Mezza *et al.*
EURASIP Journal on Audio, Speech, and Music Processing (2023) 2023:11
https://doi.org/10.1186/s13636-022-00267-2

EURASIP Journal on Audio, Speech, and Music Processing

EMPIRICAL RESEARCH

Open Access

A latent rhythm complexity model for attribute-controlled drum pattern generation

Alessandro Ilic Mezza*, Massimiliano Zanoni and Augusto Sarti

Abstract

Most music listeners have an intuitive understanding of the notion of rhythm complexity. Musicologists and scientists, however, have long sought objective ways to measure and model such a distinctively perceptual attribute of music. Whereas previous research has mainly focused on monophonic patterns, this article presents a novel perceptually-informed rhythm complexity measure specifically designed for polyphonic rhythms, i.e., patterns in which multiple simultaneous voices cooperate toward creating a coherent musical phrase. We focus on drum rhythms relating to the Western musical tradition and validate the proposed measure through a perceptual test where users were asked to rate the complexity of real-life drumming performances. Hence, we propose a latent vector model for rhythm complexity based on a recurrent variational autoencoder tasked with learning the complexity of input samples and embedding it along one latent dimension. Aided by an auxiliary adversarial loss term promoting disentanglement, this effectively regularizes the latent space, thus enabling explicit control over the complexity of newly generated patterns. Trained on a large corpus of MIDI files of polyphonic drum recordings, the proposed method proved capable of generating coherent and realistic samples at the desired complexity value. In our experiments, output and target complexities show a high correlation, and the latent space appears interpretable and continuously navigable. On the one hand, this model can readily contribute to a wide range of creative applications, including, for instance, assisted music composition and automatic music generation. On the other hand, it brings us one step closer toward achieving the ambitious goal of equipping machines with a human-like understanding of perceptual features of music.

Keywords Rhythm complexity, Variational autoencoders, Latent space regularization, Drums

1 Introduction

Researchers in psychology, neuroscience, musicology, and engineering have long tried to find quantitative mathematical models of perceptual attributes of music. However, human perception can hardly be systematized into a fixed set of disjoint categories. In fact, music similarity, expressiveness, emotion, and genre, to name a few, are elusive terms that defy a shared and unambiguous definition [1–3]. Furthermore, seeking a general consensus might be considered an ill-defined problem, as these aspects of music fruition are distinctively subjective and are strongly dependent on one's personal experience, music education, background, and culture [4–6]. Our perception of a musical performance depends on the complex interaction between multiple interrelated conceptual layers, and not a single aspect can be gauged in a vacuum: melody, harmony, rhythm, loudness, timbre, time signature, and tempo all play a joint role in the holistic perception of music and may affect how a musical piece is experienced [7]. Nevertheless, while an all-encompassing model for music perception may seem a long way off, researchers have compellingly resorted to a

*Correspondence:
Alessandro Ilic Mezza
alessandroilic.mezza@polimi.it
Department of Electronics, Information and Bioengineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

divide-and-conquer strategy to address such a multifaceted problem.

This work focuses on the long-studied aspect of rhythm complexity. Over the past few decades, numerous mathematical models have been proposed in the literature [8–17], all aimed at assessing the complexity of a pattern of rhythmic events. These algorithms ultimately provide an explicit mapping from a symbolic representation of the rhythm to a scalar value meant to quantify the degree of complexity that a human listener would perceive. These methods, however, are only able to partially model rhythm complexity. Indeed, while most listeners are able to assess the complexity of a given musical piece, an experienced musician writing or performing music can make use of a controlled degree of complexity to great artistic effect.

Data-driven methods have recently proven to be a powerful and expressive tool for multimedia generation. For example, deep learning techniques have been successfully applied to image [18–20], text [21–23], speech [24–26], and music generation [27–32]. However, the mappings provided by deep generative models for musical applications are typically implicit and lack interpetability of the underlying generative factors. In fact, they often fall short on two crucial aspects: controllability and interactivity [33]. Nonetheless, several attribute-controlled methods have been recently proposed in the literature. The hierarchical architecture of MusicVAE [29], other than achieving state-of-the-art performance in modeling long-term musical sequences, enables latent vector arithmetic manipulation to produce new samples with the desired characteristics. Following the work of [29], Roberts et al. [28] explored possible interactive applications of latent morphing via the interpolation of up to four melodies or drum excerpts. In [31], Gillick et al. introduced GrooVAE, a seq2seq recurrent variational information bottleneck model capable of generating expressive drums performances. The model, trained using Groove MIDI Dataset, was designed to tackle several drum-related tasks, including humanization, groove transfer, infilling, and translating tapping into to full drum patterns. Furthermore, Engel et al. [34] showed that it is possible to learn a-posteriori latent constraints that enable the use of unconditional models to generate outputs with the desired attributes. Hadjeres et al. [35] proposed a novel geodesic latent space regularization to control continuous or discrete attributes, such as the number of musical notes to be played, and applied it to the monophonic soprano parts of J.S. Bach chorales. In [36], Brunner et al. presented a recurrent variational autoencoder complemented with a softmax classifier that predicts the *music style* from the latent encoding of input symbolic representations extracted from MIDI; the authors thus performed style transfer between two music sequences by swapping the style codes. Tan and Herremans [37] took inspiration from Fader Networks [38] and proposed a model that allows to continuously manipulate music attributes (such as *arousal*) by independent "sliding faders." This was achieved by learning separate latent spaces from which high-level attributes may be inferred from low-level representations via Gaussian Mixture VAEs [39]. More recently, Pati and Lerch [40] proposed a simple regularization method that monotonically embeds perceptual attributes of monophonic melodies, including rhythm complexity, in the latent space of a variational autoencoder. The authors have later investigated the impact of different latent space disentanglement methods on the music generation process of controllable models [41]. Finally, it is worth mentioning that recent commercial products, such as Apple's Logic Pro Drummer, offer some degree of control over the rhythm complexity of an automated polyphonic drumming performance.

Against the backdrop of such a rich literature corpus, the contribution of this article is twofold. First, we propose a novel complexity measure that is specifically designed for drum patterns belonging to the Western musical tradition. To the best of our knowledge, this constitutes the first attempt at designing a proper polyphonic rhythm complexity measure. We validate the proposed algorithm via a perceptual experiment conducted with human listeners and show a high degree of agreement between measured complexity and subjective evaluations. Second, we present a latent vector model capable of learning a compact representation of drum patterns that enables fine-grained and explicit control over perceptual attributes of the generated rhythms. Specifically, we encode the newly proposed complexity measure in the latent space of a recurrent variational autoencoder inspired by [29, 31] and modified to enable single-knob manipulation of the target attribute. The resulting model can generate new and realistic drum patterns at the desired degree of complexity and provides an interpretable and fully-navigable latent representation that appears topologically structured according to the chosen rhythm complexity measure.

The remainder of this article is organized as follows. In Section 2, we provide an overview of the relevant literature and existing techniques for measuring the rhythm complexity of monophonic patterns. In Section 3, we propose a novel polyphonic complexity measure. In Section 4, we describe the dataset of drum patterns utilized in the present study. In Section 5, we provide the details of the listening test conducted to validate the proposed complexity measure and present the results. In Section 6, we outline the proposed latent rhythm complexity model.

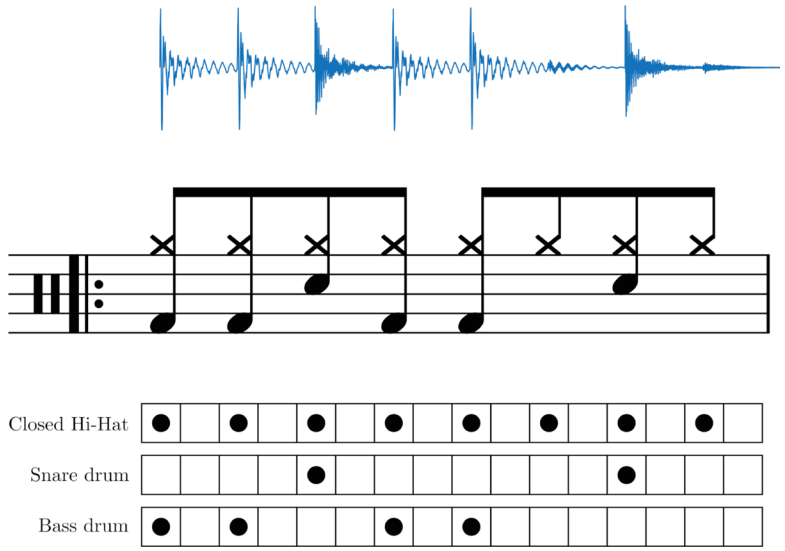

Fig. 1 Various representations of the same polyphonic drum pattern: raw audio waveform (top), drum sheet music (center), and symbolic binary representation (bottom)

In Section 7, we evaluate the performance of the proposed model on the tasks of attribute-controlled drum pattern generation and output complexity manipulation. Finally, Section 8 concludes this work.

