Stylistic Melody Generation with Conditional Variational Auto-Encoder

Junyan Jiang (junyanj) [1] **Zhiqi Wang (zhiqiw)** [1]

Abstract

Stylistic music generation requires human controls over high-level music features, e.g., rhythm, pitch, structure, and genre. While *Structure* is a key notion for both music generation and representation learning, most existing generative models for music either adopt hard-coding structures or merely implicitly inherit the structure from the training pieces. Consequently, it is still difficult to control music structure in a flexible way in order to create new pieces. In this paper, we propose several novel ways trying to represent music structure using the framework of Conditional Variational Auto-Encoder (CVAE). Furthermore, we explore the problem of *disentanglement* of the representations of music *structure* and *content* for long music pieces. We show in the results that we can extract the bar-level structure of a long music piece while creating novel pieces via swapping the representations of different pieces.

1. Introduction

Stylistic music generation is an interesting and important topic in computer music, music therapy and computational creativity. With the development of deep generative models, researchers have developed several new models for music generations, like MusicVAE (Roberts et al., 2018) and MidiNet (Yang et al., 2017), the aim of which is to train a model that create new pieces without human guidance. However, music generation in practice does not always expect computers to freestyle; we want to keep in charge of some stylistic property of the generative piece. For example, we may ask the questions like how to generate a melody given a specific chord progression, how to generate pitch contour given rhythm, or how to generate a similar piece given a sample. This requires us to develop generative models with stylistic controls.

In this project, we investigate the problem of generating *lead melody*, an important element for modern popular music. Under this setting, we can classify stylistic controls into two categories:

(1) Internal properties of a melody, e.g., rhythm, pitch contour, structure;

(2) External context, e.g., chord sequence, tonality.

A successful generative model with a specific stylistic control should be able to analyze and disentangle such feature automatically if it is an internal property, as well as to generate a new one consistent with human's controls.

Notice that these two processes are highly related to VAE, as the analysis part can be performed by the encoder and the generation part can be performed by the decoder. Therefore, we believe that VAE is a good starting point for further implementation.

Among these stylistic features, *structure* is the most difficult one to handle both in analysis tasks and generation tasks (Paulus et al., 2010; Dai et al., 2018). Music structure itself is a complex phenomenon. First, music structure is a multi-level concept (Hamanaka et al., 2018). On bar level, the structure may refer to how nearby music notes are grouped; on phrase-level, the structure may refer to local repetition and similarity; on song level, the structure may refer to the organization of different segments (e.g., verses and chorus). Secondly, structures may refer to a variety of composition technologies including repetition, duality, melodic sequence and motivic development. The complexity of music structure makes it difficult to be modeled. Currently, mainstream models including MusicVAE and MidiNet still struggle with generating melodies with nice structural properties.

The aim of the project is to design a new deep generative model based on MusicVAE (Roberts et al., 2018) for stylistic melody generation. For internal properties controls, we will focus on a new generative model with the ability to automatically analyze and separate the concept of music structure given a music piece, and the ability to generate new songs consistent with the given structure representation. The model combined VAE and self-attention, a popular architecture for modeling sequence structure. For external context, we will propose a conditional version of the system that utilizes the chord sequence and the tonality as an external control. We will provide some experimental

[1] Carnegie Mellon University, Pittsburgh, PA 15213, USA.

10708 Class Project, Spring, 2019.

results on the model, including the effectiveness of such controls and musicality of generated pieces.

2. Related Work

2.1. Stylistic Music Generation

Traditionally, rule-based expert systems, automata and Hidden Markov Models (HMM) are widely used in automatic music generation tasks (Conklin & Witten, 1995; Lo & Lucas, 2006; Thornton, 2009). Commonly, they all require human-defined style parameters for music generation. It is therefore hard for the systems to adapt to new styles in general.

