple Gradio [17] interface where users can upload timbre and rhythm prompts to realize sound gestures or convert between drum timbres. The system displays spectrogram and rhythm feature representations of the prompt and output audio, visualizing the conversion process as in Figure 1. Users can tweak fine-grained controls of the generative model (temperature, top-k, and top-p truncation [18] of the sampling procedure, classifier-free guidance weight, causal masking bias) and input a random seed to sample different variations from the same prompt. We find that truncating the predicted token probabilities results in more “typical" drum timbres and a focus on strong beats, while sampling from the full distribution allows for more diverse timbres and rhythmic ornamentations (e.g. fills). We also observe that higher guidance weights result in tighter adherence to the rhythm prompt.

4. DISCUSSION

TRIA obtains strong results in realizing drum beats from sound gestures and drum-to-drum timbre transfer, as the tight information bottleneck of the rhythm features forces the model to generate plausible drum textures given arbitrary rhythm prompts. This performance comes despite the relatively small model and dataset sizes; we hypothesize that the highly repetitive (and thus redundant) structure of drum recordings, coupled with the dominance of transients and noise-like textures, makes the masked token modeling task easier than for more structurally entropic signals such as speech.

5. REFERENCES

[1] C. Donahue, I. Simon, and S. Dieleman, "Piano genie," in *ACM IUI*, 2019.

[2] S. H. Hawley, "Pictures of midi: Controlled music generation via graphical prompts for image-based diffusion inpainting," *arXiv preprint arXiv:2407.01499*, 2024.

[3] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *Neural Information Processing Systems*, 2024.

[4] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," in *ICML*, 2024.

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

[6] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "Ddsp: Differentiable digital signal processing," *arXiv preprint arXiv:2001.04643*, 2020.

[7] J. Shier, A. R. Charalampos Saitis, and A. McPherson, "Real-time timbre remapping with differentiable dsp," in *NIME*, 2024.

[8] R. Kumar, P. Seetharaman, I. K. Alejandro Luebs, and K. Kumar, "High-fidelity audio compression with improved rvqgan," in *NeurIPS*, 2023.

[9] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "Soundstorm: Efficient parallel audio generation," *arXiv preprint arXiv:2305.09636*, 2023.

[10] H. Flores Garcia, P. Seetharaman, R. Kumar, and B. Pardo, "VampNet: Music generation via masked acoustic token modeling," in *Conference of the International Society for Music Information Retrieval (ISMIR)*, 2023.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[12] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, no. C, Mar. 2024. [Online]. Available: https://doi.org/10.1016/j.neucom.2023.127063

[13] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "Maskgit: Masked generative image transformer," in *CVPR*, 2022.

[14] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372

[15] J. D. Parker, J. Spijkervet, K. Kosta, F. Yesiler, B. Kuznetsov, J.-C. Wang, M. Avent, J. Chen, and D. Le, "Stemgen: A music generation model that listens," in *ICASSP*, 2024.

[16] H. Yang, J. Su, M. Kim, and Z. Jin, "Genhancer: High-fidelity speech enhancement via generative modeling on discrete codec tokens," in *Interspeech*, 2024.

[17] A. A. D. K. A. A. J. Z. Abubakar Abid, Ali Abdalla, "Gradio: Hassle-free sharing and testing of ML models in the wild," in *ICML Workshop on Human in the Loop Learning (HILL 2019)*, 2019.

[18] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," in *ICLR*, 2020.