Deep Musical Information Dynamics:
Novel Framework for Reduced Neural-Network Music Representation with Applications to MIDI and Audio Analysis and Improvisation

Shlomo Dubnov, Ke Chen, Kevin Huang

Department of Music, University of California San Diego

Abstract. In this paper we use a recently proposed framework called Deep Musical Information Dynamics (DMID) to explore information contents of deep neural models of music by applying bit-rate reduction to latent representations that are used to generate the musical surface. Our approach is partially motivated by rate-distortion theories of human cognition that claim that in order to deal with the complexity of sensory information some information must be lost or discarded in the act of perception. When lossy encoding is done over time, this may alter the anticipations that are formed within and across voices at different levels of representation of the musical structure. Moreover, we postulate that a goal of a musical machine learning system, and possibly human musical learning system, is learning a latent representation that "explains out" most of the Information Dynamics of the Musical surface. This assumption is explored in DMID through several experiments on symbolic (MIDI) and acoustic (spectral) music representations using a Variational Auto Encoding scheme with an additional bit-rate reduction step. Our results suggest that higher mutual information can be found between latent representations encoded with reduced rates. The DMID framework is significant for studies of computational creative music systems since it allows exploration of information relations in latent and surface levels of musical data in a quantifiable and computationally tractable manner.

Dedicated to the memory of Naftali Tishby (1952–2021)

1 Introduction

Musical Information Dynamics is a field in music analysis that is inspired by theories of musical anticipation (Meyer, 1956; Huron, 2006), which deals with quantifying the amount of information passing over time between the past and future in a musical signal (Dubnov, 2006; Abdallah & Plumbley, 2009; Marcus T. Pearce, 2012). Modeling and capturing information in a musical signal heavily depends on the representation of the data. Recently, generative deep neural models have become the primary tools for representation learning in multiple domains. The compelling idea of machine learning is that various aspects of

modeling data from the world, that include embeddings, representation, and encodings are captured using vector spaces that beautifully map structures into mathematical space. The powerful idea of generative modeling is that in order to effectively represent the world, the learning system needs to be able to effectively "imagine" it. Mathematically, this means that the goal of learning is to be able to approximately reproduce or simulate the statistics of the world from an internal representation that the learning system constructs. Finding such representation entails being in a state of homeostasis between the system and the outside environment, which in our cases comprises the external world of music and the internal minds of the musicians themselves, be it listeners or composers.

When time is involved, ideas such as Friston's free-energy applied to predictive coding (Friston, 2012) and Schmidhuber's world models (Schmidhuber, 2010) that use Reinforcement Learning (RL) to capture compressed spatial and temporal representation of the environment come to mind. We discuss these ideas briefly in the conclusion. It might be valuable to point out that the main difference between these approaches and the work presented here is that we do not assume an active learning or active inference scenario. The current paper focuses on representation learning aspects of a musical model, which amounts to specifying the goals for optimizing model parameters that reduce the uncertainty about music by forming the best latent or posterior beliefs at different levels of reduction. Extensions to active inference or planning of optimal musical actions to validate or violate musical anticipation in the process of composition or improvisation are possible directions for future research.

In order to learn generative musical models we desire a representation that comprises of multiple levels of structure that can "explain out" most of the music information dynamics. We employ deep neural networks as models of musical structure and use them to explore the temporal relations at various levels and at different fidelity of encoding. The novel framework called Deep Musical Information Dynamics (DMID)(Dubnov, 2020) considers music as a complex phenomenon comprising of parallel temporal representations at different fidelities of encoding, organized in time. Our analysis juxtaposes a reduced information (low bit-rate encoding) of the latent variables with high rate representation of the musical surface data, and applies various statistical test to it with qualitative observations. Motivated by rate-distortion theories of human cognition(R.Sims, 2016), DMID provides a framework for exploring possible relations between latent anticipations existing in the listener's mind and information dynamics of the actual musical surface. Even if this approach does not claim to represent the process of how a composer actually writes music, it can serve as a useful conceptual model for analysis of musical structures in relation to creative and listening thought processes.

Technically speaking, earlier studies of Musical Information Dynamics used the concept of Information Rate (IR) to estimate the mutual information between past and future of a musical signal (Dubnov, Assayag, & Cont, 2011). Generative models that maximize IR were shown to provide good results in machine improvisation systems (Pasquier, Eigenfeldt, Bown, & Dubnov, 2017).

Since music is constantly changing, the ability to capture structure in time depends on the way similarity is computed and predictions or anticipations are formed over time. For example, two musical measures improvised over the same sequence of chords often have different notes because of voicing considerations, rhythm changes, embellishments and more. Accordingly, finding repetitions in time require representations that are robust to statistical variations in the signal. DMID analysis offers a combined approach to learning representation and measuring the predictive information at different structural representation levels. Our goal is to show the importance of considering the trade-off between encoding complexity of musical structures and their predictive properties at different levels of musical representation (latent versus surface).

In the paper, the DMID framework is demonstrated by conducting experiments with several different music representations. For symbolic (MIDI) data, we conduct two experiments: monophonic representation to study counterpoint relations between voices, and a polyphonic (piano roll) case to study temporal structure at different reductions. The structural analysis is repeated for an audio recording using a spectral signal representation. In all cases, the surface is encoded using Variational Inference models, each adapted for the specific representation case. The variational encoding framework allows us to first establish a high-rate representation of the musical surface, which is then reduced using a bit-allocation method into a second stream of low rate encoding. This reduction serves to investigate the trade-off between fidelity and quality of music representation and its predictive properties. In a sense, the reduced representation can be considered as a conceptual or mental abstraction of music that might be needed to capture longer-term structural similarities in music that cannot be perceived when looking or listening to music in full details. To allow this combined representation-prediction trade-off, the combined loss defined in DMID includes both the information rate in terms of time evolution for each representation stream, and the accuracy of the encoding of the musical surface by that stream. Both the temporal and cross-stream relations are measured in terms of mutual information within and across the high and low rate representations. Additionally, we consider if reducing the fidelity of musical representation is manifested across voices in a polyphonic musical piece. Although this case is not directly captured by the DMID model that is defined for a single musical stream representated in parallel at different rates, it is reasonable to assume that relations between multiple voices in the same musical piece might be also revealed differently subject to their reduced encoding. Accordingly, we conduct an additional experiment using the DMID approach where we consider the case of a counterpoint, where we explore the mutual information between voices in the same musical piece, examined at different reduced representation rates. Our findings suggest that there is a mid-level representation that actually shows higher correspondence between the voices compared to full-rate or very low-rate encoding. The implications of our findings are discussed throughout the paper.

The structure of the paper is as follows: in section 2 we introduce the concept of DMID. Section 3 deals with different methods to learn musical representation

from data. In section 4 we discuss the core concept of reducing the representation encoding by using methods of lossy compression from information theory. In section 5 we introduce the methods and techniques of mutual information estimation. In section 6 we further extend the mutual information method for measuring changes in information contents over time based on the Variable Markov Oracle (VMO). In section 7 we report the experiments on monophonic and polyphonic midi and audio. Section 8 concludes the paper with a discussion of the findings in the proposed DMID framework as a general approach to study of creative musical systems.

1.1 Detailed description of the paper structure

Due to the size of the paper, we provide here a more detailed account of the paper structure. This subsection can be considered as a sort of "Table of Contents" for the paper. Readers interested in a particular topic might be able to navigate the paper in a non-sequential manner, as we tried to provide cross-references between sections to allow a non-linear reading. The contents of the papers can be summarized as follows:

- Section 2 introduces the concept of latent predictive information, which is the basis for analysis of temporal structure of the reduced or possible "mental" abstract representations of music that emerges when some details of the encoding are eliminated.
- In section 3 we propose the use of a Variational Inference method, namely the Variational Auto Encoder (VAE), as our method of choice for representation learning. For the monophonic case we use a hierarchical VAE that uses short term temporal learning with Gated Recurrent Units (GRU) at a beat and tick temporal levels. For polyphonic MIDI representation, we consider a piano-roll representation that is cut into short-time simultaneous events (frames), with multiple notes (multi-hot encoding) passed into a VAE, one frame in a time. Similarly, we treat audio as short time audio frames analyzed in terms of their frequency contents using the magnitudes of the Short Time Fourier Transform (STFT), also known as a magnitude spectrogram.
- In section 4 we introduce the technique of optimal bit allocation based on Rate-Distortion theory, where we operate under the assumption that the encoding vectors learned in the previous representation learning step are element-wise independent and obey multi-variate Gaussian statistics. This is, of course, a gross simplification, but we use it for practical algorithmic purposes to be able to choose what elements from the full-rate encoding could be eliminated for a given encoding rate.
- In section 5 a method for estimating mutual information which captures statistical dependencies between random variables averaged over multiple realizations of these variables for an arbitrary statistical distribution, is introduced. It is easiest to think about Mutual Information as a generalization of a correlation measure, which effectively captures the same statistical dependency for the case of Gaussian signal. In our DMID model, we estimate

mutual information in two different settings – first between different representation levels, i.e. between full and reduced rate representations, and second, between past and future of each representation stream. Since the statistics of our streams are complex and unknown, special techniques for measuring such statistical relations are required. Accordingly, we introduce first the so-called Mutual Information Neural Estimation (MINE) method to estimate mutual information between different representation levels. We later on use this method for studying also the mutual information between voices in polyphonic music, at different representation levels.
– In section 6 we introduce the concept of predictive information and describe an approximate predictive information estimator based on the Variable Markov Oracle (VMO). The advantage of VMO over MINE is that it allows the estimation of instantaneous values of mutual information over time, while MINE requires averaging over the whole signal. VMO uses a "trick" where a variable length string compression algorithm is used as a proxy for an entropy measure, characterising the instantaneous surprisal or uncertainty of a signal in terms of encoding the present by finding approximate repetitions in the signal past. Loosely speaking, the same way STFT allows short time-frequency analysis by compromising on the ideal notion of pure frequency as something that exists from an infinite past to an infinite future, VMO compromises on the ideal notion of mutual information as something that is averaged over full statistics of a signal, and approximates it in dynamic manner by considering approximate compression of the present based only on its past. This method of time-changing predictive information is called Information Rate, which is later applied to the analysis of predictive structure of a complete musical piece at different representation levels.
– Finally, in section 7 we report the experiments on our three types of musical situations: 1.) Monophonic Symbolic Music, 2.) Polyphonic Symbolic Music, and 3.) Audio Recording of Polyphonic Music. In section 8 the representation-prediction trade-offs are summarized into a joint objective that combines reduced latent encoding with predictive information criteria into a single mathematical expression. This learning objective is further discussed in terms of the differences between learning generative representation and active inference, relating it to creativity and the role DMID representation plays in finding the right intrinsic motivation and empowerment for a musical agent.
– Some mathematical and implementation details related to the experimental results are deferred to the Appendix.

2 Deep Musical Information Dynamics (DMID)

The materials of this section are largely based on an initial paper(Dubnov, 2020) that provides the underlying motivation for proposing a DMID model. The underlying creative cognition assumption is that the imagination of the composer, improviser, or listener is an active process that involves encoding the current

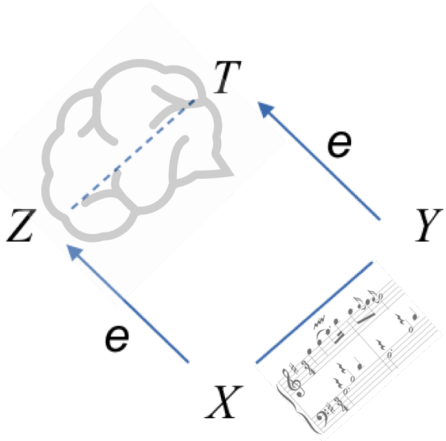


Fig. 1: Graph of statistical dependencies between the different model variables. The letter "e" represents an embedding created by encoding. Later on we will use VAE as our encoding method. Complexity of the embedding will be controlled by bit-rate allocation.

incoming musical stimulus and the anticipation of its future. In information processing terms, encoding deals with extracting salient features that allow efficient representation of the signal, while a second process is trying to predict the evolution of those features over time. While the first process often deals with the recognition of isolated sound events, it is the second aspect of prediction that is commonly recognized as the underlying mechanism for anticipations that allow for the creation of surprise, validation, or violation of expectation and building of tensions and resolutions in a musical narrative.

In order to allow a quantitative approach to the analysis of what's going on in the "musical mind"(Assayag, 1999), we propose an information theoretic model for the relation between four factors: the signal past X, the signal present Y, and their internal or mental representation in terms of past and present latent variables Z and T, respectively. This highly simplified model assumes a set of Markov chain relations that specify the so-called Data Processing Inequality relations (Cover & Thomas, 1991). A set of variables A, B, C obeys a Markov chain relation if we have $P(A, B, C) = P(A|B)P(B|C)P(C)$, or in other words, if the conditional distribution of A depends only on B while being conditionally independently of C. The relationship between the aforementioned factors X, Y, Z, and T is illustrated in Figure 1, with Markov relations between the past of the signal X that is encoded into a latent representation Z, the future of the signal Y that depends on its past X, and its approximation by decoding from a latent representation T that is predicted from past latent representation Z. The significance of Markov relation is that it specifies the model assumption about how aspects of musical information evolve and depend on each other. The

underlying assumption in latent models is that the data we observe is generated by some hidden underlying factor, and the goal of the learning system is to find the best interpretation or understanding of the data we observe by "a-posterior" finding the best latent explanation or deep representation. To rephrase our latent representation assumptions in terms of Markov relations, we can summarize the goal of our model as finding a hidden Z that is, mathematically speaking, "transmitted" over a noisy channel that renders it into X, which in turn is finally decoded into Y. In other words, the triplet X, Y, Z obeys Markov relation $p(X, Y, Z) = p(Y|X)p(X|Z)p(Z)$.