Recently, the development of deep generative models provides researchers with new insight for sequence generation. In the recent two years, deep generative models became widely adopted in music generation. For example, DeepBach (Hadjeres et al., 2017) proposed a method to generate Bach-style music. DeepBach adopts a Long Short-Term Memory (LSTM) network for melody modeling and uses a pseudo-Gibbs Sampling technique for music generation. While the results are exciting, it is worth noticing that the methods in DeepBach are hard to be reproduced on other music styles as the model incorporates strong assumptions on Bach's style into the model itself. Another model, DeepJ (Mao et al., 2018) combined a Biaxial LSTM architecture with additional genre conditioning at every layer. To get a parametric representation of the music genre, the model uses one embedding layer to learn the distributed representation and then another fully-connected hidden layer is applied. The generated samples by DeepJ did demonstrate the stylistic generation. However, the generated piece still lacks long term structures, and a large amount of data is required for training a new genre.

Other attempts include the usage of popular deep generative models. The MidiNet (Yang et al., 2017) adopts Generative Adversarial Networks (GAN) for fixed-length music generation and the MusicVAE (Roberts et al., 2018) uses the Variational AutoEncoder (VAE) model for melody generation. We will discuss about MusicVAE in detail in 2.3.

2.2. Variational Autoencoder

The variational autoencoder (Kingma & Welling, 2013) are powerful latent variable models which consist of an encoder and decoder. The encoder compresses training data into a latent space, to learn a Gaussian probability density $q_\theta(z|x)$, from which we can sample to get noisy values of representations z. The decoder uses the representation z to reconstruct the original data. Ideally, the latent variable z captures the probability characteristics of giving data points in the datasets. Recently, VAE and its variance have been applied to many fields of applications, and perform efficient inference by deep neural networks. Conditional-VAE is proposed by giving labels into input, and gained impressive performance in Attribute2Image (Yan et al., 2016). Since we noticed that harmonic context is an important factor in both perception and composition process, it can be regarded as a condition to the model and should be provided to both the encoder and decoder as a condition. Conditional VAE will be applied in our further implement.

However, one drawback of using a single latent unit is that it is sensitive to changes in one factor, while not for other factors. Therefore, disentangled representation are usually factorized so that different independent latent units encode different independent factors in data. In our experiment, the melody contains two important musical factors: the pitch contour and the rhythm pattern. They can be separated into two latent vectors and have their specific defined loss function. Traditional attempts to learn disentangled representations under supervised learning, which is unrealistic in our domain. One famous unsupervised approaches $\beta-$VAE (Higgins et al., 2017) proposed to add an extra hyper-parameter to the original VAE objective, which allows constraints on the encoding capacity of latent vectors and encourage factorization.

2.3. MusicVAE

MusicVAE (Roberts et al., 2018) adopts the idea of VAE for melody generation. In this model, the piano roll of a monophonic melody line m is fed into the encoder to get a latent variable z_m, and the decoder tries to reconstruct the melody from z_m. The encoder and the decoder use LSTM cells to process variant-length data.

The major contribution of the work is to show that VAE is a viable generation model for music melody, as the model successfully generated musically meaningful 2-bar samples by interpolating between the latent code of given melodies. However, the model fails to generate, or even reconstruct a long piece (e.g., 16-bar samples) well even if it explicitly adopts a hierarchical architecture for the decoder. Also, it is not a stylistic melody generator as no disentanglement is performed on the latent space yet. Thus, we cannot control the stylistic parameters of the generative melody in an initiative way. These two issues of MusicVAE inspired us to work on the project.

2.4. Attention-Based Music Generation

The attention mechanism is adopted in sequence to sequence models and achieved huge success in many generative tasks like machine translation (Vaswani et al., 2017). Such a mechanism can help the model focus on certain inputs to the fore while diminishing the importance of the others. (Huang et al., 2019) combined self-attention with seq2seq model, trying to generate music while keeping long-term structures and achieved better results compared to only using LSTM without self-attention. Their experiment demonstrated that training with relative attention has benefits for learning long-term structural information. However, this can not be directly used in VAE, since compared with LSTM encoder, lots of information is lost when VAE encodes inputs into latent variables.

