InstrumentGen: Generating Sample-Based Musical Instruments From Text

Shahan Nercessian* **and Johannes Imort***
Native Instruments
{firstname.lastname}@native-instruments.com

* Equal contribution

Abstract

We introduce the *text-to-instrument task*, which aims at generating sample-based musical instruments based on textual prompts. Accordingly, we propose *InstrumentGen*, a model that extends a text-prompted generative audio framework to condition on instrument family, source type, pitch (across an 88-key spectrum), velocity, and a joint text/audio embedding. Furthermore, we present a differentiable loss function to evaluate the intra-instrument timbral consistency of sample-based instruments. Our results establish a foundational text-to-instrument baseline, extending research in the domain of automatic sample-based instrument generation.

1 Introduction

The synthesis of sounds and corresponding interfaces for controlling their timbre form a seminal topic in audio research [1]. Meanwhile, generative models have been successfully applied to images and text, where their convincing ability to draw novel samples from learned data distributions has already proved to be disruptive [2]. It becomes only natural to consider the implications of such technologies when applied in the context of audio and music production.

Several generative models have been proposed for neural audio synthesis. NSynth [3] uses a WaveNet autoencoder to synthesize pitched instrument samples. GANSynth [4] considers an instantaneous frequency representation to model signal phase. Differentiable digital signal processing (DDSP) [5] and related works [6, 7] construct autoencoders with differentiable synthesizer back-ends to promote controllability. A real-time variational autoencoder design was introduced in [8]. GANstrument [1] utilizes a feature descriptor achieved through adversarial domain confusion. These models all lack an interface for controlling audio generation via text input. Consequently, we have witnessed a surge in the development of text-to-audio systems generating compelling audio examples from text prompts. One particular family of approaches rely on neural audio codecs [9, 10] representing audio compactly as a set of discrete codes whose sequence can be learned using transformer-based language models. While initial approaches targeted speech [11, 12] and environmental sounds [13], follow-on works adapt techniques for text-to-music, generating entire musical passages from text [14, 15].

In this paper, we introduce a new task which we call *text-to-instrument*, whose aim is to generate musical instruments given a text prompt in a zero-shot manner. Under this task, we explicitly model a musical instrument as a collection of waveforms sampling an instrument's time-domain response across the axes of pitch (the fundamental frequency of a note) and velocity (the intensity with which a note is played). In this paradigm, we move beyond the constraints of any single parametric synthesizer, avoiding the expressivity limitations tied to its specific implementation details. As in [1], we note that injecting prior domain knowledge into the generative process via techniques like DDSP is indeed interesting, but is complementary to this work as such approaches naturally constrain the manifold that system outputs can live on [7]. Unlike text-to-music, which predominantly involves generation of

NeurIPS 2023 Workshop on Machine Learning for Audio.

a single audio example for the text prompt at inference, text-to-instrument systems must generate an ensemble of audio examples such that they are individually stabilized in pitch and timbrally consistent with one another, so that they can be assembled into a playable instrument that can be triggered in predictable ways. In summary, our primary contributions are:

- We introduce the text-to-instrument task.
- We propose *InstrumentGen*, a catered text-to-instrument solution expanding on a state-of-the-art text-prompted generative audio model to be, alongside a contrastive language-audio pretraining (CLAP) embedding [16] used as a joint audio/text representation, additionally conditioned on instrument family, source type (i.e. acoustic, electronic, or synthetic), pitch across the entire 88-key range of a standard full-length piano keyboard, and velocity.
- We present a differentiable loss to objectively assess the intra-instrument timbral consistency (TC) of sample-based instruments for our task by generalizing a (potentially multi-scale) log mel spectrogram loss [5, 17], and use it as an evaluation metric in this work.

2 Proposed Method

InstrumentGen is based on the MusicGen [15] architecture as a foundation, which consists of a neural audio codec and a language model designed to predict acoustic tokens based on conditioning signals. To improve audio quality, we replace the original EnCodec architecture [18] used in MusicGen with the Descript audio codec (DAC) [10], which addresses the issue of codebook collapse in previous models while achieving higher audio fidelity. Additionally, we introduce a set of new conditioning signals to the system: this includes instrument family, source type, pitch, and velocity, alongside a joint language-audio embedding [16]. This conceivably allows instrument samples to be inferred from either text or audio prompts, where we focus on the former but can also perform the latter. Fig. 1 gives an overview of our method.