Using the Markov relation $Z-X-Y$, we can try to formulate a mathematical expression for the goals underlying the learning process of such a music system (Elliott, 1993). Our expression for the optimization goal comprises a combination of factors - simplicity of representation versus its prediction ability. Accordingly, we are looking for a representation that minimizes the discrepancy, or statistical difference, between signal prediction using complete information about the past X, versus its prediction capability by using a simplified encoding of the past Z. The overall quality of such error is averaged over all possible encoding pairs X, Z of the musical surface and its latent code, respectively.

$$\langle D_{KL}(p(Y|X)||p(Y|Z))\rangle_{p(X,Z)} = I(X, Y|Z) = I(X, Y) - I(Z, Y) \qquad (1)$$

In our notation, $D_{KL}(\cdot, \cdot)$ is the Kullback-Liebler (KL) divergence between different distributions, and $I(\cdot, \cdot)$ is the mutual information between their random variables. The mathematical definitions and proof of these relations is provided in the Appendix. Since $I(X, Y)$ are independent Z, minimizing the KL divergence happens when $I(Z, Y)$ is maximized, with zero KL obtained when $I(Z, Y) = I(X, Y)$. In other words, we postulate that a goal of a musical machine learning system, and possibly human musical learning systems, is learning a latent representation Z that "explains out" most of the Musical Information Dynamics $I(X, Y)$. This principle is expressed as minimization of $I(X, Y|Z)$. i.e. finding a latent Z so that there will be very little remaining information passing between the past X and the present Y of the musical surface itself.

To complete our postulate, some additional constraints on Z need to be specified, since if $Z = X$, this condition is trivially satisfied. Accordingly, we add constraints that require that the latent variable representation is as compact or as simple as possible. In the process of VAE learning, the latent representation is simplified by bringing it to be as close to a Gaussian uncorrelated noise as possible[1]. Additionally we introduce lossy compression of Z for a pre-trained VAE using a bit-allocation procedure. We term this principle of minimization in equation 1 *DMID minimization principle*, and investigate it here for different type of musical data, with their different respective VAEs, and using different reduction levels.

[1] This is the ELBO minimization principle, to be discuss in the next section.

2.1 Adding simplicity requirement of the latent representation

The above formulation says that we need to be looking for some meaningful Z, but we still do not know what such Z looks like, so minimizing KL could be trivially satisfied by taking $Z = X$. To avoid such a trivial solution, we add a constraint on Z requiring it to be the most compact or simplest latent "explanation" that is derived from X. In information theoretical terms, we can write this criteria as the minimization of $I(X, Z)$. In a sense, we look for the least amount of bits of information about X to be contained in Z. Later on we will actually need a third parameter that prevents $I(X, Z)$ from going to zero. This can be achieved by adding a fidelity requirement or bound on distortion between X and Z, that will be denoted as $D(X, Z)$. It is important to note that this distortion D is not same as a KL-divergence, but is rather some physically motivated distortion, such as Mean Square Error (MSE) or some other distortion measure between the "compressed" X, as expressed by Z, and the original X. For the moment, we will ignore this fidelity constraint and consider only the first competing relations between maximizing $I(Z, Y)$ and minimizing $I(Z, X)$. Combining the two goals, we arrive at the target function for our learning system

$$\max_{P(Z|X)} \{I(Z, Y) - \lambda I(X, Z)\} \tag{2}$$

This formulation bears close resemblance to the idea of the Information Bottleneck (IB) (Tishby & Zaslavsky, 2015). The formulation of IB is to say that a learning system tries to find the most compact representation of X that still provides most information about a different variable Y. It should be noted that IB does not allow an Auto-Encoding like unsupervised representation learning, which will be separately handled through VAE encoding. Accordingly, predictive IB looks at the next Y that is different from the reconstruction of X itself. In the following we will combine the predictive IB with the unsupervised representation learning of X.

2.2 Three factors with two competing criteria

To summarize, we have the following two criteria that combine competing goals for three factors $I(X, Z)$, $I(Y, Z)$ and $D(X, Z)$

- finding the most compact representation of present X that is most informative about the future Y (i.e. time information).
- finding the most compact representation of X from which X can be recovered with minimal distortion $D(X, Z)$ (i.e. reconstruction qualify).

We identify the first criteria as time information and the second with representation. In future sections these the two criteria are combined into one goal, using VAE for representation learning and VMO for the temporal modeling. Before going into this discussion, we need to address one additional aspect that is particular to DMID, which is the amount of predictive information present in the latent codes themselves.

2.3 Latent Predictive Information

An important goal of representation learning that stems from the DMID minimization principle is maximizing the amount of information passing between the encoded past embedding Z and the next musical surface Y. Using the second set of Markov relations shown in Figure 1, we relate temporal information between the latent variables Z and T themselves. Accordingly, we express $I(Z,Y) = I(Z,T) - I(Z,T|Y)$ as the ability to predict the future of the musical surface Y from past embedding Z, relative to predictive information in the latent embeddings themselves. From these Markov relations we see that the amount of information that past embedding Z is carrying about the future surface Y is less then the amount of information carried between the past and future of the embeddings Z and T, respectively. The term $I(Z,T)$ can be considered as latent predictive information, corresponding to some sort of a reduced "imaginary" or abstract musical anticipation.

Ignoring $I(Z,T|Y)$ (or assuming it is zero) means that hearing the next musical frame Y causes the "internal" representations of the past and the present to be independent. In such a case the expectations $I(Z,Y)$ from the reduced representation of the past Z towards the present sound Y are the same as the expectation of the actual sound Y, so that listening to the sound itself does not add any surprise. This creates an exceptional situation where a computer system that operates by maximization of the information rate (IR) in the latent states sequence $I(Z,T)$ alone is maximizing the predictive ability of the model, and no additional information or surprise is contained in the sound being produced next. Musically speaking this excludes any variations in terms of expressive inflection or sound performance aspects, if we consider here the acoustic interpretation. In the score interpretation, the surface aspects might refer to melodic and texture details, as contrasted with some underlying reduced musical structure such as prolongations or harmonic or metric rules.

This discussion points to the important difference between latent encoding and surface detail, which creates an additional surprisal factor $I(Z,T|Y) = I(Z,T) - I(Z,Y)$ that might be important for various musical situations. In other words, the total musical effect is a combination of the anticipation error of a reduced latent process and the additional information in the next realization of the musical surface. Both factors form the basis for the musical tension that emerges in the listening process and are factors that can be deliberately established during the composition design phase.

In other words, the difference between information contained in imagining the continuation of a reduced musical representation in terms of some latent variables, versus the actual realization of that next musical surface, is possibly an important aspect in the process of composition or listening.

3 Variational Inference Models for Music

The question of how to encode music in a latent space is a burgeoning research topic in the field of machine learning for computer music. In this paper we

apply the Variational Autoencoder (VAE) (Kingma & Welling, 2013) to map the explicit music surfaces into the latent variables with higher structure. An autoencoder is a structure that contains an encoder and a decoder. When used in the deep learning field, each of them is composed of some types of neural networks. During the training process, an input is encoded into a latent vector, also known as an embedding, and an output is then decoded from the same latent vector. The goal of training the autoencoder is to make the input and output as similar as possible, while trying to obtain an embedding that has some desired properties, such as dimensional reduction, sparseness and so on. Broadly speaking, autoencoders belong to the category of unsupervised learning methods, since their goal is to reveal structure that exists in the data rather then classify it.

Unlike standard autoencoders that translate the data into a more desirable representation, VAE explicitly constrains the latent variables Z so that they should be random variables distributed according to some prior $p(z)$. The input X and latent code Z can then be seen as random variables $Z \sim p(Z), X \sim p(X|Z)$. The VAE consists of an encoder probability $q_\lambda(Z|X)$, which approximates the posterior probability $p(Z|X)$, and a decoder probability $p_\theta(X|Z)$, which parameterizes the likelihood $p(X|Z)$. In practice, the approximate posterior and likelihood distributions are parameterized by weights of connections between neurons in artificial neural networks. Posterior inference is done by minimizing the KL divergence between the encoder and the true posterior. It can be proved that this optimization problem is the same as maximizing the evidence lower bound (ELBO):

$$ELBO = E[\log p_\theta(X|Z)] - KL(q_\lambda(Z|X)||p(Z)) \leq \log p(X) \tag{3}$$

Another important insight about VAE is that it is a generative model, where the latent variables Z are used to "drive" the decoder into producing novel samples of the data X. In a statistical sense, VAE is used to learn the distribution $P(X)$, rather then to simply encode X.

In this paper, we will use VAE to analyze three formats of music: (1) polyphonic music, (2) monophonic counterpoint music, and (3) music waveform in recorded or synthesized audio. Different formats of music require different structures of neural networks to encode and decode.

3.1 Hierarchical VAE for Monophonic Music

Many works (Chen, Zhang, Dubnov, Xia, & Li, 2019; Roberts, Engel, Raffel, Hawthorne, & Eck, 2018; Pati, Lerch, & Hadjeres, 2019; Chen, Wang, Berg-Kirkpatrick, & Dubnov, 2020; Chen, 2021) explore the encoding method of monophonic music that contains one note at a time. In this paper, we use a hierarchical VAE model (Pati et al., 2019; Chen, Xia, & Dubnov, 2020) for the encoding and decoding of monophonic music. As shown in Figure 2, this method consists of pre-processing a series of musical notes into a token sequence, with three types of tokens:

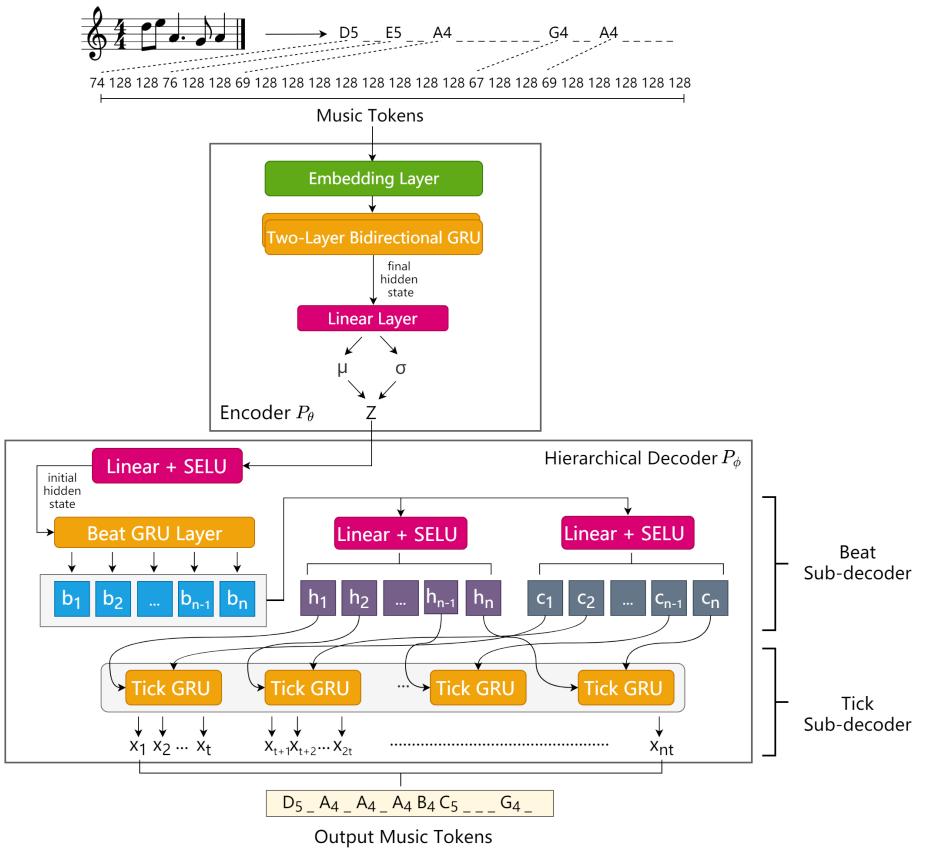


Fig. 2: The network architecture of monophonic music Variational Autoencoder.

– Pitch: encoding what pitch onset (from C0 to G10) this time step has.
– Hold: encoding that indicates if this time step continues with the last pitch.
– Rest: encoding that indicates if this time step is a rest (silence).

