CONLON: A Pseudo-Song Generator Based on a New Pianoroll, Wasserstein Autoencoders, and Optimal Interpolations

Luca Angioloni[1] Tijn Borghuis[2,3] Lorenzo Brusci[3] Paolo Frasconi[1]

[1] DINFO, Università di Firenze, Italy [2] Eindhoven University of Technology, The Netherlands
[3] Musi-co, Eindhoven, The Netherlands

Introduction

In this paper, we focus on the autonomous generation of polyphonic and multi-instrument MIDI partitures, aiming at producing relatively long *pseudo-songs* (i.e. tracks that have the duration of a song but whose temporal structure is not controlled by a compositional intent), that are effectively *usable* in a professional context.
We do this through:

- the use of a novel pianoroll-like representation PR$^{\mathrm{C}}$.
- the use of Wasserstein autoencoders (WAE).
- the definition of strategies for exploring the WAE latent space.

We introduce two new datasets, especially composed and edited by musicians made aware of creating training sets for generative models: ASF-4 in Acid Jazz, Soul and Funk and 4 instruments; HP-10 in High Pop and 10 instruments.
LPD-5 (5 instruments) was derived from the Lakh MIDI dataset [1] by Dong *et al.*

Novel representation: PR$^{\mathrm{C}}$

Binary pianorolls (PR) are among the most common representations. They are, however, a *lossy* description of MIDI data in at least two ways. First, they do not include note velocities, second, they make it impossible to distinguish between long notes and repeated occurrences of the same notes.
The solution proposed in this paper uses a second channel that explictly represents note durations as continuous variables. Our PR$^{\mathrm{C}}$ description does not suffer the ambiguity between long notes and repeated occurrences of the same note and, except for time quantization, is completely lossless. (See Fig. 1)
In a multi-track context the tracks are stacked together on the channel axis.

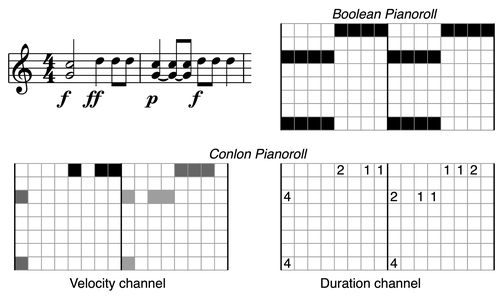

Fig. 1: A short phrase described as PR (top right) and as PR$^{\mathrm{C}}$ (bottom). Here we set quantization at 1/8.

Wasserstein Autoencoder

As a generative model, we experiment with Wasserstein autoencoders (WAE) [2], a type of autoencoder that is less subject to the "blurriness" problem typically associated with variational autoencoders (VAE). To the best of our knowledge, they have not been applied to music generation before.
WAEs penalize a measure of discrepancy $\mathcal{D}$ between the expected $p(x|z)$ and the prior $p(z)$, pushing the expectation inside the distance as in equation:

$$\min_{q(z|x)} \mathbb{E}_p \mathbb{E}_{q(z|x)}\, c(x, G(z)) + \lambda \mathcal{D}(q_z, p_z) \tag{1}$$

where c is a reconstruction loss and λ a hyperparameter to be fixed.
In all our experiments we employed the Maximum Mean Discrepancy (MMD) [3] for $\mathcal{D}$ and a Gaussian prior, and we structured the encoder and the decoder as in the DCGAN [4] architecture, based on 2D convolutional layers.

Generation Strategy

Our generation strategy is formulated as an optimization problem for exploring the autoencoder latent space in a way that prevents abrupt transitions between consecutively generated patterns, as well as regions with little variation.
A pseudo-song is generated by creating a trajectory of length T, $z_1, \ldots, z_T$ in the latent space, and applying the generator model to each latent vector to produce a corresponding sequence of patterns.
We defined 2 strategies:

- Interpolations: where we pick a start pattern and a goal pattern and use the encoder to obtain a start and goal points in the latent space; Trajectories are then computed with a linear or a spherical interpolation.
- Swirls: where latent trajectories are produced by taking real and imaginary parts of periodic complex-valued parametric functions as shown in Fig. 2 on the left.

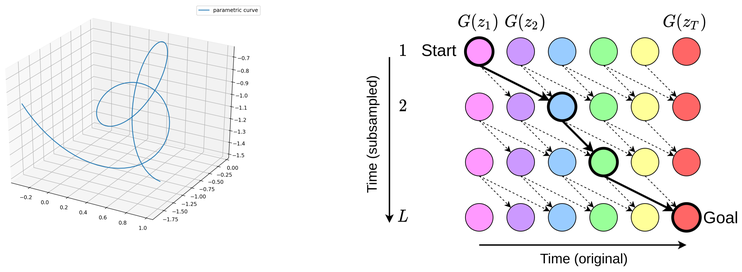

Fig. 2: Latent trajectories produced by *swirls* in a 3D space on the left. On the right, *trellis* for trajectory smoothing. The horizon H is 2 in this example. Among all paths from Start to Goal, the highlighted path is the one whose smallest edge weight is maximum.