2 Background on rhythm complexity

Over the years, several rhythm complexity measures have been proposed in the literature [42–44]. Rather than considering the music signal as a raw waveform, most of the existing methods relies on a intermediate symbolic representation of rhythm as that produced by an ideal onset detector [45]. For a given tatum[1], it is customary to derive a discrete-time binary sequence of *onsets* distributed across a finite number of *pulses*. On this grid, ones correspond to onsets and zeros correspond to silence, as depicted in Fig. 1. A pulse refers to the smallest metrical unit meaningfully subdividing of the main beat and represent one of all the possible discrete-time locations within a binary pattern that can be assigned either one or zero. Therefore, a rhythm complexity measure can be thought of as a (nonlinear) function $f_{\mathrm{p}} : \{0,1\}^{M\times N} \to \mathbb{R}$ that, given the matrix representation of a polyphonic rhythm with N pulses and M voices, yields a real-valued scalar.

Many rhythm complexity measures have been based on the concept of *syncopation* [8, 13–15, 47] i.e., the placement of accents and stresses meant to disrupt the regular flow of rhythm. Others, such as [10, 12, 48], are measures of *irregularity* with respect to a uniform meter, and several rely on the statistical properties of inter-onset intervals [9, 16, 42, 49]. Moreover, some authors have investigated entropy [50], subpattern dependencies [11], predictive coding [17], and the amount of data compression achievable [12, 51] in order to quantify the complexity of a rhythmic sequence. However, to the best of our knowledge, previous work almost entirely concerns monophonic patterns ($M = 1$) and not polyphonic rhythms ($M > 1$).

Notably, [52, 53] explore the adaptation of Toussaint's metrical [13] and Longuet-Higgins and Lee [8] monophonic complexity measures to the polyphonic case, respectively. In [52], rhythm complexity estimates are used to rank MIDI file in a database. In [53], complexity measures are used to drive an interactive music system. Crucially, however, [52, 53] consider each drum-kit voice independently of the others before pooling the results, thus disregarding the interaction between voices. Furthermore, the authors provide no validation of the proposed methods against the results of a subjective evaluation campaign conducted with human listeners.

3 A novel rhythm complexity measure for polyphonic drum patterns

Drawing from the rich literature discussed in Section 2, the simplest design for a proper polyphonic complexity measure f_{p} would first entail computing the complexity of each voice x_m in a M-voices pattern $\mathbf{x} = [x_1[n], \ldots, x_M[n]]^\top$ separately from one another by using one of the many state-of-the-art monophonic rhythm complexity measures $f(\cdot)$. Then, the overall complexity can be obtained as the linear combination

[1] In this context, *tatum* refers to "the smallest time interval between successive notes in a rhythmic phrase" [46].

$$f_{\mathrm{p}}(\mathbf{x}) := \sum_{m=1}^{M} w_m f(x_m[n]). \tag{1}$$

However, such a naive approach is bound to provide a poor complexity model as it does not take into consideration the interplay between voices that are instead meant to complement each other.

Instead, we propose to compute the linear combination of the monophonic complexity of *groups of voices* selected from those that are often found to create interlocked rhythmic phrases in the drumming style typical of contemporary Western music. Indeed, our assumption is that grouping multiple voices together allows to better capture the perceptual rhythm complexity of polyphonic patterns, as it is arguably determined by the joint interaction of multiple sources that play a certain role only in relation to others.

Given a subset of binary voices $x_1[n], \ldots, x_L[n]$ out of the M voices in $\mathbf{x} \in \{0,1\}^{M \times N}$, the kth group can be defined as

$$g_k[n] := \bigvee_{\ell=1}^{L} x_\ell[n] \tag{2}$$

Namely, $g_k[n] = 1$ if and only if at least one of the L grouped voices had an onset at pulse n. Otherwise, $g_k[n] = 0$. Applying (2) to all K groups, the given pattern $\mathbf{x}$ yields an augmented matrix representation $\mathbf{g} = [g_1[n], \ldots, g_K[n]]^\top$ of size $K \times N$, where possibly $K \gg M$. Hence, the proposed polyphonic complexity measure is given by

$$f_{\mathrm{p}}(\mathbf{g}) := \sum_{k=1}^{K} w_k f(g_k[n]) \tag{3}$$

where the weights w_k can be either, e.g, set to $1/K$ (yielding a simple average) or determined via (possibly non-negative) linear regression against the subjective results of a large-scale listening test.

We empirically determine the grouping reported in Table 1. Most notably, bass and snare drums are merged into a single group ($k = 1$). Together, they constitute the backbone of contemporary Western drumming practices, especially in the rock and pop genre. Therefore, their relationship cannot be wholly conceptualized if they are considered disjointedly. Likewise, high and low toms are merged into group $k = 6$. For their part, closed and open hi-hat appear by themselves in respective groups ($k = 2$ and $k = 3$), and we include an auxiliary group ($k = 4$) to account for those rhythms in which the hi-hat follow a regular pattern regardless of the pedal action. For instance, let us consider a 1-bar pattern where open and closed hi-hat alternate as depicted in Fig. 2. The closed hi-hat ($k = 2$) is always off-beat and thus is likely be assigned high complexity. However, the joint rhythm consists of a regular sequence of semiquavers thus making up for a rather easy-to-conceptualize rhythm. Similarly, an auxiliary group ($k = 9$) is introduced for crash and ride cymbals, as the former is often used to accent patterns played mostly on the latter. Ultimately, by measuring the complexity of joint patterns and individual voices, we expect to regularize the overall complexity estimate accounting for both regularity and novelty.

Table 1 Proposed drum-kit voice groups $g_k[n]$ used for the computation of rhythm complexity as given in (3)

Group	Voices
$k = 1$	Bass drum (0); Snare drum (1)
$k = 2$	Closed Hi-Hat (2)
$k = 3$	Open Hi-Hat (3)
$k = 4$	Closed Hi-Hat (2); Open Hi-Hat (3)
$k = 5$	High Floor Tom (4)
$k = 6$	Low-Mid Tom (5); High Tom (6)
$k = 7$	Ride Cymbal (7)
$k = 8$	Crash Cymbal (8)
$k = 9$	Ride Cymbal (7); Crash Cymbal (8)

In this study, in order to quantify the complexity of each group, we adopt Toussaint's metrical complexity measure [13]. For completeness, a detailed presentation of [13] is given in the Appendix. However, the proposed method does not intrinsically rely on any particular choice of $f(\cdot)$, and a different monophonic measure may be used for each group of voices independently of the others.

4 Groove MIDI Dataset

Groove MIDI Dataset (GMD) was released by the authors of [31] and contains 13.6 h of drums recordings performed by professional and amateur drummers on an electronic drum set. The dataset contains audio files, MIDI transcriptions, and metadata, including time signature and tempo expressed in ticks per quarter. Whereas the original recordings are of variable lengths, we limit our study to 2-bar scores. GMD comprises a total of 22619 2-bar scores, 97% of them being of time signature 4/4. Filtering out other time signatures yields a total of 21940 samples. The General MIDI standard for drum-kits provides an integer number between 1 and 255 corresponding to each drum instrument. We apply the reduction strategy proposed in [29, 31] to map the 22 drum classes included in GMD onto nine canonical voices: Bass drum (0), Snare drum (1), Closed Hi-Hat (2), Open Hi-Hat (3), High Floor Tom (4), Low-Mid Tom (5),

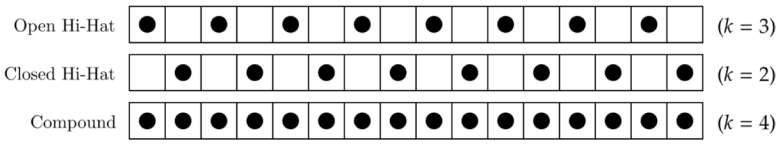

Fig. 2 Exemplary 1-bar hi-hat pattern. Taken by itself, the Closed Hi-Hat group ($k = 2$) is characterized by high syncopation, all onsets being off-beat. Functionally, however, the combined pattern played on the hi-hat ($k = 4$) is likely to be perceived as steady and regular

High Tom (6), Ride Cymbal (7), Crash Cymbal (8). We assume a tatum corresponding to a sixteenth note, regardless of tempo. This yields a total of 32 pulses every two bars. Thus, we quantize every MIDI event to the closest pulse in order to produce a symbolic representation in the form of a real-valued matrix with $M = 9$ rows and $N = 32$ columns. Finally, we obtain a binary matrix by discarding the information regarding the velocity of each onset and replacing with ones all rhythmic events with non-zero velocity.