There are also some working attempts to combine variational attention for sequence-to-sequence models (Bahuleyan et al., 2017a). Such mechanism works on the Variational Encoder-Decoder models (e.g., for machine translation), but due to different encoding targets, these models cannot be directly applied to VAEs.

3. Proposed Methods

We here propose three methods that we experiment with during the entire semester. We will discuss the advantage and disadvantage of each method in the discussion section.

3.1. Bar-Level Similarity Representation

Symbolic music can be regarded as a highly structured sequence, with repetition and similarity on different scales. It is generally hard to model music structures both in transcription problems and generation problems, and a unified representation of music structure is not yet concluded (Paulus et al., 2010).

Assume that a music melody is denoted as a sequence $\mathbf{m}_{1..n}$. Typically, a token $\mathbf{m}_i$ refers to the note sequence in one bar or one beat. In our project, we use one bar for convenience. A traditional metric for describing the music structure is the similarity matrix $\mathbf{T}$, in which each cell T_{ij} denotes the correlation between certain music elements (e.g., melody or rhythm) between $\mathbf{m}_i$ and $\mathbf{m}_j$.

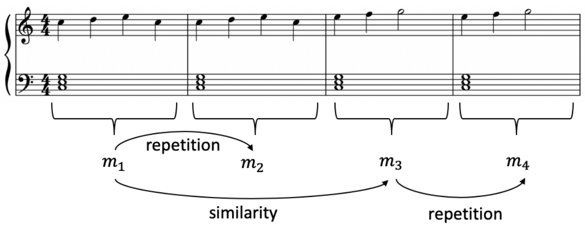


Figure 1. An example of music structure analysis of the first four bars of the song *Two Tigers*. The song contains exact repetitions as well as inexact similarity in pitch contour.

Notice that although the similarity matrix $\mathbf{T}$ is very convenient for music analysis, it is still not obvious how to generate a new piece according to $\mathbf{T}$ itself. Therefore, we here design a new form of similarity representation with the capability of both analyzing and generation. We call it Masked Self-Attention Similarity Matrix, or MSASM in short.

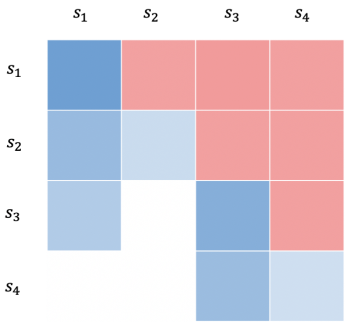


Figure 2. An example masked self-attention similarity matrix of the song *Two Tigers* (Best viewed in color). The red cells are masked out to zero. Darker blue denotes a larger attention value.

Given a melody piece $\mathbf{m}_{1..n}$, the h-th MSASM $\mathbf{A}^h$ is an $n \times n$ matrix given by

$$A^h_{i,j} = \begin{cases} \text{Softmax}\left(\frac{\mathbf{Q}^h_i \mathbf{K}^h_j + S^h_{i-j}}{T}\right) & i \geq j \\ 0 & i < j \end{cases} \quad (1)$$

Where T is the temperature hyper-parameter,$\mathbf{S}^h$ is the relative position encoder proposed by (Huang et al., 2019), and $\mathbf{Q}^h$, $\mathbf{K}^h$ are the attended variable given by

$$\mathbf{Q}^h_i = \mathbf{W}^h_Q \text{Encoder}(\mathbf{m}_i) \quad (2)$$

$$\mathbf{K}^h_i = \mathbf{W}^h_K \text{Encoder}(\mathbf{m}_i) \quad (3)$$

And $h = 1..H$ is the head number. Notice that each head may attend on a different property between two bars $\mathbf{m}_i$ and $\mathbf{m}_j$. For example, some heads focus more on the rhythmic similarity while others focus more on the pitch contour similarity.