Next we feed the tokenized data into our VAE model. The encoder is composed of an embedding layer, two bidirectional-GRU recurrent layers, and a linear layer for constructing the mean and the logarithmic standard deviation of the latent distribution parameters, in accordance to the common training paradigm of VAEs. The decoder has a hierarchical structure with two sub-decoders. The first is a beat sub-decoder that extracts the latent variable from the encoder into a short hidden state sequence. Every hidden state is analogous to single beat in music. The second tick sub-decoder then disassembles each beat's hidden state into tick sequence in time, which consists of the pitch, hold, and rest tokens. Both sub-decoders comprise of GRU layers and several linear layers. This hierarchical decoder is considered to be a representation of the structural prior in the VAE model, where the beat sub-decoder first outputs some large note groups, and the

tick sub-decoder further outputs each note by decoding these groups. By training the model, the parameters of the neural networks are tuned so as to optimize the efficiency and accuracy of the decoding. The output of the tick sub-decoder is regarded as the final output, with the reconstruction loss function being the multi-cross entropy between the input tokens and the output predictive tokens. Finally, the complete VAE training loss function consists of the reconstruction loss and the KL divergence loss between the latent state distribution and the assumed prior.

As depicted in Figure 2, a sequence of music melody (e.g. one or two music measures) is tokenized into a sequence of integers. In the example of Figure 2, the numbers 0-127 represent the pitch onset tokens from C0 to G10. The number 128 denotes the hold token, and 129 (nonexistent in the Figure 2) denotes the rest token. Then, this melodic sequence is sent into an embedding layer by converting 129 numbers into latent embeddings. This embedding sequence is fed into two bidirectional-GRU layers and output the final hidden state vector. The last part of the encoder is a linear layer (or two linear layers) to map the final hidden state vector into the mean and the logarithmic standard deviation of the latent distribution parameters. The latent variable z is sampled from this distribution.

In the decoder, the first beat sub-decoder is composed of three linear layers and one Beat GRU layer. As shown in Figure 2, we use the Scaled Exponential Linear Unit (SELU) as the activation function after the first linear layer in the decoder and two entry linear layers in the beat sub-decoder:

$$SELU(x) = \begin{cases} \lambda x, x > 0 \\ \lambda\alpha e^x - \lambda\alpha, x \leq 0 \end{cases} \tag{4}$$

where $\alpha \approx 1.673, \lambda \approx 1.051$. The latent variable z is fed into the first linear layer and taken as the initial value of the Beat GRU. The Beat GRU outputs each hidden state vector $b_1, b_2, ..., b_n$ as the initial state for each beat in the music measure. Each beat state is further mapped into two types of hidden state h and c. Then, the second tick sub-decoder utilizes each pair of (h, c) in a Tick GRU. Each beat state pair (h, c) generates t tokens within the timeline, where h is the initial value and c is the conditional value of the Tick GRU in each beat state. The final output x is a discrete symbol in each time state, obtained by a fully-connected layer to map the output of Tick GRU into a probability vector and taking the argmax[2] of it. The cross-entropy loss is computed between the original ground truth (one-hot vector) and the probability vector to train the VAE model.

In this paper, we use the number of beats $n = 8$, and the number of ticks in each beat $t = 6$ to represent two monophonic music measure. The latent dimensional size is 10, which we use to convert the melody tokens.

3.2 VAE for General Polyphonic Music

Finding a principled latent representation of polyphonic music is an open problem in music research, with several works proposing multi-track music encoding,

[2] find the index where contains the maximum probability value.

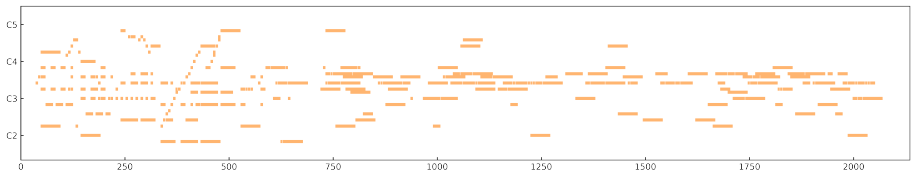


Fig. 3: An example of pianoroll. The x-axis is the timeline and the y-axis is the pitch. Orange pieces denote the notes appearing within the timeline.

decoding and a combined toolbox (Z. Wang, Zhang, et al., 2020; Z. Wang, Chen, et al., 2020; H. Dong, Chen, McAuley, & Berg-Kirkpatrick, 2020). In this paper, we use a simple piano-roll format to construct a vanilla VAE model for encoding and decoding polyphonic music. The piano-roll format for polyphonic music consists of a matrix where the x-axis (row) corresponds to time step and the y-axis (column) is the pitch map, with a binary value for each pitch class denoting presence or lack of note activation at that time instance:

$$Pianoroll(t, p) = c \in \{0, 1\} \tag{5}$$

Figure 3 shows a piano-roll example of polyphonic music.

As shown in Figure 4, our proposed polyphonic music VAE encoder simply consists of two linear layers. We first flattens the piano-roll matrix into a sequence of length $T \times P$ (T is the total number of time steps, and P is the total number of pitches). Each P-step group is a multi-hot vector for each time step. We implement a linear layer to map the input sequence to an integrated latent vector. This vector is similar to the final state output of the GRU layer in the monophonic VAE architecture. Then, we implement another linear layer to map the integrated latent vector into the mean and the standard deviation of the latent VAE distribution. Similarly, the decoder comprises two linear layers that map the latent variable back into an output piano-roll matrix. The reconstruction loss function is the binary-cross entropy loss between the input matrix and the output matrix. Following the standard VAE method, the total training loss constitutes of the piano-roll reconstruction loss and a KL divergence loss between the latent distribution and a Gaussian prior.

There are two reasons for us to use the linear layer to process the pianoroll instead of recurrent layers or convolutional layers. First, different from monophonic music, polyphonic music possibly contains more than one note in many time steps. This requires us to predict the probability of every note in every time step. In the monophonic VAE, the recurrent layers only process sequences of length T, while in the polyphonic VAE, the sequence length is increased to $T \times P$, so the training speed would be extremely low and the GPU memory cost would be unaffordable. As for the convolutional layers, it is hard for them to process such a long sequence since they only contain a limited perceptive field to capture each segment of the sequence. Second, if we directly process the multi-hot pianoroll vectors into the model without flatting the pianoroll, the

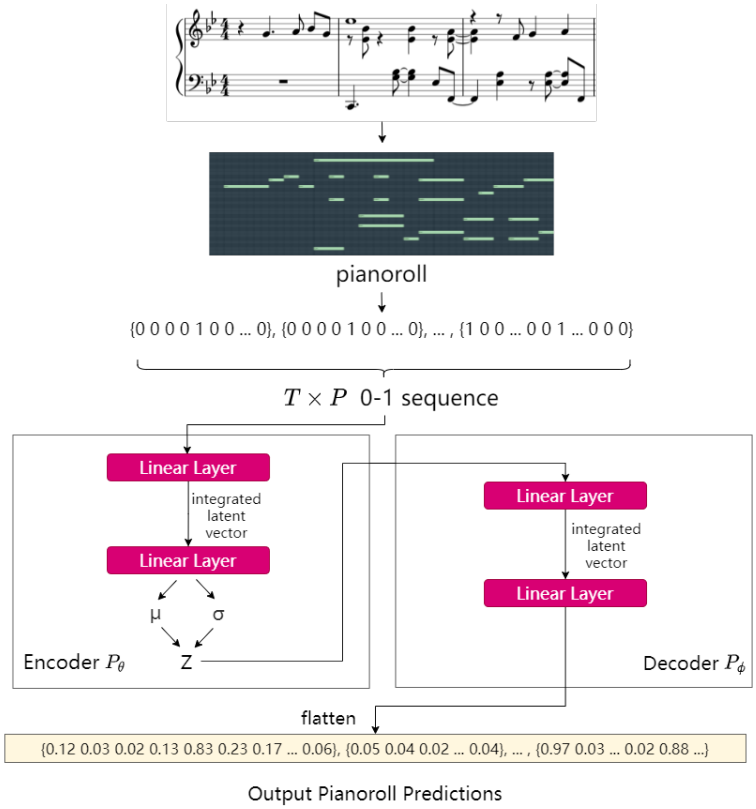


Fig. 4: The network architecture of polyphonic music Variational Autoencoder.

final output of each time step of the decoder is also a multi-hot vector. In that, we need to implement another component to predict the number of notes in each time step. This adds more parameters into the model in comparison to the monophonic VAE architecture.

Therefore, using linear layers to compose the encoder and the decoder of the polyphonic VAE is a trade-off and an easy-to-implement method. This allows the implementation fo models using similar computing resource consumption for both the monophonic and polyphonic VAE. In the experiments, we treat the polyphonic music and the monophonic music as two cases with different settings. The difference in VAE structure will not affect our findings. Currently, the construction of an efficient polyphonic VAE is still a challenging topic in research. Some works (Z. Wang, Zhang, et al., 2020; Z. Wang, Chen, et al., 2020) propose different data and network structures to achieve a higher accuracy in reconstructing the music from the latent variable. In the future work, we hope to construct a more advanced structure to process polyphonic music using a VAE.

3.3 VAE for Audio

The high dimensional and continuous nature of audio signals creates a challenge for finding an efficient latent representation for music. Several works have utilized VAE models for exploring audio representations, including VAE models for finding disentangled audio representations(Luo, Agres, & Herremans, 2019) and VAE models for modeling audio containing speech (Hsu, Zhang, & Glass, 2017). Apart from VAE, simple convolutional models are used in different music tasks such as music recommendation (Chen, Liang, Ma, & Gu, 2021) and source separation (Chen, Du, et al., 2021). In our experiments for audio representations, we chose to use a convolutional VAE to map the short-time Fourier transform (STFT) representation of audio into a lower dimensional latent representation. The input audio signal is sampled at 22050Hz and then preprocessed using STFT with a window length of 92 ms and a hop length of 12 ms. The spectrogram is split into short, fixed-length time frames with duration 279 ms resulting in an input representation with 1024 frequency bins and 24 time bins.

The encoder and decoder of our VAE have a symmetrical structure consisting of convolutional layers and linear layers. Full details for the architecture parameters are available in Appendix B. The encoder is constructed of five 2D convolution layers with max pooling layers located after the third convolution layer, and after the last convolution layer. The final layer is flattened and processed through two linear layers to obtain a mean and a logarithmic variance. From the mean and logarithmic variance, the latent variable is sampled from a normal distribution. The decoder nearly mirrors the encoder architecture. The latent variable is first passed through a linear layer to obtain the correct shape when unflattened. Following are five 2D transpose convolution layers and two upsampling layers to restore the shape of the spectrogram.

The reconstruction loss function is the mean-squared error between the input and output spectrogram, and the whole VAE training loss is constituted by the reconstruction loss and the KL divergence loss. Since mean-squared error does not accurately portray perceptive audio reconstruction quality, the reconstruction quality of audio is evaluated using log spectral distortion between the two power spectrograms S and $\hat{S}$:

$$LSD = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(S_{ti} - \hat{S}_{ti})^2} \tag{6}$$

4 Reduced Representation

The idea of musical reduction is known from various theories of music analysis, such as Schenkerian analysis (Marsden, 2010) and later generative grammar approaches like GTTM(Lerdahl & Jackendoff, 1996). These theories consider deeper musical structures, somewhat related to musical theory concept of *Ursatz* that comprises a background abstract layer from which the musical foreground

surface[3] emerges. It should be noted that Ursatz is not really predictive and has not been proven to represent music cognition, much like many of the formal grammar theories that are still arguably non-representative of true human cognitive faculties. Nevertheless, it can still considered as a prototype for music reductionist analysis. The GTTM, which is derived from Schenkerian analysis, is a generative model by definition, suggestive of another conceptual parallel to our methods. Of course these techniques are very different from the methods developed in this paper and are brought here only for motivation purposes. As we will observe later (section 7.1), both musical surface and highly compressed representations contain little predictive power, while maximal predictability happens in the mid-range of reduction quality. A different aspect of reduced representation that served as motivation, which is closer to aspects of human cognition is the study of Rate–Distortion as a way of extracting useful or meaningful information from noisy signals (R.Sims, 2016). The idea of reduced representation also has been recently explored in the context of representation learning in deep neural networks using a framework known as Information Bottleneck (Tishby & Zaslavsky, 2015).

The special case of music is that the deeper background structure serves not only to compress, but also to predict musical continuations, or in other words, the concept of *Ursatz* introduces the idea that longer musical progressions or temporal relations, which are hard to perceive on the surface, can be more evidently revealed from the background. Accordingly, the question of reduction in music is closely related to the task of prediction. In deep learning some attempts to consider predictive information through use of a bottleneck or noisy representation in temporal models such as RNNs have recently appeared in the literature(Alemi, 2019),(Z. Dong, Oktay, Poole, & Alemi, 2020) An important distinction between these works and the proposed framework is that we are not introducing rate-limitation or adding noise to the latent codes during the learning process, but rather we are applying bit-reduction to a pre-trained encoder-decoder network in order to reduce the complexity of the latent representation prior to decoding. This allows experimenting with various bit-rate regimes without retraining the network each time. For this purpose we borrow a technique of bit-allocation from Rate–Distortion theory of lossy information processing. The reduced latent representation streams will be later subject to predictive analysis to consider their temporal significance.