5 Perceptual evaluation

5.1 Experimental setup

The polyphonic complexity measure proposed in Section 3 is validated via a listening test involving eight MIDI patterns sampled from GMD. The evaluation corpus was obtained by synthesizing audio clips from MIDI files in order to control the quantization and velocity of the test patterns. In fact, the audio recordings included in GMD contain agogic and dynamic accents which are traditionally excluded from the evaluation of rhythm complexity. First, we quantized every onset to the nearest semiquaver, and the corresponding velocity values were all set to 80. Then, the resulting patterns were repeated four times to create 8-bar sequences, synthesized to wav files using a library of realistic drum samples distributed with Ableton Live 9 Lite, and finally presented to human listeners via an online form. Akin to the five-point category-judgment scales of the Absolute Category Rating method included in the ITU-T Recommendation P.808 [54], testers were asked to provide a subjective assessment of the perceived rhythm complexity on a scale of 1 (lowest complexity) to 5 (highest complexity). The eight test samples were selected as follows. First, we filtered all three folds of GMD to gather a pool of candidate patterns. Specifically, we discarded all rhythms having either less than three voices (to make sure to evaluate proper polyphonic patterns) or less than eight pulses where at least one onset is present (to exclude overly sparse temporal sequences). Then, in order to ensure an even representation across the whole range of complexity values, we evaluated (3) for every candidate pattern using uniform weights $w_k = 1$, $k = 1, \ldots, K$. We selected eight uniformly spaced target complexity values by sampling the range between the minimum and maximum complexity thus obtained. Hence, we extracted the eight drum patterns whose complexities were closest to the target ones. During the test, the order in which the patters were presented to the user was randomized and the name of each file replaced with a string of random characters. In order to minimize experimenter-expectancy effect, no audio examples were provided to the subjects before the test. Indeed, manually selecting a number of clips that aligned with the authors' a priori notion of rhythm complexity could have possibly led to confirmation bias. Instead, we opted for an experimental setup in which all test clips were presented in the same web page, and users were allowed to listen to all patterns and possibly modify previous assessments before submitting the final evaluation results. A total of 24 people took part in the experiment, mainly from a pool of university students and researchers from the Music and Acoustic Engineering program at Politecnico di Milano, Italy. All test subjects are thus expected to have some degree of familiarity with basic music theory concepts.

5.2 Results

Figure 3 shows the correlation between the proposed rhythm complexity measure and the scores attributed to each pattern by the test subjects. Blue circles represent the average perceptual complexity assigned by the users to each of the eight drum patterns. Blue vertical lines, instead, represent the standard deviation for each given sample. The red dashed line represents the linear regression model with complexity measures as covariates and average subjective assessments as dependent variables. Using a uniform weighting policy for all voice groups, the data show a Pearson linear correlation coefficient of 0.9541 and, correspondingly, a Spearman rank correlation coefficient of 0.9762, indicating a strong monotonic relationship. Furthermore, the simple linear model $y = 0.034 f_{\mathrm{p}}(\mathbf{g}) + 1.35$ can fit the average user scores with a coefficient of determination of $R^2 \approx 0.91$.

As previously mentioned in Section 3, perceptually informed group weights $w_1, \ldots, w_K$ may be determined from the collected subjective assessments. Albeit the small sample size involved in the present experiment does not allow for a robust linear regression and it is likely to lead to overfitting, we empirically found that

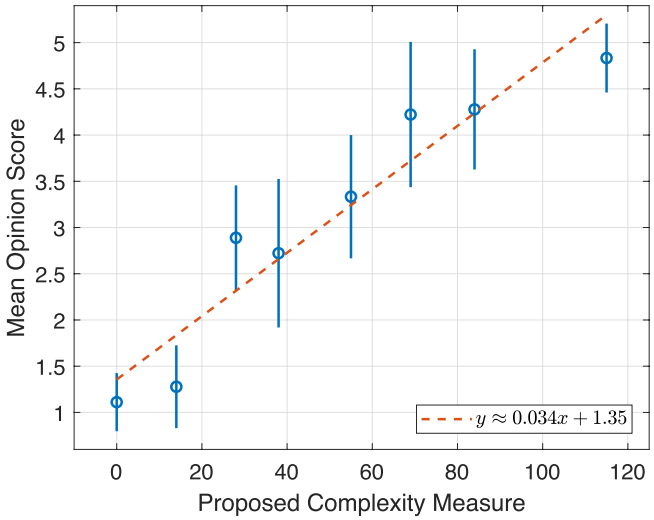

Fig. 3 Correlation between the proposed rhythm complexity measure with uniform weights $w_k = 1$, $\forall k$, and the average score assigned by human listeners to each of the eight drum patterns

setting $w_1 = 3$ for the bass and snare drum group ($k = 1$) and $w_4 = w_9 = 1/3$ for the compound hi-hat ($k = 4$) and cymbals ($k = 9$) groups leads to a Pearson coefficient of 0.983 corresponding to a linear model $y = 0.033 f'_{\mathrm{p}}(\mathbf{g}) + 0.89$ with $R^2 \approx 0.97$.

The results presented in this section indicate a high degree of agreement between subjective assessments and the proposed complexity measure. Going forward, however, additional tests on curated datasets may be needed to confirm the applicability of the voice groups identified in Section 3 to genres such as jazz and heavy metal that are characterized by peculiar drumming techniques. These experiments are left for future work.

6 Proposed attribute-controlled drum patterns generation model

6.1 Deep generative architecture

In this section, we present a new attribute-controlled generative model that enables fine-grained modeling of musical sequences conditioned on high-level features such as rhythm complexity. The deep generative model is based on the hierarchical recurrent β-VAE architecture of MusicVAE [29], and it is augmented with two auxiliary loss terms meant to regularize and disentangle the latent space, respectively.

As in [31], the recurrent encoder $q_\phi(\mathbf{z}|\mathbf{x})$ comprises a stack of two bidirectional layers, each with 512 LSTM cells. The forward and backward hidden states obtained by processing an input sequence $\mathbf{x} \in \{0,1\}^{M \times N}$ are concatenated into a single 1024-dimensional vector, before being fed to two parallel fully-connected layers. The first layer outputs the locations of the latent distribution $\boldsymbol{\mu} \in \mathbb{R}^H$, where $H = 256$. The second layer, equipped with a softplus activation function, yields the scale parameters $\boldsymbol{\sigma} \in \mathbb{R}^H_{\geq 0}$.

We implement a hierarchical LSTM decoder $p_\theta(\mathbf{x}|\mathbf{z})$ composed of a high-level conductor network and a bottom-layer RNN decoder. Namely, both the conductor and the decoder are two-layer unidirectional LSTM networks with 256 hidden cells and tanh activations. The output layer of the conductor has size 128 and that of the decoder has M sigmoid units, as many as the number of drum voices.

The input sequence $\mathbf{x}$ is split into $S = 8$ non-overlapping sections of size $M \times N/S$. The conductor network, whose goal is to model the long-term character of the entire sequence, outputs S embedding vectors which are in turn used to initialize the hidden states of the lower-level decoder.