2.1 Compressed Audio Representation

In this work, we employ the DAC as an intermediate representation of the monophonic input waveform $\mathbf{x} \in \mathbb{R}^{1\times L}$ (cf. Fig. 1), resulting in the discrete codes $\mathbf{c} \in \mathbb{R}^{C\times N}$. Here, L denotes the length of the waveform, N the sequence length of the acoustic tokens, and C the number of codebooks used. The DAC is trained on a broad spectrum of audio types, thereby making it also suitable for generating tonal one-shot instrumental sounds. We deliberately opt to model our task at a sampling rate of 44.1 kHz, as this would ultimately be a minimum requirement for real-world music production applications. We employ the corresponding pretrained model weights and fix them during training.

2.2 Language Model

For modeling the discrete audio tokens of single-shot instrumental samples, we consider a smaller, 60M parameter variant of the MusicGen transformer decoder [15], both to prevent overfitting and to provide faster inference. The resulting model consists of 12 decoder layers with 16 attention heads per layer and a transformer model dimension of 512. As in MusicGen [15], we predict and reconstruct audio from tokens of the 4 most significant [10] codebooks at each frame. Our predictor model learns to select tokens from codebooks of size 1024 using delayed pattern interleaving [15].

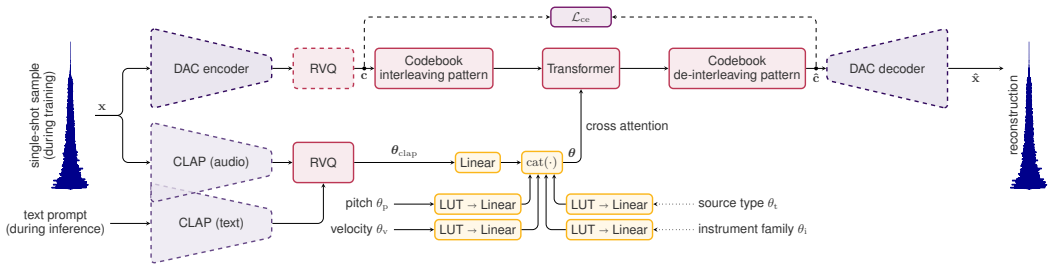

Figure 1: Overview of proposed method. Dashed lines indicate modules with fixed and pretrained weights during training. Source type and instrument family are not provided during inference.

2.3 Conditioning

Categorical conditioning We use a categorical conditioning scheme for instrument family θ_{i}, source type θ_{t}, pitch θ_{p}, and velocity θ_{v}, that consists of a lookup table (LUT) and a fully connected layer that maps the dimension of the categorical feature space to the inner dimension of the language model. When used, the instrument family and source type attributes in our dataset serve as supplementary metadata-driven timbral cues for generation. For pitch, we model the $P = 88$ range of notes spanned by a standard full-length keyboard, corresponding to musical instrument digital interface (MIDI) note numbers 21-108, and note that this is a significant expansion relative to the chroma feature used in [15]. We consider $V = 5$ velocity layers, according to the values that are actually present within our training dataset, corresponding to MIDI velocities 25, 50, 75, 100, and 127.

Joint text and audio conditioning Wu et al. [16] introduced a CLAP framework, which is designed to process both audio and text data and generate corresponding embeddings by using two separate encoders: one for audio and another for text. Both embeddings are further processed by a 2-layer MLP with ReLU activation to bring them into a fixed dimension of 512. The resulting model leverages a contrastive loss function, and is trained on musical signals to encourage the audio and text embeddings to be similar when they come from matching pairs. The audio encoder uses the HTS-AT (transformer-based) architecture [19], while the text encoder leverages RoBERTa [20].

When CLAP embeddings are used, we quantize them via residual vector quantization (RVQ) with learned codes, yielding $\boldsymbol{\theta}_{\mathrm{clap}}$. Since the various conditioning signals used here reflect global cues $\boldsymbol{\theta}$ for steering audio generation, they are fused with the transformer decoder by means of cross-attention.

3 Timbral Consistency Measure

Our task necessitates that waveforms comprising a generated instrument are timbrally consistent to one another. We provide an objective means to measure this relation, considering that signals within an instrument may vary in pitch. Specifically, mel-frequency cepstral coefficients (MFCCs) are ubiquitously known as timbral descriptors, as a cepstral lifter whose quefrencies correspond to the first handful of MFCCs can be used to estimate a signal's spectral envelope [21]. We integrate this notion into a differentiable objective function which generalizes a multi-scale log mel spectrogram loss [17], aggregating over all possible pairwise comparisons within the instrument. For an ensemble of waveforms $\mathbf{X} \in \mathbb{R}^{K \times L}$ comprising an instrument (with $K = PV$), the TC measure is defined as

$$\mathcal{L}_{TC}(\mathbf{X}) = \sum_{i=1}^{K} \sum_{j=i+1}^{K} \sum_{s=1}^{S} \|\mathbf{y}_i - \mathbf{y}_j\|_1^1 \tag{1}$$

over S scales, where for an arbitrary audio waveform $\mathbf{x}_k$ from $\mathbf{X}$ we have

$$\mathbf{y}_k = \mathbf{D}_{M_s}^{-1} \mathbf{D}_{m_s} \log \left[\mathbf{B}_s \left| \mathcal{F}_s(\mathbf{x}_k) \right|^p \right] \tag{2}$$