4.1 Bit-rate limited encoding

As mentioned above, in this study we consider a particular case of reduced representation that is based on Rate–Distortion theory. Rate–Distortion theory offers an optimal solution for finding the most compact (least rate) encoding for a given limit on the distortion or reconstruction error. Equivalently, Distortion–Rate finds the best encoding in terms of least distortion for a given rate. Algorithm for optimal bit allocation according to the Rate–Distortion theory are so called

[3] Musical surface is musical data that contains the actual notes and sounds.

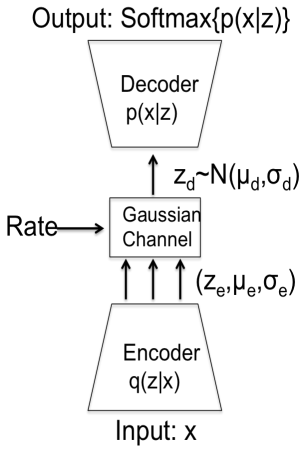


Fig. 5: Noisy channel between encoder and decoder

bit-allocation methods that we describe below. By using a rate-limited channel between encoder and decoder of the VAE we are able to control the complexity of the encodings using a bit-allocation algorithm. In our case, we use Rate as the free parameter to find the least Distortion codes, under the assumption that latent codes in VAE are distributed as multi-variate uncorrelated Gaussians. The Rate–Distortion function that provides the lower limit on the achievable rate R as function of the maximal allowed distortion level D, is given by

$$R(D) = \begin{cases} \frac{1}{2} log_2 \frac{\sigma^2}{D}, & \text{if } 0 \leq D \leq \sigma^2 \\ 0, & \text{if } D > \sigma^2. \end{cases} \tag{7}$$

Where R is the rate, and D is the distortion value. This Rate–Distortion function can be converted into a Distortion–Rate function $D(R) = \sigma^2 2^{-2R}$ that gives the lower limit on distortion D that is achievable for a given rate R. This ideal lower limit (i.e. least distortion) can be efficiently achieved for a particular type of signals that is known in communication theory as "multivariate Gaussian channel". We adopt this channel model for our experiments without further justification. What is special about this type of channel is that an optimal bit-reduction can be achieved by using the so called reverse water-filling algorithm (Cover & Thomas, 2006). This algorithm starts with a predefined bit-rate R, and successively allocates one bit at a time to the strongest component, repeating the process until all bits in the bit-pool are exhausted. One should note that channels (i.e. latent variables in our case) with variance less then allowed distortion, or channels that run out of bits for a given rate, are given zero bits and thus are eliminated from the transmission.

Schematic representation of the channel inclusion in the auto-encoder architecture is given by Figure 5. Encoding the latent components at a rate R changes

the mean and variance of the VAE as follows (Berger, 1971)

$$Q(z_d|z_e) = Normal(\mu_d, \sigma_d^2) \tag{8}$$

$$\mu_d = z_e + 2^{-2R}(\mu_e - z_e) \tag{9}$$

$$\sigma_d^2 = 2^{-4R}(2^{2R} - 1)\sigma_e^2 \tag{10}$$

This process requires some explanation: for a given rate R we obtain the bit-rate for each of the latent variable according to the reverse water-filling procedure. This gives us different number of bits for each latent dimension, where the high variance dimensions tend to grab the bits first, often leaving the weak (small variance) latent variables with zero bits. Next we proceed by sampling a value from the encoder distribution according to the original VAE mean and variance parameters, μ_e and σ_e^2. Then, using the rate R and the original mean and variance parameters for each latent variable, we derive a new mean and variance μ_d and σ_d^2. We use these probability parameters to sample a reduced bit-rate value and use it as our new input to the decoder. One can see that latent variables that are allocated zero bits need not be transmitted[4], while channels allocated a very high rate will transmit an (almost) unaltered value of the latent variable that was sampled in the VAE encoder.

Depending on the specific experiment and data-type, the last softmax layer needs to be processed to determine how it translates into specific notes or spectrogram segments. The details of each representation are discussed in the corresponding experimental section. The VAE-based noisy channel is common to all cases, done during the encoding-decoding part prior to the final data reconstruction step.

5 Estimating the Reduction Quality

In DMID framework we consider the effect of reduced representation on the ability to capture the information in the musical surface represented by a full-rate encoding. By changing the bitrate, we have a theoretical "knob" that changes the fidelity of the representation for the current musical frame. In order to be able to assess the effect of reduced representation across musical voices, or the effect of past reduced representation Z on the future surface Y, we need to employ other statistical methods to estimate the change in mutual information between these variables.

It is important to note that when we talk about full-rate encoding, we still refer to the encoding of the latent representation that can be decoded at acceptable level of quality or precision to reconstruct the musical surface. In mathematical notation, we may write $Z_{full-rate} = Encode(X)$ and $T_{full-rate} =$

[4] More precisely, the value that the decoder needs is mean value of that latent variable that is independent of the particular instance being transmitted. This mean value can be obtained a-priori and thus can be "hard coded" into the decoder ahead of time, with no need to transmit it.

$Encode(Y)$, such that the reconstructions $\hat{X} = Decode(Z_{full-rate})$ and $\hat{Y} = Decode(T_{full-rate})$ are sufficiently similar to the original X and Y. We assume that the pre-trained VAE model actually accomplishes this level of encoding-decoding quality, and then we proceed to reduce the rate of Z, making $\hat{X}$ and $\hat{Y}$ less exact or further away from their original X and Y. Since we assume that the VAE learning is done sufficiently well, in some of the discussion we might refer to the representation of musical surface data X and Y as a full-rate encoding, but it should be clear that all the analyses are done on the latent encoding of musical data, and not on the musical surface itself, and in the case of measuring predictive information by VMO, the analysis is performed on a discrete (symbolized) version of that encoding, as explained in section 6.3.

5.1 Mutual Information Neural Estimation

Theoretically, computing the mutual information between two variables is a hard problem. The computation is only tractable if two variables are discrete or if two variables' probability distributions are known. For the latent variables sampled from a VAE, we cannot directly compute the mutual information since it is intractable to marginalize the probabilities from the whole latent space.

Mutual Information Neural Estimation (MINE) is a framework that allows us to use a neural network output to approximate the mutual information between two variables. Suppose that we want to estimate the mutual information between x and z, we first construct the joint datapoints $(x_1, z_1), (x_2, z_2), ..., (x_n, z_n)$, where (x_i, z_i) is the tuple we hypothesize these two variables are related. Then, we create a new group of datapoints $(x_1, z_{a_1}), (x_2, z_{a_2}), ..., (x_n, z_{a_n})$, where $(a_1, a_2, ..., a_n)$ is a randomly shuffled index sequence from 1 to n. Next, we feed these two groups of data points into a neural network T_θ to converge by a loss function:

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} T_\theta(x_i, z_i) - log(\frac{1}{n}\sum_{i=1}^{n} e^{T_\theta(x_i, z_{a_i})}) \tag{11}$$

where minimizing this loss is to find a tighter lower bound that can approximate the true mutual information, by Donsker-Varadhan representation of the KL divergence (Donsker & Varadhan, 1983). Generally, the mutual information estimation is done by finding a mapping between two sets of datapoints, so it is a "point-wise" mutual information that is not easily applicable to time-series or predictive information. In the following experiments, we will use MINE to to analyze these latent variables in different cases.

5.2 Bit-Rate Reduction versus VMO Symbolization

It should be noted that in our experiment we apply two separate steps of information reduction:

1. Bit-rate reduction using bit-allocation techniques applied to the latent encodings of the VAE.

2. Symbolization step in VMO that is used for finding repetitions of variable length

The symbolization step, which can be considered conceptually as something similar to Vector Quantization, is needed for technical reasons, in order to use FO that is a string matching method, to find repeated sub-sequences of various length. To our best knowledge, there are no other existing method that performs symbolization based on maximization of Information Rate[5]. In application of MINE, no symbolization step is required. Estimating predictive information using symbolization with variable memory oracle is the topic of the next section.

6 Predictive Information in Variable Memory Models

Predictive Information refers to a measure of mutual information between two random variables, one representing the past and one representing the present (or the future) of some random process. In our formulation, we have two or more parallel streams of such random variables. One stream is the musical surface, where predictive information $I(X, Y)$ is measured between two time instances X and Y of the musical score or audio recording. We can also apply predictive information to sequences of latent states, wanting to measure the amount of information passed over time in the reduced representation stream $I(Z, T)$ between past Z and future T.

6.1 From Sound (or Musical Score) to Information

One of the difficulties in estimation of predictive information is that the mutual information for the next MIDI measure or the next audio frame based on its immediately preceding measure or audio frame carries too many details to be able to see the large picture. Going from signal to symbols, or changing the representation from a complex musical score to a smaller set of representative symbols or tokens is a common practice in time series modeling. Moreover, from the temporal modeling aspect of how the future depends on the past, which is a standard Markov assumption for random processes, the fixed memory assumption is too limited to be able to capture the long term dependencies that exist in music. It is possible of course to capture temporal relations in sequences of continuous measurement in real valued signals or series of feature vectors using auto-regressive models such as Dynamic Texture Models based on Kalman filters (Barrington, Chan, & Lanckriet, 2009), Deep Kalman, or Recurrent Neural Network models with latent embedding (Fraccaro, Sønderby, Paquet, & Winther, 2016). In our work we pursue an alternative route of symbolization prior to information modeling, or in other words, we turn the complex polyphonic midi or the features extracted from an audio signal into continuous vectors first, and then

[5] In other words, what we are seeking symbolic representations that are most informative in time and thus are best at "explaining out" the Latent Information Dynamics of music

group them into a finite set of representative labels or symbols to allow finding the variable length motifs or sub-sequences of different durations. We know effectively that music has a long memory and that phrases or structural musical elements extend beyond the immediate past. In order to take into account partial repetitions that may occur in the musical piece, we explored the use of Factor Oracle (FO) (Allauzen, Crochemore, & Raffinot, 1999), which constructs an automaton-like structure that links forward positions along a sequence to find its factors.

Note 1. The term "factor" is used in FO in a very specific manner to denote any sub-sequence of the original sequence. Due to ambiguity of the term "factor" the reader should not confuse this with the general use of the word in other parts of this paper., which is a term used for sub-sequences.

For the sake of completeness, we will briefly describe the FO method below. For a full account the reader is referred to (Dubnov et al., 2011). One should note that in variable Markov Oracle (VMO) (C. Wang & Dubnov, 2014), which is an extension of FO to real-valued sequences of vectors (time series), the symbolization step is done during the sequential modeling by setting up a threshold for similarity between the time series data. As such, VMO symbolization is done by taking into account the predictive aspects of the discretized sequence, choosing the best predictive version as measured by its Information Rate. We will discuss this below, as finding the best symbolization is critical to capturing the information dynamics of the musical data. To sum up the motivation behind use of VMO, the following steps are outlined:

- Initial representation of music or audio as feature vectors or multi-hot midi encoding as a piano-roll
- Latent embedding of the initial input representation using VAE
- Reduced representation of the latent vectors using bit-allocation
- symbolisation and long-term memory modeling using VMO

6.2 Factor Oracle model for repeated sequence discovery

An example of FO structure, in the format of VMO, for a sequence of labeled states is shown in Figure 6. More examples and details for this structure is demonstrated in (C. Wang & Dubnov, 2014). The meaning of the labels depends on the type of data used to represent music, for instance labels can represent clusters containing similar audio features, or clusters of similar complex note combinations in a MIDI file. Finding the right set of labels for representation of complex MIDI data or audio signal is one of the critical aspects of this analysis, which will be described in the next section where we talk about estimation of Music Information Dynamics.

Given a sequence of labels, forward links are used to retrieve any of the subsequences from X. An oracle structure has two types of forward links; the first is an internal forward link which is a pointer from state $t-1$ to t labeled by the symbol x_t, denoted as $\delta(t-1, x_t) = t$. The other forward link is an external

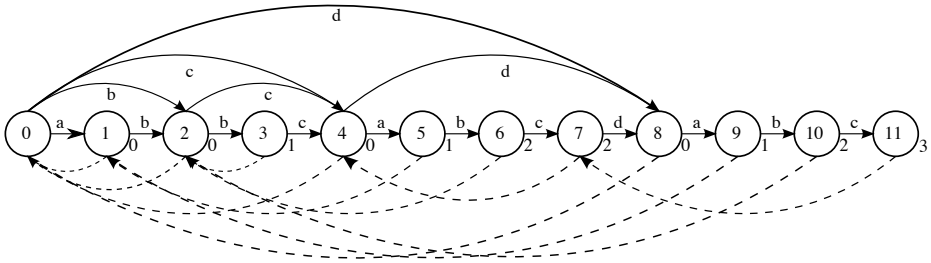


Fig. 6: A *VMO* structure with symbolized signal $\{a, b, b, c, a, b, c, d, a, b, c\}$, upper (normal) arrows represent forward links with labels for each frame and lower (dashed) are suffix links. Values outside of each circle are the lrs value for each state.

forward link which is a pointer from state t to $t+k$ labeled by x_{t+k} with $k > 1$. An external forward link $\delta(t, x_{t+k}) = t + k$ is created when

$$\begin{aligned} x_{t+1} &\neq x_{t+k} \\ x_t &= x_{t+k-1} \\ \delta(t, x_{t+k}) &= \emptyset. \end{aligned}$$

In other words, an external forward link is created between x_t and x_{t+k} when the sequence of symbols $\{x_{t+k-1}, x_{t+k}\}$ is first seen in X with x_t and x_{t+k-1} sharing the same label. In our IR estimation, and in the machine improvisation application, the forward links are not used and are provided here for the sake of completeness being part of the Factor Oracle automaton[6].