The latent code $\mathbf{z} \in \mathbb{R}^H$ is randomly sampled from a multivariate Gaussian distribution $p(\mathbf{z})$ parameterized by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. Then, it is passed through a fully-connected layer followed by a tanh activation function to compute the initial states of the conductor network. For each of the S segments, the 128-dimensional embedding vector yielded by the conductor is in turn passed through a shared fully connected layer to initialize the hidden states of the lower-level decoder. The concatenation between the previous output and the current embedding vector serves as input for the decoder to produce the next section. The decoder autoregressively generates S sections that are thus concatenated into the complete output sequence.

As commonly done for β-VAEs, the base model is trained by minimizing the following objective [55]

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}\big[\log p_\theta(\mathbf{x}|\mathbf{z})\big] + \beta D_{\text{KL}}\big(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z})\big), \quad (4)$$

where $D_{\text{KL}}(\cdot\,||\,\cdot)$ denotes the Kullback-Leibler divergence (KLD) and the real-valued parameter $\beta < 1$ favors reconstruction quality over enforcing a standard normal distribution in the latent space [21].

6.2 Latent space regularization

The deep generative model described in Section 6.1, despite having proven able to produce coherent long-term musical sequences, is unaware of the perceptual aspects of target rhythms. To incorporate this information into our latent vector model, similarly to [36, 40, 56], we propose a multi-objective learning approach. Namely, we force the base model to jointly learn the rhythm complexity of input patterns along with minimizing the classic β-VAE loss function given in (4). Our goal is to regularize the latent space in a way that would allow for continuous navigation and semantic exploration of the learned model. This is achieved by including the following auxiliary loss function

$$\mathcal{L}_{\text{reg}} = \text{MSE}(f_{\text{p}}(\mathbf{g}), z_i), \quad (5)$$

where $f_{\text{p}}(\mathbf{g})$ is a polyphonic rhythm complexity measure such as the one described in Section 3, and $z_i \in \mathbb{R}$ is ith element of the latent code $\mathbf{z}$. This way, we are effectively constraining the ith latent space dimension to become topologically structured according to the behavior of the target perceptual measure. Hence, sampling latent codes along such dimension allows for the explicitly manipulation the complexity of output patterns in a way close to human understanding.

Since the latent vectors are encouraged to follow a multivariate standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ by the KLD term in (4), we standardize the complexity values of training data by subtracting the sample mean and dividing by the standard deviation. This yields the zero-mean and unit-variance complexity distribution shown in Fig. 4, which is compatible with being encoded into the ith univariate component of $\mathbf{z}$. Finally, we use the training data statistics thus obtained to apply the same standardization at inference time.

6.3 Latent space disentanglement

Having regularized z_i as described in the previous section, there are yet no guarantees that some information regarding rhythm complexity had been incorporated into other latent space dimensions. In fact, rhythm complexity is typically measured by gauging onset locations, which ultimately carry most of the same information the base model is trying to encode in the H latent dimensions. In particular, we would like the remaining $H - 1$ dimensions of the latent space to be relatively invariant with respect to changes in input complexity. Indeed, explicit and interpretable control over the desired output behavior becomes unfeasible when multiple latent variables are redundant and affect the same aspects of the overall rhythm complexity model. In the context of feature learning for generative applications, such a desirable property is often referred to as latent space *disentanglement* [57].

Notably, β-VAE was originally introduced to favor disentanglement [55]. However, this is mainly achieved for large values of β, as later observed in [58]. Therefore,

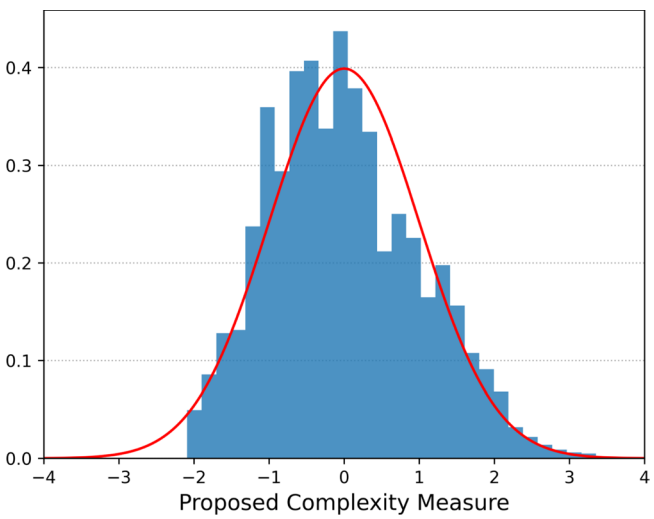

Fig. 4 Histogram of the proposed rhythm complexity measure evaluated on all 2-bar training patterns in GMD. Complexity values are standardized to obtain a distribution with zero mean and unit variance. For comparison, the probability density function of a standard normal distribution $\mathcal{N}(0, 1)$ is overlaid in red

inspired by prior work on predictability minimization [59, 60] and attribute manipulation by means of sliding faders [37, 38, 40], we propose to augment the base model with an auxiliary adversarial loss term promoting latent space disentanglement. Let $\mathbf{z}_\star \in \mathbb{R}^{H-1}$ be the vector of all remaining latent variables in $\mathbf{z}$ except for z_i. We define an adversarial regressor $\hat{f}_\mathrm{p}(\mathbf{z}_\star)$ that is tasked to estimate the input rhythm complexity $f_\mathrm{p}(\mathbf{g})$ by minimizing the following loss functions

$$\mathcal{L}_\mathrm{adv} = \mathrm{MSE}(f_\mathrm{p}(\mathbf{g}), \hat{f}_\mathrm{p}(\mathbf{z}_\star)). \tag{6}$$

In order to reduce the amount of information regarding $f_\mathrm{p}(\mathbf{g})$ that is embedded into $\mathbf{z}_\star$, we connect the encoder and the regressor via a gradient reversal layer (GRL) [60] that flips the sign of the gradients during backpropagation. Therefore, the encoder will learn a latent representation $\mathbf{z}_\star$ that is minimally sensitive with respect to the input complexity as it is now trained adversarially with respect to the regressor.

In this study, we implement the adversarial regressor as a feed-forward neural network with two hidden layers with 128 units followed by ReLU activations and a linear output layer yielding a scalar value. The block diagram of the complete model is depicted in Fig. 5.

6.4 Model training

The proposed latent vector model is optimized for 300 epochs using Adam [61], a batch size of 128, and the following compound objective function

$$\mathcal{L} := \mathcal{L}_\mathrm{VAE} + \alpha\, \mathcal{L}_\mathrm{reg} + \gamma\, \mathcal{L}_\mathrm{adv}, \tag{7}$$

where α and γ are scalar weights for the attribute-regularization and adversarial terms, respectively.

The learning rate is set to 10^{-3} and exponentially decreased to 10^{-5} with a decay rate of 0.99. We set the regularization weight $\alpha = 1$ for the entire training. Conversely, β and γ are annealed during early training to let the model focus more on pattern reconstruction then on structuring the latent representation. Namely, we set $\beta = 10^{-4}$ and $\gamma = 10^{-6}$ for the first 40 epochs. Throughout the following 250 epochs, β is linearly increased up to 0.25 with a step of 10^{-3} per epoch, and γ is increased up to 0.05 with a step of $2 \cdot 10^{-4}$. Furthermore, we randomly apply teacher forcing on the recurrent decoder with a probability of 50%.

7 Performance evaluation

In this section, we evaluate the proposed latent vector model on several attribute-controlled generation tasks. In Section 7.1, we discuss the effects of latent space regularization. In Section 7.2, we show that the proposed adversarial component is effectively disentangling the latent representation. In Section 7.3, we investigate the capability of the proposed method to alter the rhythm complexity of input patterns in a controlled way. Finally, in Section 7.4, we test the model on the task of attribute-controlled generation from randomly sampled latent vectors.

7.1 Latent space regularization

In Fig. 6, we depict the latent vectors obtained by encoding GMD test data. For the sake of visualization, we only plot two latent dimensions, i.e., z_i and z_l. In this example, we regularize the first dimension, i.e., $i = 0$, and choose $l = 127$. The color assigned to each point represents the rhythm complexity measured on the respective input patterns using (3): brighter colors correspond to higher complexity. The clearly noticeable color gradient indicates that the complexity values have been monotonically encoded along the regularized dimension. Furthermore, the latent complexity distribution appears to be continuously navigable from low to high values by traversing the latent space toward the positive direction of z_0.