For each scale s, $\mathcal{F}_s$ denotes the respective short-time Fourier transform, $\mathbf{B}_s$ denotes an M_s-band mel transformation matrix, and $\mathbf{D}_{m_s} \in \mathbb{R}^{M_s \times M_s}$ denotes a discrete cosine transform basis matrix that has been masked with zeros for row indices greater than m_s MFCCs. When $m_s = M_s$ (i.e. all available MFCCs are considered), $\mathbf{D}_{M_s}^{-1} \mathbf{D}_{M_s} = \mathbf{I}$. With $m_s < M_s$, we effectively apply cepstral liftering that can neutralize the effect of the pitched excitation within the signal (cf. Fig. 2). We consider $S = 1$, $M_s = 80$, and $m_s = 13$, and $p = 1$ for demonstrative purposes in this work.

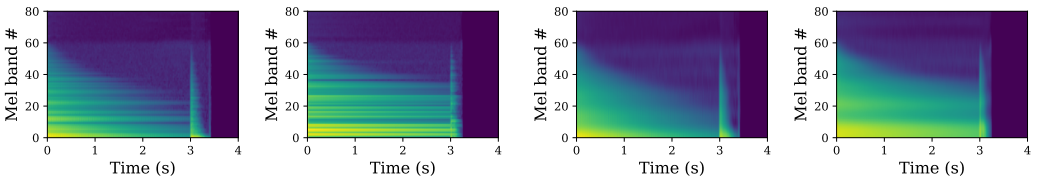

Figure 2: Log mel spectrograms of `guitar_acoustic_010-024-100` (C1, left) and `guitar_acoustic_010-049-100` (C#3, right) (a) without and (b) with liftering. Liftering can allow examples of differing pitch to be more readily compared to one another in the timbral sense.

4 Experimental Results

We train models on the NSynth dataset [3], pruning it according to our specified 88-key pitch range. We resample the dataset, captured at 16 kHz, to a target rate of 44.1 kHz, viewing it as a proxy in lieu of an equally comprehensive full-band alternative. Models are trained to minimize the cross-entropy $\mathcal{L}_{\text{ce}}$ between predicted codes $\hat{\mathbf{c}}$ and ground truth $\mathbf{c}$, over 1M training steps with AdamW optimizer, a batch size of 32, and a cosine-annealed schedule with initial learning rate of 10^{-3} as in [15].

We consider two models to evaluate the fundamental capabilities of the system under investigation. In the first model, instrument timbre is specified from a closed set of instrument family and source type fields available as metadata in the dataset. Our second model is truly a text-to-instrument model, where we leverage CLAP to enable text input at inference. In this case, the instrument family and source type attributes help guide training (subject to dropout with 70% probability), but do not need to be specified at inference. Pitch and velocity are expected as inputs during training and inference across all models in order to generate a complete sample-based instrument. We explicitly note that we did not condition models on more specific instrument metadata because training and evaluation datasets constitute disjoint sets of instruments, and cannot compare to existing text-to-music systems because they are constrained in their specificity of pitch and velocity needed to carry out our task.

We compare models quantitatively (cf. Tab. 1), generating instrument samples reflecting those in the NSynth test set according to a fixed input cue representing each instrument (i.e. a single metadata configuration or CLAP embedding). As is customary [14, 15], we report Fréchet audio distance (FAD) leveraging VGGish and average CLAP score. To measure the efficacy of our pitch conditioning, we estimate the median pitch of generated samples using YIN [22], reporting median absolute deviation in semitones ($\text{MAD}_{\text{pitch}}$) relative to ground truth. Moreover, we measure the average TC across all instruments. Despite potential estimation errors, the detected pitches of our outputs are generally well within a semitone of their targets. The CLAP-based model noticeably improves upon the simple metadata-driven model in terms of CLAP and TC. Its TC metric approaches that of ground truth, with some gap left to close in future works. Interestingly, the metadata-driven model yields marginally lower FAD, which may be due to VGGish not being specifically fine-tuned to musical audio signals.