An oracle structure carries two kinds of links, forward links and suffix links. A suffix link is a backward pointer that links state t to k with $t > k$, without a label and is denoted by $sfx[t] = k$.

$$\begin{aligned} sfx[t] = k \iff &\text{the longest repeated suffix of} \\ &\{x_1, x_2, \ldots, x_t\} \text{ is recognized in } k. \end{aligned}$$

Suffix links are used to find the longest repeated suffix in X. In order to track the longest repeated suffix at each time index t, the length of the longest repeated suffix at each state t is computed by the algorithm described in (Lefebvre, Lecroq, & Alexandre, 2003) and is denoted by $lrs[t]$. The parameter lrs is part of the on-line construction algorithm of the oracle automaton (Lefebvre et al., 2003).

6.3 Estimation of Music Information Dynamics using VMO

Variable Markov Oracle (VMO) is a machine improvisation method based on the Factor Oracle (FO) string matching algorithm. The VMO suffix tree data structure allows for various media-data generation (C. Wang & Dubnov, 2014). In

[6] The function of the forward links is retrieval of factors of X, searching from the beginning of a string X following the forward links path

order to operate on complex musical or audio features, as well as any other multivariate time series data, VMO quantizes a signal $x_1^T = x_1, x_2, \ldots, x_t, \ldots, x_T$, into a symbolic sequence $s_1^T = s_1, s_2, \ldots, s_t, \ldots, s_T$, over a finite alphabet $s \in S$. Relating this to the deep information dynamics model, we denote $X = x_{past} = x_1^{T-1}$ and $Y = x_{present} = x_T$. The computation of Information Dynamics is done using the Information Rate (IR) measure $I(X, Y)$ of mutual information between past and present of the musical signal. The novelty of the current approach compared to previous VMO application is that it is being applied for the first time to a sequence of latent states from a machine learned representation instead of standard human-engineered features. Moreover, using the bit-rate reduction step, we use VMO to estimate $I(Z, T)$ for a reduced representations.

IR is estimated by extending the FO algorithm to real-valued sequences, which in our case are latent states from VAE, with similarity being accounted up to a given threshold. Changing the threshold from low to high can be effectively thought of as symbolization that is done at different quantization levels from fine to coarse.

At each threshold value, a different oracle graph is estimated by the FO algorithm. For each such oracle, a compression method of Compror (Compression Oracle) algorithm C is used as an estimator of $I(X, Y) = H(Y) - H(Y|X) \approx C(Y) - C(Y|X)$. Here the entropy H is substituted with string compression algorithm C, and $C(Y) = log_2(|S|)$ is taken as the number of encoding bits for individual symbols over alphabet S, and $C(Y|X)$ is the number of bits in a block-wise encoding that recursively points to repeated sub-sequences (C. Wang & Dubnov, 2015a).

As mentioned above, the essential step for symbolization in VMO is finding a threshold value, θ, that effectively partitions the latent variables into discrete categories. The threshold θ is used to determine if the incoming x_t is similar to one of the frames following the suffix link started at $t - 1$. VMO assigns two frames, x_i and x_j, the same label $s_i = s_j \in \Sigma$ if $||x_i - x_j|| \leq \theta$. In (C. Wang & Dubnov, 2015b), VMO was introduced as an efficient method for estimating IR. At each bit-rate (which will be specified in section 7), we apply VMO analysis to estimate the information dynamics of that representation. This gives us an estimated value of $I(Z, T)$ at different bit-rates, as discussed in the experiments section.

To summarize, VMO algorithm comprises of a combination of two innovations:

- Extension of the FO algorithm that operates on discrete sequences with precise matching to an approximate matching up to a given threshold
- Use of Compror to estimate the information rate for that threshold

By using an exhaustive search over possible threshold values, the one with higher information rate is retained as the best representation. As mentioned in the introduction, one of the advantages of using VMO for mutual information estimation is that it allows instantaneous time-varying estimates of IR based on the local information gain of encoding a signal based on linking it to its similar past. This

differs from other methods of mutual information estimation like MINE that averages over the whole signal.

6.4 Estimation of Latent-Predictive Information

In order to estimate the reconstruction quality of musical surface from past encoding, we need to estimate the mutual information between past embedded Z and the future musical surface Y, $I(Z,Y)$. For convenience, we will call this *Latent-Predictive Information*, to distinguish it from Predictive Information that is measured within the same data stream. So we want to measure the ability to predict one stream of data, in our case the surface, from another stream of data, in our case the past embedding. This notion of Latent-Predictive Information bears some resemblance to notions of Transfer Entropy (Granger, 1969) that is measuring the amount of directed (time-asymmetric) transfer of information between two random processes, also related to notions of Granger Causality (Schreiber, 2000).

In our method, during the first encoding step we already minimized a related quality $I(X,Z)$ (and $I(Y,T)$ as well) by training a VAE. This process of learning the latent codes was done in the training step by optimization of instantaneous representation quality. This encoding is further modified using bit-allocation, reducing the information between the surface data and its latent encoding. This embedding gives us two different random variables Z and Y (or X and T, respectively). For practical purposes, in the experiments we consider $Y = Z_{full-rate}$, and $Z = Z_{limited-rate}$ by doing the two steps of VAE encoding, followed by bit-rate reduction.

To summarize, the Deep Information Dynamics model introduced the concept of predictive-reconstruction surprisal $I(Z,T|Y) = I(Z,T) - I(Z,Y)$ comprising of the difference between prediction quality of the next element in the embedding space, as measured by IR of the sequence of latent variables, and the reconstruction quality of the future musical surface from previous latent state. It should be noted that in terms of mathematical definitions of mutual information, both factors require averaging over complete set of latent and surface data values, resulting in a single number of summary statistic that is time independent. We use MINE (Mutual Information Neural Estimation) (Belghazi et al., 2018) for estimating $I(Z,Y)$ (or in practice $I(Z_{full-rate}, Z_{limited-rate})$. In contrast to VMO, the MINE method averages over the whole set of data points in Z and Y, resulting in a single number. Accordingly, the surprisal experiments reported in the experimental section comprise two different types of analysis:

1. time varying IR profile showing predictive information of latent encodings at different bit allocation rates
2. a constant number representing an average predictive quality of a future surface frame from past bit-rate limited latent encoding

Since VMO and MINE are based on very different algorithms, the units of IR and encoding MINE estimates are not compatible, so we analyze them separately.

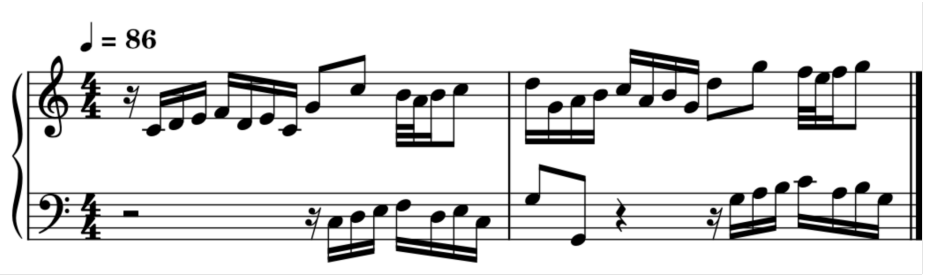


Fig. 7: The example of Bach Invention No.1. Notice that it only shows two bars, so the terminating bar is not the true end.

7 Experiments

We apply the DMID framework to three music cases: (1) polyphonic music, (2) monophonic counterpoint music, and (3) music waveforms from recorded or synthesized audio. Specifically, in each case, we use the corresponding variational autoencoder model to obtain the latent variables of musical pieces, and then use bit-allocation to control the rate of information that can be transmitted. By controlling the rate of the bit-allocation, we can get the latent variables that contain different amounts of original information. Furthermore, we use VMO and mutual information neural estimation (MINE) to analyze different relations in time and across representation levels of these latent variables.

7.1 Monophonic and Two Voices Counterpoint Case

In this experiment we test the DMID formalism in two settings: same voice prediction versus across voices prediction. An interesting aspect of using reduced representations of musical surface is its ability to reveal relations between parallel voices, such as in the case of musical counterpoint. Each voice is separately analyzed in terms of its own representation and prediction structure. Next, we explore the question of whether reduced representations reveal latent structures and relations between the two voices. The motivation for this experiment is relatively simple – in many musical voice leading practices, the voices are related through some latent aspect, such as belonging to the same harmony that extends the structure further then immediate consonance/dissonance relations between the parallel notes. Accordingly, we express whether higher levels of mutual information can be found between reduced representations rather than within the surface itself, and whether the melodic logic of one voice is influenced by another voice. This experiment, although in part deviating from the strict formalism of DMID, is suggested as an additional insight into using bit-rate reduction to reveal musical structure, both within and across voices.

In this experiment, we use the dataset of 14 two-part inventions composed by Johann Sebastian Bach. The MIDI files are collected from the *Complete Bach*

MIDI Index[7]. All files are accurately quantized according to the original piano scores at 120 BPM (beat per minute). In each invention, there are two voices consisting of monophonic music. Therefore, we use the hierarchical VAE from section 3.1 to encode and decode each voice of each invention. Figure 7 shows the first two bars in Bach's Invention No.1.

We split the dataset into 10 training pieces and 4 testing pieces. To augment the data, we shift the key of each training invention into its neighbouring 8 tonalities (i.e., moving up four tonalities and moving down four tonalities). We did not shift for all 12 tonalities to avoid the extreme high and low notes in the shifted music pieces. As a result, 10827 bars (measures) are used for training and 1485 bars are used for testing. In terms of the choice of training hyperparameters, we set the batch size to 128, learning rate to 1e-4 (with Adam Optimizer (Kingma & Ba, 2015)), beat number to 8, and the tick number per beat to 6, so the total number of time steps is 48 ($\frac{1}{3}$ corresponding to demisemiquaver to correctly handle triplets). We use early stopping to find the best reconstruction accuracy in the training set. The final trained model performs at 99.07% average reconstruction accuracy between the input tokens and the output tokens in the testing set. This accuracy is the same as the state of the art performance achieved in (Pati et al., 2019; Chen, Wang, et al., 2020).

After completing the initial encoding, we use bit-allocation to control the rate of the information that is transmitted from the original latent variables to a reduced representation stream. For each voice of every test invention, we obtain a latent variable representing the voice melody of one bar. We set the bit-rate to be 10, 100, 1000, and 10000 to adjust the amount of information reduction relative to the original latent variables. Then, we decode these allocated latent variables back into a music surface, creating new reconstructed pieces. From Table 1, we can see that with bit-allocation, latent variables lose different amounts of information according to the bit-rate. When the rate is 10, the reconstruction accuracy drops to the lowest value of 87.07%. As we increase the rate, more information will be transmitted through the channel, thus the adjusted latent variables approach the original latent variables. The accuracy increases from 87.07% to 99.07%. When the rate is 10000, the adjusted latent variables are much more similar to the original ones. The resulting accuracy is the same as that of the original latent variables.

After experimenting with adjusting the latent variables according to different rates, we try to figure out the predictive ability of these latent variable in three generative scenarios:

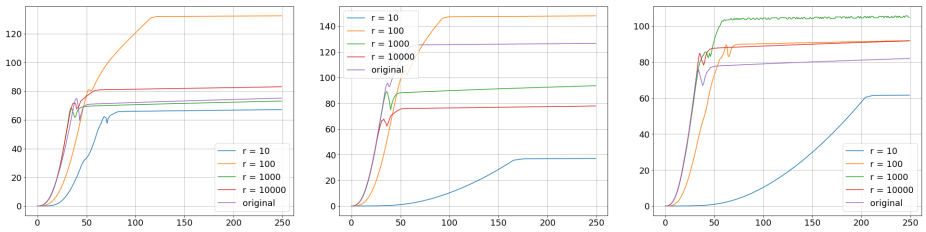
(1) Using the past measure to predict the future measure;
(2) Using the first voice measure to predict the second voice measure (in the same position);
(3) Using the second voice measure to predict the first voice measure.

The reason for us to select these three scenarios is because musically, the adjacent measures share many progressive relationships, and in the counterpoint

[7] http://www.bachcentral.com/midiindexcomplete.html

Table 1: The reconstruction accuracy by controlling the rate of the bit-allocation.

Rate	**10**	**100**	**1000**	**10000**	**Original**
Reconstruction Acc.	87.07%	90.75%	97.68%	99.07%	99.07%



(a) The rate and mutual information training results for the (1) scenario (past-future).