In Fig. 7, instead, the coloring is determined according to the rhythm complexity measured on the output patterns generated by the decoder. Whereas the variational

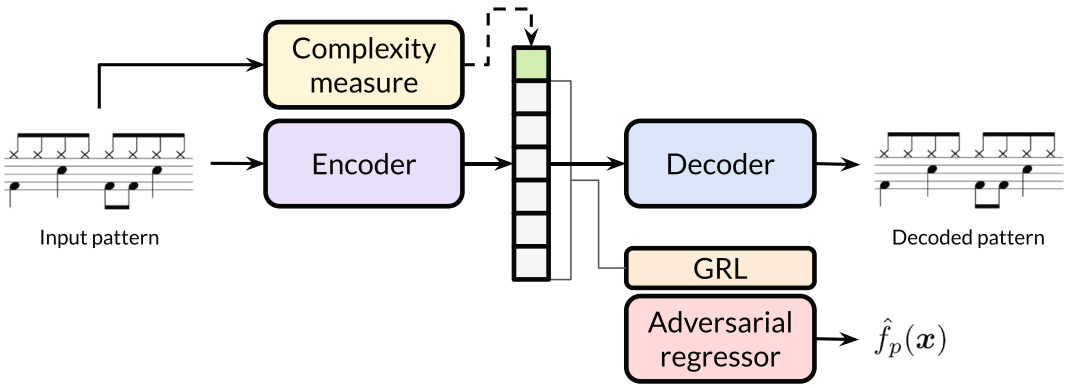

Fig. 5 Proposed latent vector model for attribute-controlled drum pattern generation

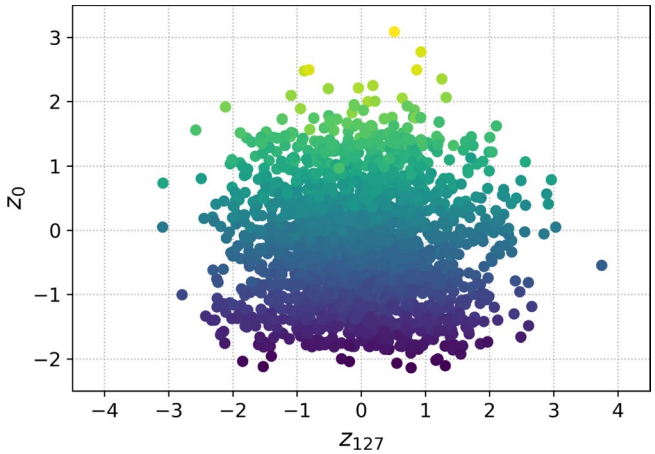

Fig. 6 Latent rhythm complexity distribution of input drum patterns

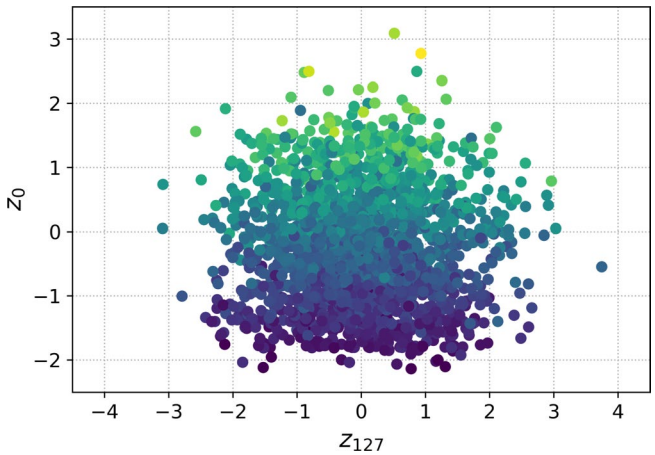

Fig. 7 Latent rhythm complexity distribution of generated drum patterns

decoding process appears to affect output measures, we may notice that the overall complexity distribution retain a high degree of agreement with that of the input data.

7.2 Latent space disentanglement

The adversarial component introduced in Section 6.3 is meant to penalize any leakage of information regarding rhythm complexity into the non-regularized latent space dimensions. To assess the effectiveness of the proposed method, we conduct a simple ablation study. We train two generative models: the first is implemented as described in Section 6; the second follows the same specifications except for the exclusion of the adversarial term from the loss function in (7).

Quantifying latent space entanglement is a challenging task, as unwanted redundancy and the intertwining of latent variables might not follow a simple and easy-to-identify behavior. Therefore, similarly to what proposed in [62], our approach was to measure latent space entanglement via nonlinear regression. We define a nonlinear regressor $r(\cdot)$ meant to estimate the measured complexity $f_{\mathrm{p}}(\mathbf{g})$ from the non-regularized latent code portion $\mathbf{z}_\star$. These codes are obtained by passing GMD data through the two pre-trained generative models under consideration. For the sake of simplicity, let us denote with $\mathbf{z}_\star$ the partial codes from the proposed model and with $\tilde{\mathbf{z}}_\star$ the ones from the baseline without adversarial term.

We implement each regressor as a two-layer feed-forward neural network with 128 units and ReLU activations. The two networks are optimized using $\mathbf{z}_\star$ and $\tilde{\mathbf{z}}_\star$ extracted from training data. The regressors are thus evaluated on the test fold of GMD. We argue that a lower regression performance corresponds to a more disentangled latent representation.

The regressor $r(\tilde{\mathbf{z}}_\star)$ achieves a coefficient of determination $R^2 = 0.5$, quantifying the percentage of the variation

in test data complexity that is predictable from the independent variables $\tilde{\mathbf{z}}_\star$. This clearly suggests that, without the proposed adversarial component, a non-negligible amount of information regarding rhythm complexity leaked into the non-regularized latent space and can be thus predicted. Conversely, the coefficient of determination of $r(\mathbf{z}_\star)$ drops to $R^2 = 0.1$ when including the adversarial loss term, revealing that the latent space has been effectively disentangled. Ultimately, this enables intuitive control over the output complexity that can be now altered in a *fader*-like fashion [38] simply by varying the scalar value of z_i.

7.3 Rhythm complexity manipulation

In this section, we demonstrate how the proposed model could allow for a fine-grained manipulation of target attributes of the generated samples. In particular, we encode each rhythm $\mathbf{x}$ in the test fold of GMD and extract the corresponding latent vectors $\mathbf{z}$. Then, we fix $\mathbf{z}_\star$ and let z_i vary according to $z_i + j\Delta z$, where $j \in \mathbb{Z}$ and $\Delta z = 0.5$. For each new latent code obtained this way, we task the decoder to generate the corresponding pattern. Hence, we compare the difference between the complexity of the unaltered output and that of the newly generated ones. Figure 8 shows the violin plot for $j \in [-5, 5]$, depicting for each shift $j\Delta z$ the distribution of the resulting changes in the complexity of the decoder output for all samples in the test set. Remarkably, we obtain a Pearson correlation coefficient of 0.90 between the desired and resulting complexity increments.

By keeping $\mathbf{z}_\star$ fixed throughout the experiment, we argue that the generated rhythms would be most similar to the original one. However, the more the target complexity is altered, the greater will be the deviation from the original pattern. To support these claims, we compute the average Hamming distance $\mathcal{H}(s_0, s_j)$ between the unaltered output pattern ($j = 0$) and the ones generated with the desired complexity increment $j\Delta z$. Namely, we convert each drum voice into a string of ones and zeros and measure the number single-character edits needed to change one pattern into the other. Arguably, a higher Hamming distance indicates a more significant modification of the original output pattern. In Fig. 9, we show $\mathcal{H}(s_0, s_j)$ as a function of $j\Delta z$ with $\Delta z = 0.1$. Notably, the average distance monotonically increases as the target complexity increment moves away from zero. In fact, the patterns with the least amount of complexity manipulation appear to be the most similar to the reference rhythm with an average of approximately 7.4 edits per sample. Conversely, the maximum distance is achieved for $j\Delta z = 2.5$, where we observe an average of 21.3 edits per sample.