In the decoding phase, we utilize the similarity matrix to generate a new sample that shares a similar music structure to the provided one. Given a hidden representation $\mathbf{z}_{1..n}$ of each bar, we reweight them according to the attention values in $\mathbf{A}$ to get a structure-aware representation:

$$\mathbf{z}^h_i = \mathbf{W}^h_V \mathbf{z}_i \quad (4)$$

$$\mathbf{z}'^h_i = \begin{cases} \mathbf{z}^h_i & i = 1 \\ \mathbf{z}^h_i A^h_{i,i} + \sum_{j=1}^{i-1} \mathbf{z}'^h_j A^h_{i,j} & i > 1 \end{cases} \quad (5)$$

$$\mathbf{z}'_i = \text{Concat}(\mathbf{z}'^1_i, ..., \mathbf{z}'^H_i) \quad (6)$$

It is worth noticing that the formula is not exactly the same as in the original transformer model (Vaswani et al., 2017), where ${\mathbf{z}'}_j^h A_{i,j}^h$ are replaced by $\mathbf{z}_j^h A_{i,j}^h$. This is because we are allowing chained similarity, for example, if bar 3 is similar to bar 2 and bar 2 is similar to bar 1, it does not necessarily imply bar 3 is similar enough to bar 1, but bar 3 still requires information in bar 1 for reconstruction as they have some correlation. We will provide a probabilistic explanation of this approach in 3.1.1.

3.1.1. A Probabilistic View

A trivial way to extend the MusicVAE model to arbitrary music length is to apply a 1-bar MusicVAE on each bar of the music:

$$\begin{aligned} \mathbf{s}_i &= \text{Encoder}(\mathbf{m}_i) && (7) \\ \mathbf{z}_i &\sim P_{\mathbf{s}_i} && (8) \\ \mathbf{m}'_i &= \text{Decoder}(\mathbf{z}_i) && (9) \end{aligned}$$

Where $\mathbf{z}_i$ is the variational variable and $\mathbf{s}_i$ is the encoded parameter that controls the variational distribution $\mathbf{z}_i$.

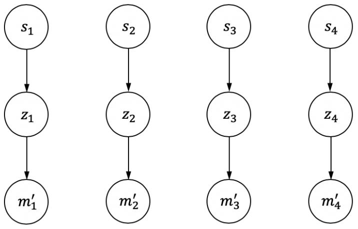


Figure 3. The graphical model of the trivial extension of Music VAE.

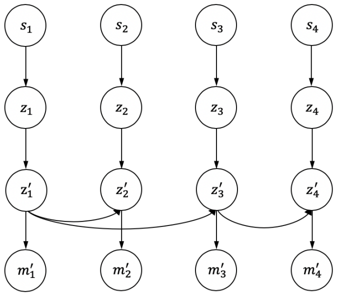


Figure 4. An example graphical model of the proposed model. This graph corresponds to the song *Two Tigers* as shown in Figure 1

An obvious problem for the model is that the model is not aware of any structural information of the music. It assumes independence between the representations of each two bars, as shown in figure 3. Thus, if we resample the variables $\mathbf{z}_{1..n}$, any inter-bar structural property will be lost. However, if we instead sample $\mathbf{z}'$ given by equation (4) for decoding instead of $\mathbf{z}$, we can retain such dependency.

One implementation problem of the model is the fact that the MSASM can easily degenerate to a diagonal matrix. In this case, all the dependencies between bars disappear, and the model becomes a trivial case. To prevent this from happening, we pose additional penalties L_d to the MSASM on the diagonal entries to encourage the model to explore the correlation between one bar and its previous ones:

$$L_d(\mathbf{A}) = \frac{1}{nH} \sum_{h=1}^{H} \sum_{i=1}^{n} A_{i,i}^h \qquad (10)$$

The total loss function is given by:

$$L = L_{recons} + \lambda_1 D_{KL}(p(\mathbf{z}' \mid \mathbf{m}) || p(\mathbf{z}')) + \lambda_2 L_d(\mathbf{A}) \qquad (11)$$

Where L_{recons} is the reconstruction loss and λ_1, λ_2 are the hyper-parameters.