(b) The rate and mutual information training results for the (2) scenario (1st voice-2nd voice).

(c) The rate and mutual information training results for the (3) scenario (2nd voice-1st voice).

Fig. 8: The training results of different MINE models for different bit-allocation rates. The x-axis denotes the training epoches. And the y-axis denotes the mutual information value.

compositions, the first and the second melody in the same measure share many dependencies. The mutual information between these measure pairs are important for us to determine if the model captures the music information dynamics along with the music development. In detail, each conditional latent variable is an adjusted one (i.e. after applying bit-allocation), and each predicted latent variable is an original one. We group every two latent variables into a two-tuple (z_{con}, z_{pred}). For example, in the first scenario, z_{con} is an adjusted latent variable for one musical measure, and z_{pred} is an original latent variable for the following measure. In the second scenario, z_{con} is an adjusted latent variable for single first voice musical measure, and z_{pred} is an original latent variable for the second voice in the same position. We applied the Mutual Information Neural Estimation (MINE) to estimate the mutual information between z_{con} and z_{pred}, conducting four experiments by controlling the rate to 10, 100, 1000, and 10000. As shown in Figure 8, all four MINE networks have converged with corresponding tuple data by achieving a stable mutual information value after several epochs of training. All models except the one of rate=10 have an information rate drop around the 25th-30th epoch but return to increase around the 45th epoch. In total, for the three scenarios we show $3 \times 4 = 12$ experimental results. In the appendix, we show an example of four reconstructions for the first four bars of Bach Invention No.10, where we change the bit-allocation rate from 10 to 10000 to generate the first voice, while the second voice remains the same (the original one).

Figure 8 and Table 2 show the experimental results for all the scenarios. The x-axis of Figure 8 is the epoch trained by the MINE network, where we can see

Table 2: The experiment results of controlling the rate to measure the mutual information between the conditional latent variables and the predictive latent variables. The first column shows different predictive scenarios. The other columns show the mutual information at different rates.

Scenario	R=10	R=100	R=1000	R=10000	Original
past-future	67.14	132.42	73.09	83.01	75.12
1st voice-2nd voice	36.96	148.05	93.64	77.85	126.63
2nd voice-1st voice	61.64	91.89	104.92	91.82	82.04

that all models have converged. The y-axis is the mutual information value by bits at different rates. In the first scenario (past - future), we can see that when the rate is 100, the mutual information achieves the highest value (132.42 bits). When the rate is 10000, the mutual information achieves the second highest value (83.01 bits). In the second scenario (first voice - second voice), we can see that a rate of 100 and a rate of 1000 achieve the top and the second highest mutual information values (148.05 and 93.64 bits, respectively). In the third scenario (second voice - first voice), we can see that a rate of 1000 and a rate of 100 achieve the top and the second highest mutual information values (104.92 and 91.89 bits, respectively), while a rate of 10000 also achieves a competitive value (91.82 bits).

From the results, we conclude that reducing the representation quality by lowering the latent bit-allocation can effectively improve the predictive information (lower the surprisal) of each voice, as well as prediction across voices. Mathematically, we showed that the level of mutual information between conditional latent variables and predictive latent variables depends on the level of reduction in the latent representation. Although the mutual information peaks at different rates for each scenario, there seems to be an overall preference towards a mid bit-rate range of 100-1000. Compared with the direct prediction using the original variables as conditions, the bitrate-reduced latent variables seem to contain more information with respect to the predictive target. Therefore, a reduced representation may serve as a better choice for some structural design in terms of temporal conditions or voice relations in predictive/generative models (i.e., we could use the reduced representation as a condition to generate new music pieces). We also find that when the rate is 100 or 1000, the reconstruction accuracy drops to 90.75% and 97.68%, respectively. This means that we have to sacrifice some of the reconstruction accuracy in exchange for better predictive ability. This trade-off requires a more meaningful study that we plan to research in the future. From the perspective of music analysis, these results suggest that different structural representation and different levels of abstraction that are used in various music analysis and music composition techniques, may have an important effect on the type of temporal or multi-voice predictive quality of music.

7.2 Polyphonic Music Case

In this section we apply deep information rate analysis to a MIDI file of the Prelude and Fugue No. 2 in C Minor, BWV 847, by J.S. Bach. The piece was chosen due to the different nature of its musical texture in the Prelude and Fugue and availability of musical analysis. The experiments qualitatively explore the relation between complexity of the encoding and its predictive properties, for different bit-rates of the encoding representation, and the human analysis.

In the first step of the experiment, we trained a VAE on a set of MIDI files from the LABROSA[8] training set, containing piano performances in the classical music genre. Our VAE model had an input layer that corresponded to one bar (16 steps) of polyphonic music. The input encoding was done as multi-hot representation, with 128 note numbers (pitches) and duration captured into a single vector by a toolkit *midiToNoteStateMatrix*[9]. The VAE had a single hidden layer of size 500 [10]. The trained weights of the VAE encoder were saved for later use, and the encoder part of VAE was used to create the representation for the next step of music analysis.

In the second step of the experiment, we input into the network the J.S. Bach's Prelude and Fugue and used the bit-allocation method to reduce the representation complexity of the hidden layer in VAE encoding. Next, VMO was used to for symbolization of the sequence of the encodings throughout the whole piece. The symbolisation step is technically needed to be able to apply the string matching methods of the Factor-Oracle algorithm that is part of the VMO package and method, as described in section 6.3.

VMO analysis allows estimating the instantaneous information rate along the piece.

Figure 9 shows the VMO-estimation of the instantaneous information rate of the latent representation $I(Z,T)$ at different bit-rates. As a visual reference we provide in the figure musicological analysis of the harmonic structure of the prelude and thematic materials of the Fugue[11] in sub-figure B, and score rendering in sub-figure C. Unfortunately at the time of this article's writing, the page seems not to be available anymore, so we provide just the main findings of that analysis.

Moreover, similar to the analysis method of the two-voice inventions in section 7.1. we used MINE to compute the average predictive mutual information across levels of representations, namely between a bit-rate limited encoding of a single previous bar and a full-rate encoding of the following bar, averaged over the whole piece. This provides an estimate of the $I(Z.Y)$ term, as shown in Figure 10. The top sub-figures A, B, and C show the number of bits allocated

[8] This is a dataset of MIDI data that was collected from the Classical Piano MIDI Page, https://www.labrosa.org/projects/piano/, containing bothe MIDI and synthesized audio from piano recordings using a Yamaha Disklavier piano.

[9] https://www.kaggle.com/chetanmj23/pop-music-collection

[10] The parameters of the VAE were chosen empirically by subjective judging the quality of VAE reconstruction

[11] http://bachwelltemperedclavier.org/pf-c-minor.html

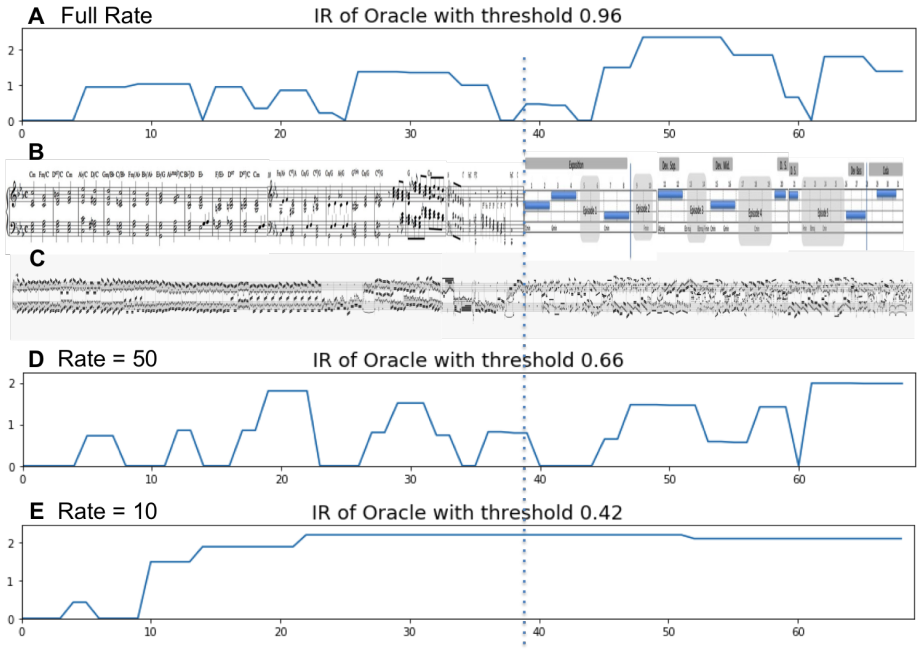


Fig. 9: Analysis of information rate using VMO at A: Full Rate, D: Rate 50, E: Rate 10. Sub-figure B shows harmonic analysis of the Prelude and thematic analysis of the Fugue, Sub-figure C shows the actual score.

to each of the 500 latent dimensions[12]. Due to the nature of the plot, case C looks continuous, but what the line represents is an equal allocations to all latent states, while in cases A and B it can be seen that only some of the latent dimensions are awarded bits, while others are set to zero. These results show that the representation rate is monotonically related to the mutual information it carries towards predicting the next bar at full rate. This is different from the inverted U relation we saw across voices in the previous section 7.1, where the voices where strongly related at a mid-range of the representation bit-rate. This difference in behavior can be explained by the different nature of the experiments. In the two voice invention study a mild reduction in bit-rate reveals more correspondence between the reduced representation of the two voices. In the polyphonic case of Bach's Prelude and Fugue, the analysis indicates that it is actually more difficult to predict the next instance of the musical surface from its reduced past instance, since the reduced past carries less information about the precise musical details to follow. In should be noted that we could not apply the VMO method to measure instantaneous predictive information across levels of representation, since VMO operates on a single data stream over time. De-

[12] A full rate was considered be be a rate of 10,000, which receives an allocation of 20 bit per dimension.

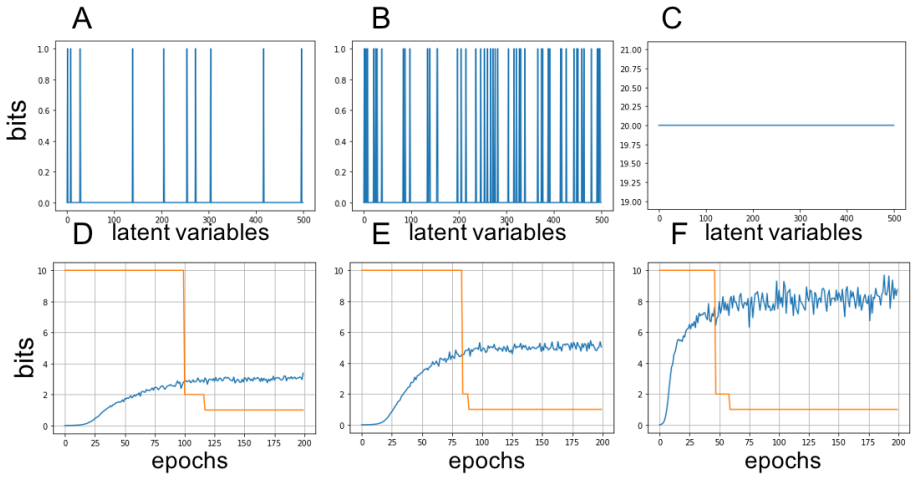


Fig. 10: Estimation of $I(Z, Y)$ that represents the predictive quality of a past bit-limited representation towards the next frame of the full-rate representation. Sub-figures A,B,C show the bit-rate allocation at rates 10, 50, and 10000, respectively. The x-axis corresponds to the 500 latent state variables, with portions showing no-bit allocation basically not being transmitted or accounted for in the latent representation. Sub-figures D,E,F show the MINE estimate as function of the training epochs, for these rates. The orange lines are changes in the learning rate that effectively signal the epoch where the MINE estimator reaches close to its final mutual information estimate value.

veloping computational measures for instantaneous information transfer across sequences or levels of representation is a topic for future research.

7.3 Discussion

Although the results are largely qualitative, one can see that changing the representation quality has a dramatic effect on the information dynamics of the latent representation. On one extreme, when the rate is too low (like the case of Rate=10), after the initial surprise of introducing musical materials, most of the Prelude and Fugue "sound the same", as represented by an almost constant flat line at the top of Figure 9, Sub-figure E. The more subtle changes can be observed between sub-figures A and D, with Full Rate and Rate=50, respectively.