7.4 Attribute-controlled generation

Finally, we evaluate the proposed latent vector model in a purely generative mode. We sample 1000 random latent codes from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and task the decoder to autonomously produce new patterns. We let z_0 vary from -2.576 to 2.576, thus accounting for the complexity of 99% of the samples in the GMD test fold. Hence, we compute the correlation between z_0 and the complexity of the newly generated patterns. Figure 10 shows a clear linear relationship between desired and output rhythm complexity. Despite a Pearson correlation coefficient of 0.9163, however, we notice the tendency of the system to reduce the output complexity with respect the target z_0 value. This, in turn, is confirmed by the slope of the best-fit linear regression model $y \approx 0.78\, z_0 - 0.18$ being less than one. Moreover, this trend is accompanied by an increment in the output complexity variance as z_0 increases. These phenomena might be explained by considering that the training fold of GMD consists of data from spontaneous drumming performances and offers a limited representation of high-complexity patterns. As a result, the decoder may have been biased toward generating more

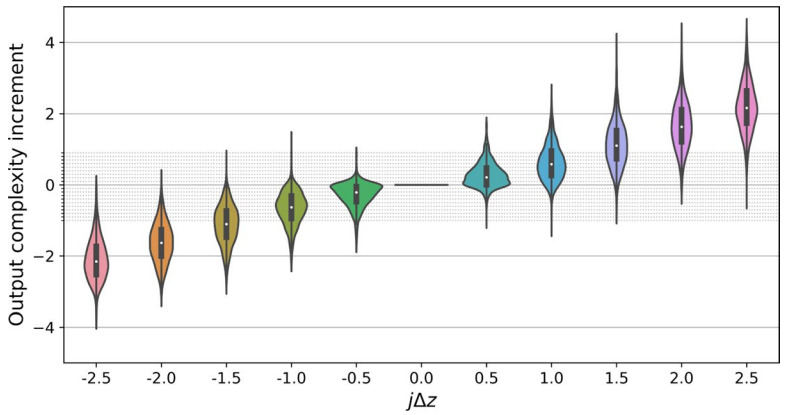

Fig. 8 Violin plot of the measured output complexity increments as a function of $j\Delta z$

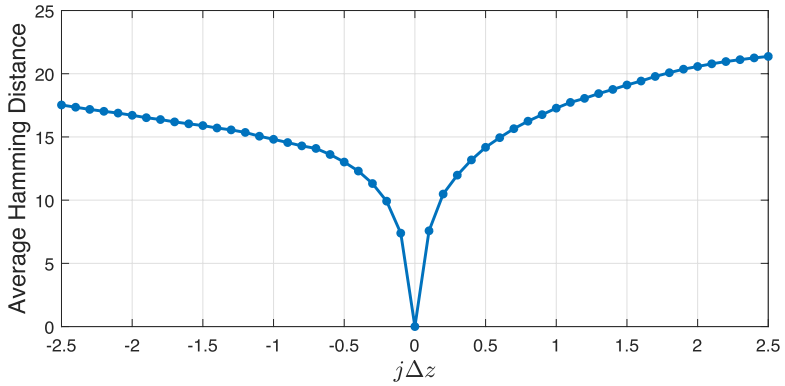

Fig. 9 Average Hamming distance as a function of $j\Delta z$. As the target complexity varies, attribute-controlled rhythms show a monotonically increasing degree of dissimilarity with respect to the unaltered output pattern

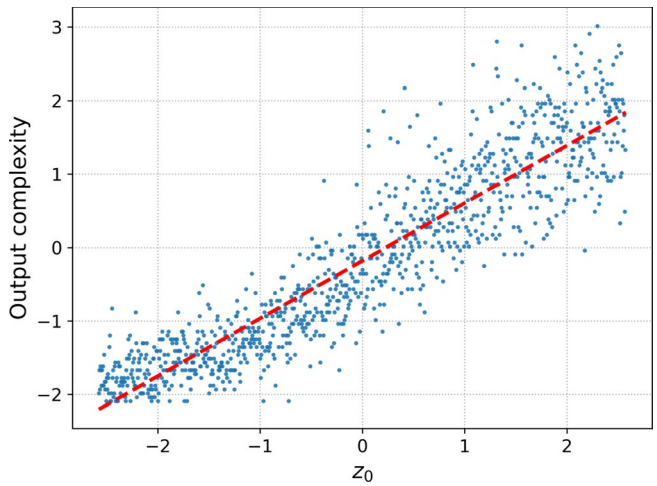

Fig. 10 Correlation between z_0 and the rhythm complexity of drum patterns generated from randomly-sampled latent vectors

conventional lower-complexity rhythms. Nevertheless, we argue that, in a practical application, this effect may be straightforwardly compensated by incorporating a suitable multiplicative factor into z_0, thus reestablishing an identity-like mapping between desired and measured complexity.

8 Conclusions and future work

In this article, we presented a novel latent rhythm complexity model designed for polyphonic drum patterns in the style of contemporary Western music. The proposed framework is based on a multi-objective learning paradigm in which variational autoencoding is supplemented by two additional loss terms, one for latent space regularization and the other targeting disentangled representations. In particular, the model is simultaneously tasked with predicting and embedding the value of a given musical attribute along one of its latent dimensions. This way, the ensuing latent space is encouraged to become semantically structured according to the target high-level feature, thus enabling straightforward interpretation and intuitive navigation. Moreover, we showed that decoding the latent representations thus obtained grants explicit control over the complexity of newly generated drum patterns. To achieve this, we introduced a new polyphonic rhythm complexity measure. To the best of our knowledge, the present work constitutes the first attempt at defining a proper complexity measure for polyphonic rhythmic patterns. The proposed measure was validated through a perceptual experiment which showed a high degree of correlation between measured complexity and that assessed by human listeners, as indicated by a Pearson coefficient above 0.95. Our method, being based on the linear combination of state-of-the-art monophonic measures applied to groups of functionally related drum voices, allows for great flexibility when it comes to measuring and weighting the contribution of individual (groups of) voices and may serve as a starting point for future research.

Endowing machines with an explicit understanding of perceptual features of music has the potential to enrich the capabilities of many AI-driven creative applications, including assisted music composition and attribute-controlled music generation. Besides, our work proves that regularizing a latent vector model according to target perceptual attributes may structure the resulting latent representations in a humanly interpretable way. Therefore, this approach might readily complement those applications involving the semantic exploration of musical content, such as music database navigation, recommender systems, and playlist generation.

Future work entails a large-scale survey to further validate the promising results presented in this article. This way, it would also be possible to determine the optimal complexity measure for each group of voices and derive a set of perceptually informed parameters for the proposed method. Moreover, examining the interplay between different yet related voices is not a concept solely pertaining to drums. In fact, we envision an adaptation of the proposed method to encompass, e.g., string quartets or four-part harmony chorales in which distinct voices are clearly identifiable and yet cannot be fully modeled independently of the others. Finally, building upon the existing work on the perception of monophonic rhythm complexity, the present study focuses on fixed-length quantized binary patterns. This means that only onset locations are considered, whereas dynamics, accents, time signature, and temporal deviations smaller than a tatum are disregarded. Although this choice is motivated by a divide-and-conquer modeling approach that regards these aspects of rhythm to be (at least partially) independent of each other, the validity of these assumptions is yet to be proved. In fact, one may argue that the temporal distribution of agogic and dynamic accents is likely to affect the complexity of a rhythmic pattern beyond the simple location of its onsets. Similarly, ditching quantization in favor of a continuous-time representation would fundamentally change the definition of syncopation, which could in turn entail a range of different psychoacoustic effects. Ultimately, these compelling questions remain open and must become the foundation of future research on the perception of rhythm.

Appendix

Toussaint's Metrical Complexity Measure

Introduced in [13], Toussaint's metrical complexity is based on the concept of syncopation. The measure, which was developed for monophonic binary patterns, entails assigning a weight to each pulse depending on their position in a regular metrical structure of length N. This is achieved by means of the hierarchy of pulses' strength proposed by Lerdahl and Jackendoff [63].

Such a hierarchy is obtained by iteratively adding units of weight to pulses spaced according to regular subdivisions at different metrical levels. For instance, a 6/8 rhythm can be either divided into three units of length two or two units of length three, and weights are assigned accordingly. For an arbitrary number of pulses, Lerdahl and Jackendoff's hierarchy thus derives from a (non-unique) tree-like structure built upon the prime factorization of N. For the sake of simplicity, however, we only consider the case of rhythms whose length is a power of two, which is relevant for our study of 2-bar 4/4 time scores.