Lastly, we curate a set of text prompts, generating corresponding instruments using our CLAP-based system. We report average CLAP score correlating generated instruments to their text prompts, $\text{MAD}_{\text{pitch}}$, and TC (cf. Tab. 2). We compile 1-5 scale mean opinion scores (MOS) across members of our organization for quality ($\text{MOS}_{\text{quality}}$), text correspondence ($\text{MOS}_{\text{text}}$), and TC ($\text{MOS}_{\text{TC}}$). We refer readers to our supplementary materials, available at `https://instrumentgen.netlify.app`.

5 Conclusion

In this work, we introduced *text-to-instrument*, and proposed a neural audio codec language model that is catered for the task. We highlighted the fundamental difference between text-to-instrument and other related tasks, whereby the former must generate several samples corresponding to the text prompt that are timbrally consistent to one another. Moreover, we suggested a differentiable objective for measuring the timbral consistency of generated musical instruments. We established a baseline which can generate timbrally consistent sample-based instruments, where we have enabled disentanglement of pitch and timbre via the cross-attention of various conditioning signals incorporated within our system. Future work will consider additional techniques to further improve the fidelity of our system.

Model	FAD↓	CLAP↑	$\text{MAD}_{\text{pitch}}$↓	TC↓
Instrument family/source type	**1.631**	0.487	**0.045**	1.813
CLAP	1.692	**0.691**	**0.045**	**1.236**
Ground truth	–	–	–	**1.053**

Table 1: Evaluation over the NSynth test set.

Model	CLAP↑	$\text{MAD}_{\text{pitch}}$↓	TC↓	$\text{MOS}_{\text{quality}}$ ↑	MOS_{text} ↑	MOS_{TC} ↑
CLAP	0.235	0.162	0.873	3.094	3.620	3.465

Table 2: Evaluation over a curated set of text prompts.

References

[1] G. Narita, J. Shimizu, and T. Akama, "GANStrument: Adversarial Instrument Sound Synthesis with Pitch-Invariant Instance Conditioning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, June 2023.

[2] H. Chang, H. Zhang, J. Barber, A. Maschinot, J. Lezama, L. Jiang, M.-H. Yang, K. P. Murphy, W. T. Freeman, M. Rubinstein, Y. Li, and D. Krishnan, "Muse: Text-To-Image Generation via Masked Generative Transformers," in *Proceedings of the International Conference on Machine Learning*, July 2023.

[3] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proceedings of the International Conference on Machine Learning*, Aug. 2017.

[4] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial Neural Audio Synthesis," in *Proceedings of the International Conference on Learning Representations*, May 2019.

[5] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," in *Proceedings of the International Conference on Learning Representations*, April 2020.

[6] D.-Y. Wu, W.-Y. Hsiao, F.-R. Yang, O. Friedman, W. Jackson, S. Bruzenak, Y.-W. Liu, and Y.-H. Yang, "DDSP-Based Singing Vocoders: A New Subtractive Based Synthesizer and A Comprehensive Evaluation," in *Proceedings of the 23rd Conference of the International Society for Music Information Retrieval*, December 2022.

[7] S. Nercessian, "Differentiable WORLD Synthesizer-Based Neural Vocoder With Application To End-To-End Audio Style Transfer," in *Proceedings of the 154th Audio Engineering Society Convention*, May 2023.

[8] A. Caillon and P. Esling, "RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis," November 2021. arXiv:2111.05011 [cs, eess].

[9] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An End-to-End Neural Audio Codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, November 2021.

[10] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-Fidelity Audio Compression with Improved RVQGAN," June 2023. arXiv:2306.06546 [cs, eess].

[11] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "AudioLM: a Language Modeling Approach to Audio Generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, June 2023.

[12] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers," January 2023. arXiv:2301.02111 [cs, eess].

[13] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually Guided Audio Generation," in *Proceedings of the International Conference on Learning Representations*, May 2023.

[14] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "MusicLM: Generating Music From Text," Jan. 2023. arXiv:2301.11325 [cs, eess].

[15] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and Controllable Music Generation," June 2023. arXiv:2306.05284 [cs, eess].

[16] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, June 2023.

[17] G. Fabbro, V. Golkov, T. Kemp, and D. Cremers, "Speech Synthesis and Control Using Differentiable DSP," October 2020. arXiv:2010.15084 [cs, eess].

[18] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *Transactions on Machine Learning Research*, September 2023.

[19] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2022.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," July 2019. arXiv:1907.11692 [cs].

[21] H. Terasawa, M. Slaney, and J. Berger, "A Statistical Model of Timbre Perception," in *Proceedings of the Statistical and Perceptual Audition Workshop at Interspeech*, September 2006.

[22] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, January 2002.