If we interpret information dynamics as measures of expectations formed by our imagination or mental representation that are inherently lossy, then the points at which the information rate changes at different bit-rate levels should be indicative of moments that carry creative or experiential significance at different levels of reduction. For instance one can see that the major events, such as transition to new materials in bars 25-27 cause a drop in information rate. Also,

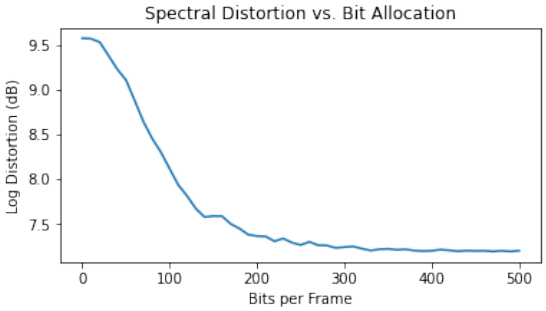


Fig. 11: Log Spectral Distortion of Bach Prelude at Varying Bit Allocation Rate

Table 3: The Reconstruction Distortion with respect to Bit-Allocation Rate

Rate	**10**	**100**	**200**	**500**	**Original**
Log Spectral Distortion (dB)	9.562	8.160	7.377	7.185	7.185

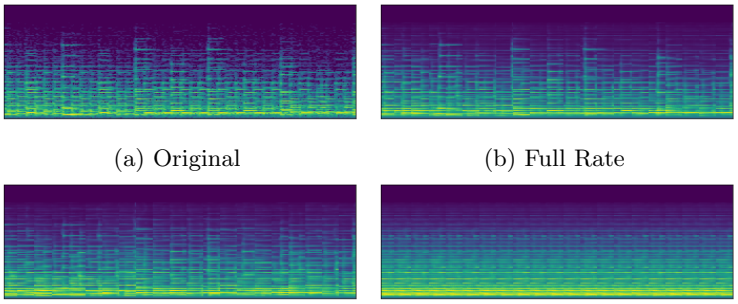
the development of thematic material in the Fugue starting around bars 44-45 causes an increase in information rate in both full rate and at a rate of 50 [13]. For high bit-rates, the graph of IR seems to be relatively robust, and for very low-rates the quality of musical reconstruction seems to be too low to capture significant musical information. For rate of 10 we see that after a short initial period where materials were not repetitive (low information rate), the rest of the music is perceived as one long repetition (high IR).

7.4 Audio Case

For our experiments with audio, we trained and evaluated VAE on audio files from the LABROSA data-set, as we did with the polyphonic MIDI experiment. The VAE used had a single fully connected hidden layer of size 100. More details on the architecture are available in Appendix B. Training was performed using an Adam Optimizer with a cyclical annealing schedule weight on the KL divergence loss (Fu et al., 2019) to promote stronger reconstruction, and early stopping to select the best model on the validation set. The experiments described below were performed on a WAV file of J.S. Bach's Prelude No.2 in C Minor, BWV 847.

Audio is reconstructed from the STFT spectrograms via the Griffin Lim algorithm (Griffin & Lim, 1984). Reconstruction of the audio is then evaluated using the log spectral distortion between the original and reconstructed audio tracks as defined in section 3.3. Full rate reconstruction of the the test audio file provides a log spectral distortion of 7.185 dB. We then measured the reconstruction quality of audio as we applied bit-allocation to control the amount

[13] The units of bit-rate reduction are total bits per measure



(a) Original (b) Full Rate

(c) Rate = 100 (d) Rate = 20

Fig. 12: Spectrograms of two measure excerpt at various bit-allocation rates

of bits of the latent variable that are transmitted per audio frame. Figure 11 depicts the relationship between the amount of latent bits per frame and the reconstruction quality measured in log spectral distance. As bit rate decreases, we notice a decrease in reconstruction quality as indicated by the increase in log spectral distortion. At around 200 bits, the perceived reconstruction error plateaus, making the bit-reduced reconstructions near indistinguishable from the full rate reconstruction. Figure 12 depicts short spectrogram clips for audio reconstruction at low, middling, and full bit-allocation rates. The spectrogram of the full bit-allocation rate closely matches the structures seen in the original structure. Around 100 bits per frame, noticeable artifacts appear throughout the spectrogram and create unwanted harmonics. At bit-allocation rates lower than 50 bits, much of the original structure is difficult to discern as fewer differences are allowed between audio frames, resulting in a sound that resembles a constant resonant hum that "averages" the tonal qualities of the piece into one complex timbre.

As the audio frames are of relatively short duration, the time structure of the signal requires predictions based on a long memory. Accordingly, VMO was used to measure the structure of the predictive information in a musical track and in its latent encoding. Cosine distance between latent encodings is used when computing thresholds in the quantization of states for VMO. Figure 13 shows the information dynamics of an audio track at varying bit-allocation rates. In the audio domain, we make similar observations regarding the relationship between bit-allocation rate and the patterns in information rate as those made with MIDI data in section 7.2. At low bit-rates, the quality of audio reconstruction was much too low for VMO to capture meaningful information dynamics. Higher bit-rates exhibit more meaningful variation in information rate. For instance, in the first 80 frames we see much more varying activity in the higher frequency bins, matched by the relatively low and rapidly fluctuating information rates. Although most of the analysis is qualitative at this point, we plan to explore the

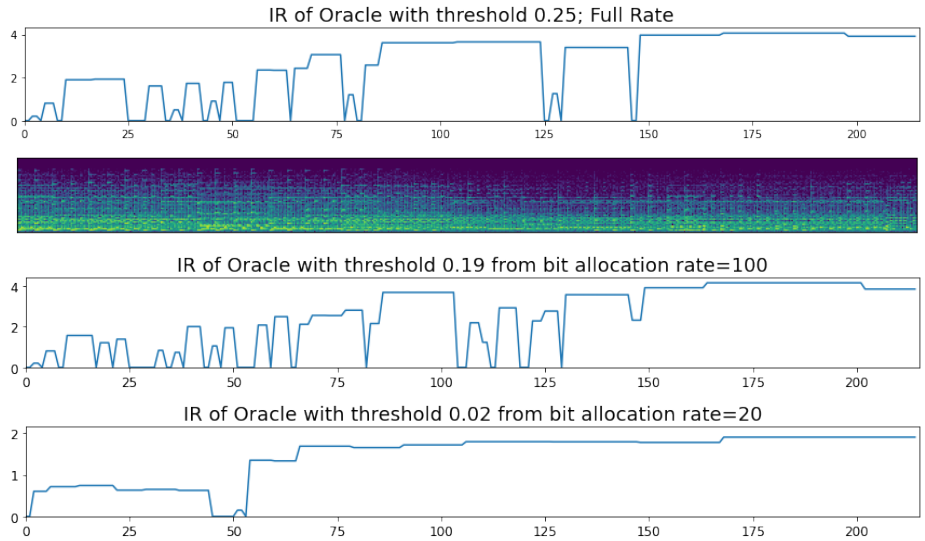


Fig. 13: Analysis of Information Rate using VMO of Audio of Bach Prelude

utility of bit-rate reduction for generative purposes, similar to the monophonic generation experiments reported in section 7.1.

8 Summary and Discussion

In this paper, we presented a theoretical framework of musical creativity and surprisal formulated in terms of information theoretical relations between full-rate (high-fidelity) encoding of the musical data, and a lower complexity latent encoding that models reduced informational aspects of musical structure. Some of the motivation for such reduction comes from limited capacity of mental or imaginary musical representations that might play role in listening or composition. The cognitive assumption is that musical creativity and musical perception take into account a trade-off between abstraction or simplified representation of music, which captures more salient or structural aspects of music, and the perceptual sensibility to the musical surface that is abundant in its sonic or musical notation detail.

8.1 Novel Learning Principle for Creative Systems

In this paper, we proposed a DMID principle that minimized the amount of information dynamics happening on the musical surface (mutual information between past X and future Y of the musical signal), given that we have the ability to extract a musical reduction or form an embedding (latent representation Z derived from past surface X) that captures both the temporal structural

relations between musical past and present, and is able to predict the future of the musical surface in detail. This principle was expressed mathematically as $\min\{I(X, Y|Z)\}$. This approach explicitly points to the importance of understanding the relation between deep encoding of musical surface, measured in terms of the ability of latent representation to capture the surface detail, and the structural aspects of the reduced representation, measured in terms of Information Dynamics of these latent states. The encoding of the musical surface was done using a Variational Auto-Encoder (VAE). An interesting insight into the DMID principle comes from the relation between the mutual information of the latent codes Z comprising the musical reduced representation and musical surface X. In (Alemi et al., 2017) it was shown that maximization of Evidence Lower Bound (ELBO) is equivalent to minimization of $I(X, Z)$ for the VAE encoder combined with minimization of the decoder reconstruction error $D(X, Z)$. This provides yet another insight into combined objective of representation learning and DMID, expressed as a loss $\mathcal{L}$

$$\min \mathcal{L} = \min(-ELBO(\beta) + \gamma D_{KL}(p(Y|X)||p(Y|Z))) \tag{12}$$

$$= \min(I(X, Z) + \beta\langle D(X, Z)\rangle - \gamma(I(Z, T) - I(Z, T|Y))) \ . \tag{13}$$

where we omitted the $I(X, Y)$ term that does not depend on the model parameters. This gives us an insight into a combined principle for a learning creative musical system that is formulated in terms of multiple simultaneous goals:

- finding the most efficient embedding or smallest latent encoding (minimal $I(X, Z)$)
- best reconstruction of the surface data from this embedding (minimal $D(X, Z)$)
- best anticipation of the embedding expressed in terms of highest information dynamics in the latent space (maximizing $I(Z, T)$)
- and finally, least discrepancy or least surprisal in the latent embeddings when the next musical surface is revealed (minimal $I(Z, T|Y)$)

To the best of our knowledge, no such conceptual or computational framework had been previously offered, which we believe provides important insights for design and understanding of creative musical systems. It also should be noted that DMID framework explicitly takes into account the structural importance of temporal relations for learning musical representations in a generative setting. Generative models are the necessary first step in planning or active learning approaches where an agent learn to operate in a world that has latent causes that it has to infer, and actions it can do on the world. These active inference approaches have recently been suggested as models for creativity and fun (Schmidhuber, 2010),(Salge, Glackin, & Polani, 2013). There are many common conceptual parallels between DMID and agent model, where it was demonstrated that a reduced representation of the data extracted from the world where an agent operates is important for successful agent operations. The learning framework and formalism in these models is one of Reinforcement Learning, where compact and simple policy is trained not only to solve a required task, but also the agent can be trained inside of its own dream environment generated by its world model.

DMID is much more limited in this respect, as it focuses on modeling the musical "world", rather then learning a behavior to execute a specific task, or operate through intrinsic motivation or sense of agent empowerment when the task is not defined. It should be noted that only after a correct representation is learned by the agent and a generative model is created, reinforcement learning may allow making inferences about sequences of latent states of the world that form optimal policies for agent's behavior in that world. In active learning literature the temporal aspects of data are modeled through agent behavior. Accordingly, time structures are captured by learning policies or by choosing the sequence of states, usually motivated by minimizing surprisal for the agent operating in that world. In the case where there is a lack of sensory input, maximising the options for the agent by putting it into a position of having most number of options is know as "empowerment". Exploring the policies or intrinsic motivations of such agents is an important next step in music research, and we see great potential of combining representations that are learned using DMID with active inference models for musical agents. Accordingly, we envision more interesting insights and practical results from using DMID in the active learning setting.

The results presented in the paper are meant to support this framework by demonstrating the analytical potential of our approach, also outlining the algorithmic tools used at each step. The code and examples related to the experiment reported in this paper can be found at https://github.com/RetroCirce/DMID-Analysis-Music-Dynamics

Acknowledgment

I would like to thank the reviewers for the detailed and insightful comments. This work was partially supported by Cygames, Inc.