At metrical level zero, the hierarchy is initialized to $\mathbf{h}^{(0)} = [1, 1, \ldots, 1]^\top$. Then, for each level $l = 1, \ldots, \log_2(N)$, weights are updated according to

$$\mathbf{h}^{(l)}[n] \leftarrow \mathbf{h}^{(l-1)}[n] + 1, \quad n = 2^l t + 1, \tag{8}$$

for $t = 0, \ldots, N/2^l - 1$. For a 16-pulse pattern, the rule in (8) yields

$$\mathbf{h}^{(4)} = [5, 1, 2, 1, 3, 1, 2, 1, 4, 1, 2, 1, 3, 1, 2, 1]^\top, \tag{9}$$

where four is the highest metrical level. This means that the first pulse is regarded as the "strongest," followed by the ninth, whereas, e.g., the second, fourth, sixth, and eighth are considered "weak" beats.

Hence, given a monophonic pattern $\mathbf{x} \in \{0, 1\}^N$ and a hierarchy $\mathbf{h} \in \mathbb{N}^N$, a quantity known as *metricity* is defined as the inner product

$$m(\mathbf{x}) := \mathbf{h}^\top \mathbf{x}. \tag{10}$$

In [13], metricity it is assumed to be inversely proportional to rhythm complexity. Inconveniently, however, it is also directly proportional to the number of onsets present in $\mathbf{x}$. To address both aspects at once, Toussaint's metrical complexity measure is defined as

$$f(\mathbf{x}) := M_\mathbf{x} - m(\mathbf{x}), \tag{11}$$

where $M_\mathbf{x}$ is the maximum metricity attainable with the number of onsets present in $\mathbf{x}$ (regardless of their original position) under the fixed hierarchy $\mathbf{h}$. Formally, given all possible binary patterns with N pulses, we can write

$$M_\mathbf{x} := \max_\kappa \left\{ m(\mathbf{x}'_\kappa) : \sum_{n=1}^{N} \mathbf{x}'_\kappa[n] = \eta(\mathbf{x}) \right\}, \tag{12}$$

where $\kappa = 1, \ldots, 2^N$ and $\eta(\mathbf{x}) := \sum_{n=1}^{N} \mathbf{x}[n]$ corresponds to the number of onsets in the target pattern $\mathbf{x}$. By taking into account the quantity $M_\mathbf{x}$, Toussaint's complexity measure becomes effectively independent

of the number of onsets, a favorable property if one does not want to attribute higher complexity to a pattern solely because it happens to have a large number of rhythmic events.

Toussaint's metrical complexity is only one of many rhythm complexity measures that can be found in the literature. For an overview, we refer the readers to [44].

Abbreviations

GMD	Groove MIDI Dataset
KLD	Kullback-Leibler divergence
MIDI	Musical Instrument Digital Interface
MSE	Mean squared error
VAE	Variational autoencoder

Acknowledgements
The authors would like to acknowledge Clément Jameau for the preliminary work conducted on the topic and for implementing part of the codebase. AIM wishes to thank Jacopo Cavagnoli for the valuable discussion concerning the subjective nature of rhythm complexity and the perception of drums.

Authors' contributions
AIM conceptualized the study, developed the method, designed and conducted the listening experiment, and wrote the original draft. MZ and AS supervised the work and reviewed the manuscript. All authors read and approved the final manuscript.

Funding
Not applicable

Availability of data and materials
The dataset analyzed during the current study is available in the Groove MIDI Dataset repository: https://magenta.tensorflow.org/datasets/groove. The data generated for the listening test are available at https://github.com/ilic-mezza/polyphonic_rhythm_complexity.

Declarations

Ethics approval and consent to participate
All participants to the listening experiment gave their consent to take part in the study and no personal data was collected.

Competing interests
The authors declare that they have no competing interests.

Received: 31 July 2022 Accepted: 21 December 2022
Published online: 17 February 2023

References

1. A. Flexer, T. Grill, The problem of limited inter-rater agreement in modelling music similarity. J. New Music. Res. **45**(3), 239–251 (2016)
2. M. Sordo, Ò. Celma, M. Blech, E. Guaus, in *Proc. of the 9th International Conference on Music Information Retrieval,* Philadelphia, 2008. The quest for musical genres: do the experts and the wisdom of crowds agree? (2008), p. 255–260
3. S. Yang, C.N. Reed, E. Chew, M. Barthet, Examining emotion perception agreement in live music performance. IEEE Trans. Affect. Comput. (2021). https://ieeexplore.ieee.org/document/9468946/
4. J.L. Walker, Subjective reactions to music and brainwave rhythms. Physiol. Psychol. **5**(4), 483–489 (1977)
5. T.E. Matthews, J.N.L. Thibodeau, B.P. Gunther, V.B. Penhune, The impact of instrument-specific musical training on rhythm perception and production. Front. Psychol. **7**, 1–16 (2016)
6. S.J. Morrison, S.M. Demorest, Cultural constraints on music perception and cognition. Prog. Brain Res. **178**, 67–77 (2009)
7. M. Leman, *Music, Gestalt, and Computing: Studies in Cognitive and Systematic Musicology* (Springer, Berlin-Heidelberg, 1997)
8. H.C. Longuet-Higgins, C.S. Lee, The rhythmic interpretation of monophonic music. Music. Percept. **1**(4), 424–441 (1984)
9. D.-J. Povel, P. Essens, Perception of temporal patterns. Music. Percept. **2**(4), 411–440 (1985)
10. S. Arom, G. Ligeti, *African Polyphony and Polyrhythm: Musical Structure and Methodology* (Cambridge University Press, Cambridge, 1991)
11. A.S. Tanguiane, A principle of correlativity of perception and its application to music recognition. Music. Percept. Interdiscip. J. **11**(4), 465–502 (1994)
12. I. Shmulevich, O. Yli-Harja, E. Coyle, D.-J. Povel, K. Lemström, Perceptual issues in music pattern recognition: complexity of rhythm and key finding. Comput. Hum. **35**(1), 23–35 (2001)
13. G. Toussaint, in *Bridges: Mathematical Connections in Art, Music, and Science*, Towson, 2002. A mathematical analysis of african, brazilian, and cuban clave rhythms. (Bridges Conference, Winfield, 2002), p. 157–168
14. L.M. Smith, H. Honing, in *Proc. of the 2006 International Computer Music Conference*, New Orleans, 2006. Evaluating and extending computational models of rhythmic syncopation in music. (Michigan Publishing, Ann Arbor, 2006), p. 688–691
15. W.T. Fitch, A.J. Rosenfeld, Perception and production of syncopated rhythms. Music. Percept. **25**(1), 43–58 (2007)
16. G.T. Toussaint, in *Proc. of the 12th International Conference on Music Perception and Cognition & the 8th Conference of the European Society for the Cognitive Sciences of Music*, Thessaloniki, 2012. The pairwise variability index as a tool in musical rhythm analysis. (School of Music Studies, Aristotle University of Thessaloniki, Thessaloniki, 2012), p. 1001–1008
17. P. Vuust, M.A.G. Witek, Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music. Front. Psychol. **5**, 1–14 (2014)
18. A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, K. Kavukcuoglu, in *Proc. of the 30th International Conference on Neural Information Processing Systems*, Barcelona, 2016. Conditional image generation with PixelCNN decoders. (Curran Associates Inc., Red Hook, 2016), p. 4797–4805
19. C. Ledig, L. Theis, F. Huszár, J. Caballero, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, in *Proc. of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 2017. Photo-realistic single image super-resolution using a generative adversarial network. (IEEE, Piscataway, 2017), p. 105–114
20. A. Razavi, A. van den Oord, O. Vinyals, in *Proc. of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, 2019. Generating diverse high-fidelity images with VQ-VAE-2. (Curran Associates Inc., Red Hook, 2019), p. 1–11
21. S.R. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, S. Bengio, in *Proc. of the 20th SIGNLL Conference on Computational Natural Language Learning*, Berlin, 2016. Generating sentences from a continuous space. (Association for Computational Linguistics, Stroudsburg, 2016), p. 10–21
22. Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. Le, R. Salakhutdinov, in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, 2019. Transformer-XL: Attentive language models beyond a fixed-length context. (Association for Computational Linguistics, Stroudsburg, 2019), p. 2978–2988
23. T. Brown et al., Language models are few-shot learners. Adv Neural Inf Proc Syst **33**, 1877–1901 (2020)
24. A. van der Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, K. Kavukcuoglu, Wavenet: a generative model for raw audio. (2016). arXiv:1609.03499
25. W.-N. Hsu, Y. Zhang, J. Glass, in *Proc. of the 18th Annual Conference of the International Speech Communication Association*, Stockholm, 2017. Learning latent representations for speech generation and transformation. (Curran Associates Inc., Red Hook, 2017), p. 1273–1277
26. K. Akuzawa, Y. Iwasawa, Y. Matsuo, in *Proc. of the 19th Annual Conference of the International Speech Communication Association*, Hyderabad, 2018. Expressive speech synthesis via modeling expressions with variational autoencoder. (Curran Associates Inc., Red Hook, 2018), p. 3067–3071
27. S. Oore, I. Simon, S. Dieleman, D. Eck, K. Simonyan, This time with feeling: learning expressive musical performance. Neural Comput. Applic. **32**(4), 955–967 (2020)