3.2. Structured Condition

Human composers always refer to some contextual information when producing a music melody. The contextual information may vary depending on the situation, but the most common ones are the harmonic context, i.e., the chord sequence and the tonality, while the chord sequence controls the local, short-term context of a melody and the tonality controls the long-term context. As the structure of harmonic context may reflect the structure of melody to some extent, such condition might be useful for an explicit representation of music structure.

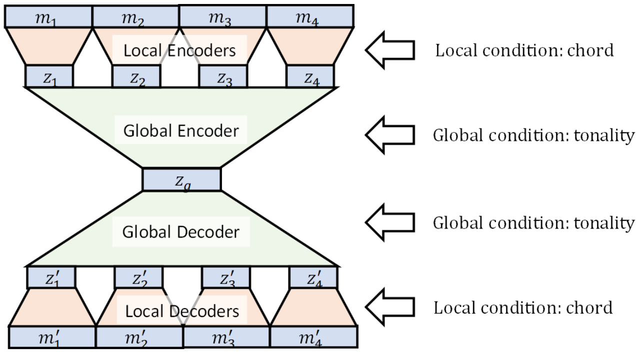
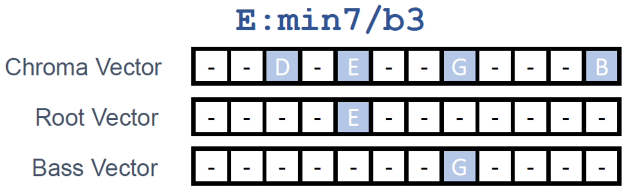


Figure 5. The model structure of the hierarchical CVAE.

We here adopt the Conditional-VAE (CVAE) architecture for our model. The modification from the VAE to its conditional version is simple. For both the encoder and the decoder, we add a new input providing the whole context information. However, when the VAE is hierarchical, we can feed the long-term and the short-term context into different parts of the encoder and decoder. In our implementation, we feed the short-term context into local encoders (decoders) and feed the long-term context into global encoder (decoders), as shown in Figure 5.

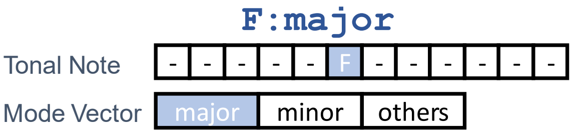
The major issue here is how to represent the chord sequence and the tonality into a numerical vector. For chord symbols, we here adopt a representation method provided by (Raffel et al., 2014), by denoting a chord into a multi-hot 36-D vector representing its root, bass and pitch class (i.e., the chroma vector).



E:min7/b3

Chroma Vector	-	-	D	-	E	-	-	G	-	-	-	B
Root Vector	-	-	-	-	E	-	-	-	-	-	-	-
Bass Vector	-	-	-	-	-	-	-	G	-	-	-	-

Figure 6. An example representation of the chord `E:maj7/b3`.



F:major

Tonal Note	-	-	-	-	-	F	-	-	-	-	-	-
Mode Vector	major	minor	others									

Figure 7. An example representation of the tonality `F:major`.

To represent the tonality, we are performing some simplification to the problem. We only retain 24 tonality classes (12 major classes and 12 minor classes) that are most common for modern popular songs and discard other rare modes like Dorian and Mixolydian. Also, we do not distinguish between natural minor, melodic minor and the harmonic minor. Under this setting, we use a one-hot representation for the 24 tonality classes.