Appendix

A Proof of Equation 1

In our notation, $D_{KL}(\cdot,\cdot)$ is the Kullback-Liebler distance between different distributions, and $I(\cdot,\cdot)$ is the mutual information between their random variables.

$$D_{KL}(P,Q) = \int p(x) \log \frac{p(X)}{q(X)} dX \tag{14}$$

with Mutual information defined as

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \tag{15}$$

and signal entropy given by

$$H(X) = -\int p(X) \log p(X) dX \tag{16}$$

Another useful relation is between KL distance and Mutual Information, which can be derived from the above definitions,

$$I(X,Y) = D_{KL}(p(X,Y), p(X)p(Y)) = D_{KL}(p(X|Y)p(Y), p(X)p(Y)) \tag{17}$$

In other words, Mutual Information $I(X,Y)$ measures the KL distance between a joint distribution $P(X,Y) = p(X,Y)$ and marginal distribution $Q(X,Y) = p(X)p(Y)$. Taking into account the Markov relations $Z - X - Y$, we have the following conditional independence relations between our variables $p(Y,X,Z) = p(Y|X,Z)p(X,Z) = p(Y|X)p(X,Z)$. Using the definition of KL divergence and the averaging over $P(X,Z)$ we have

$$\begin{aligned}
&\langle D_{KL}(p(Y|X)||p(Y|Z))\rangle_{p(X,Z)} \\
&= \int p(X,Z)(\int p(Y|X) \log \frac{p(Y|X)}{p(Y|Z)} dY) dX dZ \\
&= \int p(X,Z,Y) \log(\frac{p(Y|X)p(X)}{p(X)p(Y)} \frac{p(Y)p(Z)}{p(Y|Z)p(Z)}) dY dX dZ \\
&= \int p(X,Z,Y) \log(\frac{p(Y,X)}{p(X)p(Y)} - \int p(X,Z,Y) \log(\frac{p(Y)p(Z)}{p(Y,Z)} \\
&= I(X,Y) - I(Z,Y)
\end{aligned}$$

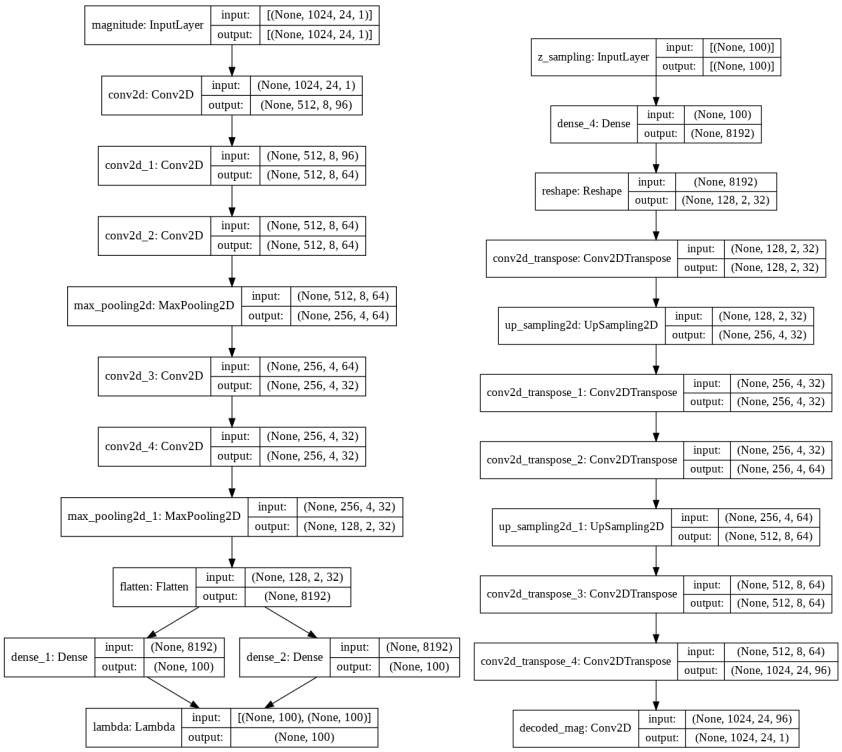
where we used the KL and Mutual information relations shown above.

The last remaining relation we need to show is that $I(X,Y|Z) = I(X,Y) - I(Z,Y)$. This can be proved by considering the definition of mutual informaiton as

$$\begin{aligned}
I(X,Y) &= H(Y) - H(Y|X) \\
I(Z,Y) &= H(Y) - H(Y|Z)
\end{aligned}$$

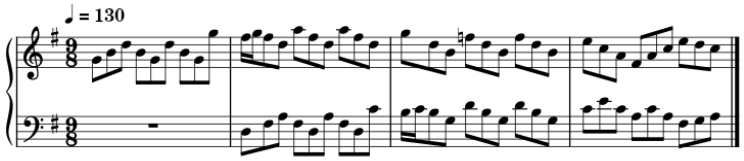
and using the Markov relation $H(Y|X,Z) = H(Y|X)$ to see that $I(X,Y|Z) = H(Y|Z) - H(Y|X,Z) = H(Y|Z) - H(Y|X) = H(Y) - H(Y|X) - H(Y) + H(Y|Z) = I(X,Y) - I(Z,Y)$

B Audio VAE Architecture

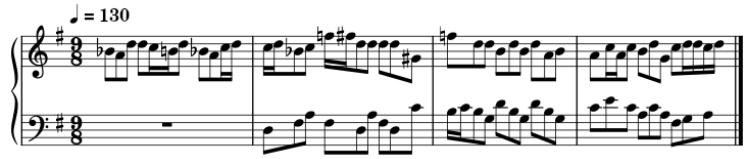


Audio VAE Encoder (left) and Decoder (right). The encoder consists of five 2D convolutional layers and two max pooling layers. Two linear layers generate the mean and logarithmic variance used to sample the output of the encoder. The decoder also consists of five transpose convolution layers and two upsampling layers to restore the output to the original shape of the input.

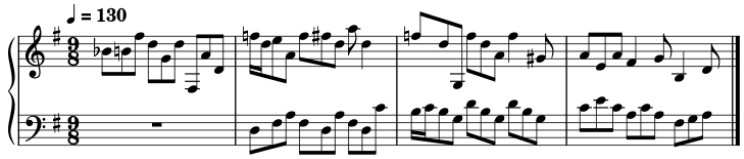
C Manipulating Latent Variables by Bit-allocation and Demos



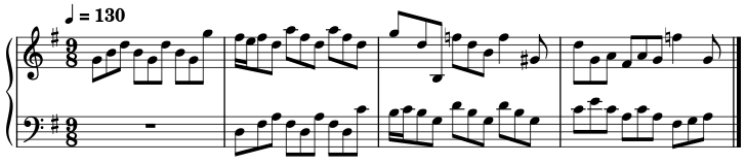
(a) The original four bars of Bach Invention No.10



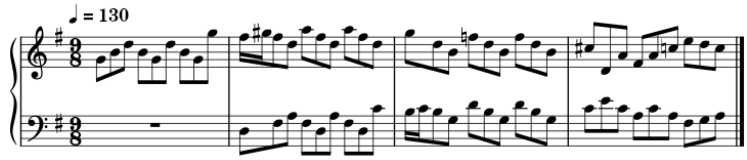
(b) The reconstructed four bars of Bach Invention No.10 under rate = 10.



(c) The reconstructed four bars of Bach Invention No.10 under rate = 100.



(d) The reconstructed four bars of Bach Invention No.10 under rate = 1000.



(e) The reconstructed four bars of Bach Invention No.10 under rate = 10000.

The reconstruction musical bars from the latent variables with different bit-allocation rate. The music in the first voice is generated from different bit-allocation rate, while the second voice remains the same (the original one).

References

Abdallah, S., & Plumbley, M. (2009, June). Information dynamics: Patterns of expectation and surprise in the perception of music. *Connect. Sci*, *21*(2-3), 89–117.

Alemi, A. A. (2019). Variational predictive information bottleneck. In *Aabi.*

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2017). An information-theoretic analysis of deep latent-variable models. *CoRR*, *abs/1711.00464*.

Allauzen, C., Crochemore, M., & Raffinot, M. (1999). Factor oracle: A new structure for pattern matching. In *SofsemÕ99: Theory and practice of informatics* (pp. 295–310).

Assayag, S. D. O., Gérard; Dubnov. (1999). Guessing the composer's mind: Applying universal prediction to musical style. In *Proceedings of the international computer music conference, (icmc).*

Barrington, L., Chan, A. B., & Lanckriet, G. (2009). Dynamic texture models of music. In *in ieee icassp* (pp. 1589–1592).

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Hjelm, R. D., & Courville, A. C. (2018). Mutual information neural estimation. In *Proceedings of the 35th international conference on machine learning, (icml)* (Vol. 80, pp. 530–539). PMLR.

Berger, T. (1971). *Rate distortion theory; a mathematical basis for data compression.* Prentice-Hall Englewood Cliffs, N.J.

Chen, K. (2021). Controllable monophonic music generation via latent variable disentanglement. *Master Thesis Archive, University of California San Diego*.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-kirkpatrick, T., & Dubnov, S. (2021). Zero-shot audio source separation through query-based learning from weakly-labeled data. In *Proceedings of the thirty-sixth AAAI conference on artificial intelligence, AAAI 2022.* AAAI Press.

Chen, K., Liang, B., Ma, X., & Gu, M. (2021). Learning audio embeddings with user listening data for content-based music recommendation. In *International conference on acoustics, speech and signal processing, ICASSP* (pp. 3015–3019). IEEE.

Chen, K., Wang, C., Berg-Kirkpatrick, T., & Dubnov, S. (2020). Music sketchnet: Controllable music generation via factorized representations of pitch and rhythm. In *Proceedings of the 21st international society for music information retrieval conference, ISMIR.*

Chen, K., Xia, G., & Dubnov, S. (2020). Continuous melody generation via disentangled short-term representations and structural conditions. In *IEEE 14th international conference on semantic computing, ICSC* (pp. 128–135). IEEE. doi: 10.1109/ICSC.2020.00025

Chen, K., Zhang, W., Dubnov, S., Xia, G., & Li, W. (2019). The effect of explicit structure encoding of deep neural networks for symbolic music generation. In *International workshop on multilayer music representation and processing, MMRP 2019.* IEEE.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory.* N. Y.: John Wiley & Sons, Inc.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory 2nd edition (wiley series in telecommunications and signal processing).* Wiley-Interscience. Hardcover.

Dong, H., Chen, K., McAuley, J. J., & Berg-Kirkpatrick, T. (2020). Muspy: A toolkit for symbolic music generation. In *Proceedings of the 21th international society for music information retrieval conference, ISMIR 2020* (pp. 101–108).

Dong, Z., Oktay, D., Poole, B., & Alemi, A. A. (2020). *On predictive information in rnns.*

Donsker, M., & Varadhan, S. (1983). Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, *36*(2), 183–212.

Dubnov, S. (2006). Spectral anticipations. *Computer Music Journal*, *30*(2), 63-83.

Dubnov, S. (2020). Deep music information dynamics. In *The 2020 joint conference on ai music creativity.*

Dubnov, S., Assayag, G., & Cont, A. (2011). Audio oracle analysis of musical information rate. In *Ieee international conference on semantic computing (icsc)* (pp. 567–571).

Elliott, D. J. (1993). Musicing, listening, and musical understanding. *Contributions to Music Education*(20), 64-83.

Fraccaro, M., Sønderby, S. K., Paquet, U., & Winther, O. (2016). *Sequential neural models with stochastic layers.*

Friston, K. (2012). The history of the future of the bayesian brain. *Neuroimage*, *62*(2), 1230-3.

Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., & Carin, L. (2019). *Cyclical annealing schedule: A simple approach to mitigating kl vanishing.*

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*, 424-438.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236-243. doi: 10.1109/TASSP.1984.1164317

Hsu, W.-N., Zhang, Y., & Glass, J. (2017). *Learning latent representations for speech generation and transformation.*

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation.* MIT Press.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd international conference on learning representations, ICLR.*

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. In *Proceedings of the 2nd international conference on learning representations, ICLR.*

Lefebvre, A., Lecroq, T., & Alexandre, J. (2003). An improved algorithm for finding longest repeats with a modified factor oracle. *Journal of Automata,*

Languages and Combinatorics, *8*(4), 647–657.

Lerdahl, F., & Jackendoff, R. S. (1996). *A generative theory oftonal music*. MIT press.

Luo, Y., Agres, K., & Herremans, D. (2019). Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders. In *Proceedings of the 20th international society for music information retrieval conference, ISMIR* (pp. 746–753).

Marcus T. Pearce, G. A. W. (2012). Auditory expectation: The information dynamics of music perception and cognition. *Topics in Cognitive Science*, *4*(4), 625–652.

Marsden, A. (2010). Schenkerian analysis by computer: A proof of concept. *Journal of New Music Research*, *39*(3), 269-289.

Meyer, L. B. (1956). *Emotion and meaning in music*. Univ. of Chicago Press.

Pasquier, P., Eigenfeldt, A., Bown, O., & Dubnov, S. (2017). An introduction to musical metacreation. *Computers in Entertainment*, *14*(2).

Pati, A., Lerch, A., & Hadjeres, G. (2019). Learning to traverse latent spaces for musical score inpainting. In *Proceedings of the 20th international society for music information retrieval conference, ISMIR* (pp. 343–351).

Roberts, A., Engel, J. H., Raffel, C., Hawthorne, C., & Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. In *Proceedings of the 35th international conference on machine learning, ICML* (Vol. 80, pp. 4361–4370). PMLR.

R.Sims, C. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181-198.

Salge, C., Glackin, C., & Polani, D. (2013). *Empowerment – an introduction.*

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Transactions on Autonomous Mental Development*.

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461–464.

Tishby, N., & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. Retrieved from `http://arxiv.org/abs/1503.02406`

Wang, C., & Dubnov, S. (2014). Guided music synthesis with variable markov oracle. In *The 3rd international workshop on musical metacreation, 10th artificial intelligence and interactive digital entertainment conference.*

Wang, C., & Dubnov, S. (2015a). Pattern discovery from audio recordings by variable markov oracle: A music information dynamics approach. In *Ieee international conference on acoustics, speech, and signal processing, ICASSP.*

Wang, C., & Dubnov, S. (2015b). The variable markov oracle: Algorithms for human gesture applications. *IEEE MultiMedia*, *22*(04), 52-67.

Wang, Z., Chen, K., Jiang, J., Zhang, Y., Xu, M., Dai, S., ... Xia, G. (2020). POP909: A pop-song dataset for music arrangement generation. In *Proceedings of the 21st international society for music information retrieval conference, ISMIR.*

Wang, Z., Zhang, Y., Zhang, Y., Jiang, J., Yang, R., Zhao, J., & Xia, G. (2020).

PIANOTREE VAE: structured representation learning for polyphonic music. In *Proceedings of the 21st international society for music information retrieval conference, ISMIR.*