28. A. Roberts, J. Engel, S. Oore, D. Eck, Learning latent representations of music to generate interactive musical palettes. Paper presented at the 2018 ACM Workshop on Intelligent Music Interfaces for Listening and Creation, Tokyo, 2018.
29. A. Roberts, J. Engel, C. Raffel, C. Hawthorne, D. Eck, in *Proc. of the 35th International Conference on Machine Learning*, Stockholm, 2018. A hierarchical latent vector model for learning long-term structure in music, vol. 80. (Curran Associates Inc., Red Hook, 2018), p. 4364–4373
30. C.-Z.A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. Dai, M. Hoffman, M. Dinculescu, D. Eck, Music transformer: Generating music with long-term structure. (2019). arXiv:1809.04281
31. J. Gillick, A. Roberts, J. Engel, D. Eck, D. Bamman, in *Proc. of the 36th International Conference on Machine Learning*, Long Beach, 2019. Learning to groove with inverse sequence transformations, vol. 97. (Curran Associates Inc., Red Hook, 2019), p. 2269–2279
32. P. Dhariwal, H. Jun, C. Payne, J.W. Kim, A. Radford, I. Sutskever, Jukebox: a generative model for music. (2020). arXiv:2005.00341
33. J.-P. Briot, F. Pachet, Deep learning for music generation: challenges and directions. Neural Comput. Applic. **32**(4), 981–993 (2020)
34. J.H. Engel, M.D. Hoffman, A. Roberts, Latent constraints: learning to generate conditionally from unconditional generative models. Paper presented at the 5th International Conference on Learning Representations, Toulon, 2017.
35. G. Hadjeres, F. Nielsen, F. Pachet, in *2017 IEEE Symposium Series on Computational Intelligence*, Honolulu, 2017.GLSR-VAE: geodesic latent space regularization for variational autoencoder architectures. (Curran Associates Inc., Red Hook, 2017), p. 1–7
36. G. Brunner, A. Konrad, Y. Wang, R. Wattenhofer, in *Proc. of the 19th International Society for Music Information Retrieval Conference*, Paris, 2018. MIDI-VAE: modeling dynamics and instrumentation of music with applications to style transfer. (2018), p. 747–754
37. H.H. Tan, D. Herremans, in *Proc. of the 21st International Society for Music Information Retrieval Conference*, Montréal, 2020. Music fadernets: controllable music generation based on high-level features via low-level feature modelling. (2020), p. 109–116
38. G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, M. Ranzato, Fader networks: manipulating images by sliding attributes. Adv. Neural. Inf. Proc. Syst. **30**, 5969–5978 (2017)
39. Z. Jiang, Y. Zheng, H. Tan, B. Tang, H. Zhou, in *Proc. of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, 2017. Variational deep embedding: an unsupervised and generative approach to clustering. (AAAI Press, Cambridge, 2017), p. 1965–72
40. A. Pati, A. Lerch, Attribute-based regularization of latent spaces for variational auto-encoders. Neural Comput. Applic. **33**(9), 4429–4444 (2021)
41. A. Pati, A. Lerch, in *Proc. of the 22nd International Society for Music Information Retrieval Conference*, Online, 2021. Is disentanglement enough? On latent representations for controllable music generation. (2021), p. 517–524
42. F. Gómez, A. Melvin, D. Rappaport, G.T. Toussaint, in *Renaissance Banff: Mathematics, Music, Art, Culture*, Banff, 2005. Mathematical measures of syncopation. (Canadian Mathematical Society, The Banff Centre, PIMS, 2005), p. 73–84
43. F. Gómez, E. Thul, G. Toussaint, in *Proc. of the 2007 International Computer Music Conference*, Copenhagen, 2017. An experimental comparison of formal measures of rhythmic syncopation. (Michigan Publishing, Ann Arbor, 2007), p. 101–104
44. E. Thul, G.T. Toussaint, in *Proc. of the 9th International Society for Music Information Retrieval Conference*, Philadelphia, 2008. Rhythm complexity measures: a comparison of mathematical models of human perception and performance. (2008), p. 663–668
45. M. Müller, *Fundamentals of music processing: audio, analysis, algorithms, applications* (Springer, Basel, 2015)
46. J. Bilmes, in *Proc. of the 1993 International Computer Music Conference*, Tokyo, 1993. Techniques to foster drum machine expressivity. (Michigan Publishing, Ann Arbor, 1993), p. 276–283
47. M. Keith, *From Polychords to Pólya: Adventures in Musical Combinatorics* (Vinculum Press, Princeton, 1991)
48. G. Toussaint, in *Meeting Alhambra, ISAMA-BRIDGES Conference Proceedings*. Classification and phylogenetic analysis of african ternary rhythm timelines. (2003), p. 25–36
49. E. Grabe, E.L. Low, Durational variability in speech and the rhythm class hypothesis. Pap. Lab. Phonol. **7**(1982), 515–546 (2002)
50. P.C. Vitz, T.C. Todd, A coded element model of the perceptual processing of sequential stimuli. Psychol. Rev. **76**(5), 433–449 (1969)
51. A. Lempel, J. Ziv, On the complexity of finite sequences. IEEE Trans. Inf. Theory. **22**(1), 75–81 (1976)
52. G. Sioros, C. Guedes, in *Proc. of the 12th International Society for Music Information Retrieval Conference*, Miami, 2011. Complexity driven recombination of MIDI loops. (2011), p. 381–386
53. G. Sioros, A. Holzapfel, C. Guedes, in *Proc. of the 13th International Society for Music Information Retrieval Conference*, Porto, 2012. On measuring syncopation to drive an interactive music system. (2012), p. 283–288
54. Subjective evaluation of speech quality with a crowdsourcing approach. Rec. ITU-T P.808. (International Telecommunication Union, Geneva, 2018)
55. I. Higgins, L. Matthey, A. Pal, C.P. Burgess, X. Glorot, M.M. Botvinick, S. Mohamed, A. Lerchner, beta-VAE: learning basic visual concepts with a constrained variational framework. Paper presented at the 5th International Conference on Learning Representations, Toulon, 2017.
56. T. Adel, Z. Ghahramani, A. Weller, in *Proc. of the 35th International Conference on Machine Learning*, Stockholm, 2018. Discovering interpretable representations for both deep generative and discriminative models, vol. 80. (Curran Associates Inc., Red Hook, 2018), p. 50–59
57. Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives. IEEE Trans. Pattern. Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
58. C.P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, A. Lerchner, Understanding disentangling in β-VAE. Paper presented at the 2017 NIPS Workshop on Learning Disentangled Representations, Long Beach, 2018.
59. J. Schmidhuber, Learning factorial codes by predictability minimization. Neural Comput. **4**(6), 863–879 (1992)
60. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks. J. Mach. Learn. Res. **17**(1), 2096–2030 (2016)
61. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization. Paper presented at the 3rd International Conference on Learning Representations, San Diego, 2015.
62. C. Eastwood, C.K. Williams, A framework for the quantitative evaluation of disentangled representations. Paper presented at the 6th International Conference on Learning Representations, Vancouver, 2018.
63. F. Lerdahl, R.S. Jackendoff, *A Generative Theory of Tonal Music, Reissue, with a New Preface* (MIT Press, Cambridge, 1996)

Publisher's Note