Smoothness can be achieved by maximizing the minimum distance between consecutive reconstructions and constraining the final length to L solving the following optimization problem:

$$\max_{t_1,\ldots,t_L} \min_{i=1,\ldots,L-1} \delta(G(z_{t_i}), G(z_{t_{i+1}})) \tag{2}$$

$$\text{s.t. } 1 \le t_i < t_{i+1} \le T \quad i = 1, \ldots L-1 \tag{3}$$

$$t_{i+1} - t_i \le H \quad i = 1, \ldots L-1 \tag{4}$$

where δ is a distance function on patterns and H a lookahead horizon. (See Fig. 2 on the right)

Quantitative Evaluations

When only a limited amount of human expert time is available, it becomes difficult to cover all different dimensions on which alternative methods can be compared. Rather that allowing non experts in our surveys, it may be preferable to complement human evaluation with a number of automatically computed metrics.

Reconstruction Error

We aim to compare WAEs fed by PR vs WAEs fed by PR$^{\mathrm{C}}$. Precision and recall are defined on the binary classification problem where the ground truth consists of Bernoulli variables $y(t, n, i) = 1$ if there is a note-on event at time t for note n and instrument i. We considered as predictions the binary quantities $\widehat{y}(t, n, i) = 1$ if the reconstructed value of the velocity at (t, n, i) is above the smallest velocity in the training set. For PR description, the predicted note-on event was the first element in the merged row of consecutive predictions. We further considered the MAE in predicting velocities and durations. Test set results comparing PR and PR$^{\mathrm{C}}$ (everything else being equal) are reported in Tab. 1.

	ASF-4				HP-10				LPD-5			
	P	R	V	D	P	R	V	D	P	R	V	D
PR	6.1	51.1	32.2	2.5	4.1	58.9	28.8	3.1	1.4	89.7	41.0	5.5
PR$^{\mathrm{C}}$	32.1	53.9	24.3	1.0	37.0	58.0	23.4	1.8	35.5	60.2	19.5	2.6

Tab. 1: Test set precision (P), recall (R), mean absolute errors on velocity (V) and duration (D) for PR and PR$^{\mathrm{C}}$.

Note Shattering

Results in Tab. 1 indicate that PR yields good recall but very low precision, and has a higher error on velocity and duration. This can be partially explained by the presence of a high number of shattered notes. To verify this hypothesis we computed the note number growth due to shattering. For each note in the ground truth, identified by the triplet (n, i, T), being n the pitch, i the instrument, and $T = [t_{\mathrm{ON}}, t_{\mathrm{OFF}}]$ the temporal interval, we counted the number of notes in the reconstruction that have the same pitch n, instrument i and whose note-ON time falls within T. We then summed these counts over all notes in the test set. In the absence of shattering, the total count equals the original number of notes. We found that WAE-PR increased the number of notes by 19%, 12%, 38% on ASF-4, HP-10, and LPD-5, respectively. By comparison, the increase factors were only 5%, 3%, and 10% for WAE-PR$^{\mathrm{C}}$.

Listening Experiments

To validate the CONLON approach, we conducted listening experiments with a group of 69 professional musicians. These experiments showed that they find pseudo-songs generated with WAEs and PR$^{\mathrm{C}}$ descriptions more usable than pseudo-songs generated with other state of the art systems and PR descriptions (see Tab. 2), also they find pseudo-songs generated by WAEs with PR$^{\mathrm{C}}$ descriptions more usable than pseudo-songs generated by the same architecture with PR descriptions (see Tab. 3), and find the development over time of pseudo-songs generated with our system coherent rather than incoherent with respect to Harmony, Rhythm, Melody, and Interplay of instruments (see Tab. 4).

Method	HP-10	LPD-5
CONLON	1.17	1.45
MuseGAN	1.83	1.55
Concordance	0.64	0.01
Significance	$p < 0.0005$	ns

Tab. 2: Mean ranks assigned by subjects to the usability of interpolations generated with our system (CONLON) and MuseGAN [5]. $m = 75$ pairs were ranked.

Description	ASF-4	HP-10	LPD-5
PR$^{\mathrm{C}}$	1.08	1.31	1.5
PR	1.92	1.69	1.5
Concordance	0.72	0.15	0
Significance	$p < 0.0005$	$p < 0.001$	ns

Tab. 3: Mean ranks assigned by subjects to the usability of pseudo-songs generated with PR$^{\mathrm{C}}$ and PR. $m = 78$ pairs were ranked.

	Aspect			
	Harmony	Rhythm	Melody	Interplay
Coherent	49%	67%	42%	51%
Neutral	35%	20%	29%	20%
Incoherent	16%	13%	29%	29%
Significance	$p < .005$	$p < .0005$	$p < .005$	$p < .0005$

Tab. 4: Coherence of CONLON pseudo-songs as judged by subjects, with respect to harmony, rhythm, melody, interplay of instruments. $m = 69$ judgements were collected.

References

[1] Colin Raffel. "Learning-Based Methods for Comparing Sequences, with Applications to Audio-To-Midi Alignment and Matching". PhD thesis. Columbia University, 2016.

[2] Ilya O. Tolstikhin et al. "Wasserstein Auto-Encoders". In: *6th International Conference on Learning Representations*. 2018.

[3] Mikolaj Binkowski et al. "Demystifying MMD GANs". In: *6th International Conference on Learning Representations*. 2018.

[4] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *4th International Conference on Learning Representations*. 2016.

[5] Hao-Wen Dong and Yi-Hsuan Yang. "Convolutional Generative Adversarial Networks with Binary Neurons for Polyphonic Music Generation". In: *Proceedings of the 19th International Society for Music Information Retrieval Conference*. 2018, pp. 190–196.