3.3. Content-Structure Disentanglement

One critical problem of the bar-level similarity representation is the fact that the content representation is still not compact as different latent codes for different bars may share similar information, which is redundant. In other words, the model does not perform *disentanglement* between structure and content. To solve this issue, we propose a new model that try to learn a more compact representation for both structure and content.

We introduce a hierarchical β-VAE (Higgins et al., 2017) whose architecture is shown in Figure 9. The main idea of the model is to introduce a variational attention matrix $\mathbf{A}$ as a representation for the structure of a melody. In variational attention (Bahuleyan et al., 2017b), the attention matrix is converted into a variational variable, which is sampled from a continuous representation space.

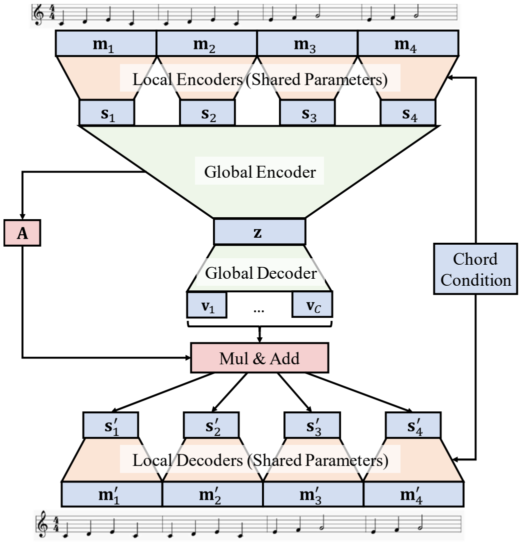


Figure 8. An overview of the hierarchical VAE.

The workflow of the model is as follows. To begin with, we have an input melody $\mathbf{m} = [\mathbf{m}_1, ..., \mathbf{m}_T]$ with T bars. Each $\mathbf{m}_t$ is first fed into a local encoder to get its bar-level representation $\mathbf{s}_t$. The global encoder takes in all $\mathbf{s}_t$ and outputs the distribution of attention matrix $\mathbf{A}$ (the structure representation) and the global latent variable $\mathbf{z}$ (the content representation of music theme). For the decoding process, a global decoder is first applied to the sampled global latent variable $\mathbf{z}$ and outputs C fixed-length latent components $[\mathbf{v}_1, ..., \mathbf{v}_C]$ where C is a hyperparameter. We want each component $\mathbf{v}_c$ to represent a distinct semantic factor of the original melody (e.g., the rhythm or pitch of the theme). Then, the sampled attention matrix $\mathbf{A}$ is used to calculate the weighted sum of the different latent components $\mathbf{v}_c$ to recover the bar-level representation $\mathbf{s}'_t$:

$$\mathbf{s}_t'^{(h)} = \sum_{c=1}^{C} A_{t,c}^{(h)} \mathbf{v}_c \qquad (12)$$

$$\mathbf{s}'_t = \mathbf{W}_d^\top \mathrm{Concat}(\mathbf{s}_t'^{(1)}, ..., \mathbf{s}_t'^{(H)}) \qquad (13)$$

Here, H denotes the number of heads for the attention and $\mathbf{W}_d$ is a linear transformation matrix. This process can been understood as *rearranging different semantic factors* $\mathbf{v}_c$ *of music content using the structure representation* $\mathbf{A}$. Finally, the local decoders are applied to recover the melody $\mathbf{m}'_t$ from $\mathbf{s}'_t$ for each bar. The loss function $\mathcal{F}(\phi, \theta; \mathbf{m}, \mathbf{c})$ of the model is given by

$$\begin{aligned}\mathcal{F} = \mathbb{E}_{\mathbf{A}\sim q_\phi^{(\mathbf{A})}(\cdot|\mathbf{m},\mathbf{c}),\mathbf{z}\sim q_\phi^{(\mathbf{z})}(\cdot|\mathbf{m},\mathbf{c})}&\left[\log p_\theta(\mathbf{m}|\mathbf{z},\mathbf{A},\mathbf{c})\right]\\ -\beta_z D_{KL}&\left[q_\phi^{(\mathbf{z})}(\mathbf{z}|\mathbf{m},\mathbf{c})\|p_{\mathbf{z}}(\mathbf{z})\right]\\ -\beta_a D_{KL}&\left[q_\phi^{(\mathbf{A})}(\mathbf{A}|\mathbf{m},\mathbf{c})\|p_{\mathbf{A}}(\mathbf{A})\right]\end{aligned} \tag{14}$$

where $\mathbf{c}$ is the condition for the model, ϕ and θ are the parameters of the encoder and the decoder, respectively.

One thing worth noting is why we introduce the variational attention loss term to the model. We find that in practice, if $\mathbf{A}$ is not variational, then the model will become problematic in a way that $\mathbf{A}$ will form a *shortcut* for information to transfer from encoder to decoder, and $\mathbf{A}$ itself will be enough to recover the whole melody line without any information stored in $\mathbf{z}$. The variational attention (Bahuleyan et al., 2017b) is introduced to solve this issue by adding some noise to the attention matrix before decoding. Notice that although the distribution of $\mathbf{A}$ can be modeled using a Dirichlet distribution as each attention vector sums up to 1, it is not feasible because the reparameterization trick is hard to perform on Dirichlet distributions. Instead, we still use a Normal distribution for $\mathbf{A}$ prior and then perform $\exp(\cdot)$ and normalization to recover the real attention matrix from the sampled latent variable.

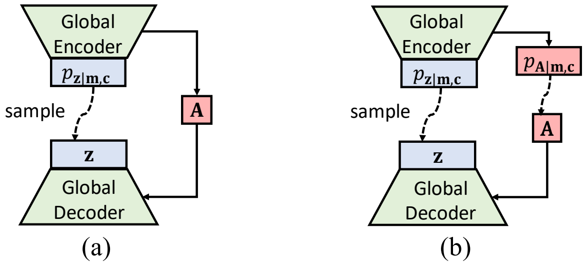

Global Encoder
$p_{\mathbf{z}|\mathbf{m},\mathbf{c}}$
sample
A
z
Global Decoder
(a)
Global Encoder
$p_{\mathbf{z}|\mathbf{m},\mathbf{c}}$
$p_{\mathbf{A}|\mathbf{m},\mathbf{c}}$
sample
A
z
Global Decoder
(b)

Figure 9. Visualization of the reparameterization part (a) without variational attention and (b) with variational attention.

In our implementation, the global decoder is simply a linear transformation, and the global encoder is a stacked Bi-directional Long Short-Term Memory (Bi-LSTM) network. The parameters of the variational distributions $\mathbf{A}$ and $\mathbf{z}$ are calculated by a linear transformation to the outputs and the hidden states of the Bi-LSTM, respectively. We adopt the conditional version of MusicVAE (Roberts et al., 2018) proposed by Yang et al. (2019) for local encoders and decoders, using chord progressions as the condition for both of them. All local encoders share the same parameters, and the same for the local decoders.

4. Experiment

4.1. Dataset

Lack of datasets always poses severe issues for computer music. In this project, we will mainly use MIDI datasets for pop songs with vocal melodies, including the Nottingham dataset (Boulanger-Lewandowski et al., 2012). To make up for the insufficiency of training datasets, we also created a new dataset from more than 1000 Chinese Karaoke songs with symbolic annotations from the state-of-the-art Music Information Retrieval algorithms (Jiang et al., 2010; Böck et al., 2016; Mauch et al., 2015). The annotations are not 100% accurate, but our experiments still acquire good results with them. The whole dataset is augmented by pitch shifting with the range from -6 semitones to +5 semitones.

In the meanwhile, we collected another dataset from the website `www.hooktheory.com`. The website contains more than 10,000 public music tabs annotated by the crowd. Most annotations contain phrase-level predominant melody lines along with its chord and tonality context. Since the dataset is phrase-level, it becomes a more suitable dataset for the local structure experiment compared to the Nottingham and the Karaoke dataset, since the latter requires random splitting to get the phrase-level music pieces and we cannot guarantee that the cutting points are precisely the start/end of a phrase. Our final dataset contains 16,000 16-bar pop music samples.

4.2. Results

4.2.1. Experiment on Structured Condition

We first experiment the model with structured condition. We want to know if (1) the structured condition is effective in the context of VAE; (2) if the music structure in harmonic context can be used as an explicit representation of melody structure.

To show the effectiveness of condition, we train a model on 2-bar melody using hierarchical VAE where each local encoder encodes one beat of the melody and each global encoder output a global latent code to summarize the whole song. By providing harmonic context, we want the global latent code to focus more on the inner properties of the melody, e.g., pitch contour and rhythmic pattern, instead of the ones that depend heavily on the harmonic context.

Figure 10 show that when we alter the condition of the HVAE decoder and kept the latent code unchanged, the model will generate a melody with a similar contour to the original one while some pitches are slightly adjusted to be consistent with the harmonic context. This means the conditions are effective in the sense that the latent code does not represent the melody in a hard way by memorizing the exact pitch. Instead, it memorizes some abstract properties

Chord Condition	**Tonality Condition**	**KL Loss**	**Recon. Loss**
No	No	0.2254	0.0433
Yes	No	0.2154	0.0504
No	Yes	0.2191	0.0783
Yes	Yes	0.2143	0.0253

Table 1. KL Loss and Reconstruction loss on training over 2-bar melodies, with or without chord condition or tonality condition

of the melody like how the pitch contour goes.

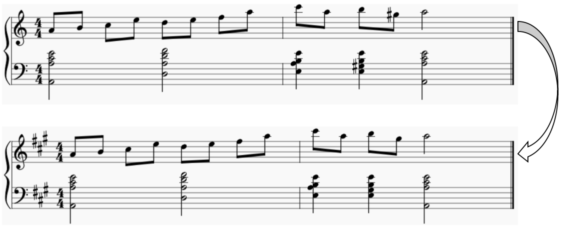


Figure 10. An example of manipulation of the structured condition. The top score contain the original music piece and the harmonic conditions. The bottom score contain the altered conditions and the generated melody with the altered chords and keys as the condition.[2]

However, when it comes to structure representation, we find that the harmonic conditions are not a suitable constraint to generate structural melodies. The reason is that the conditions are not powerful enough. By altering them, each note pitch only changes up to 1 to 2 semitones. This is barely enough to change the global structure of the melody. Also, we find the model trained with and without the conditions actually acquire similar performance with a similar KL loss, which means that the structure of the conditions does not help much on getting a more compact content representation of the melody. This observation gives us the idea that the structure information is actually stored inside of the latent variable, instead of the external conditions.

4.2.2. Experiment on Bar-Level Similarity Representation

We train the model proposed in 3.1 on 16-bar melody with the following hyper-parameters: (a) number of heads $H = 16$, (b) bar-level latent code dimension $z_{dim} = 512$, (c) $(\lambda_1, \lambda_2) = (1.0, 0.3)$. After training, the model acquires a high reconstruction accuracy (97.59%) on the validation set.

One advantage of the model is the fact that each attention matrix has a clear meaning corresponding to the self-similarity between two bars of the song. Therefore, we can visualize the attention matrix as shown in Figure 12.

However, the model has its limitation in two aspects. First, we find that the structure extracted by the model is incomplete. When we resample the latent variable, some of the global structure information is kept, but some others are lost. See Figure. 11 for such an example. Second, the model does not have a compact representation of the music content as each bar still has an individual latent variable, which might contain redundant information. Also, we cannot get a disentangled representation of music content from the latent representation.

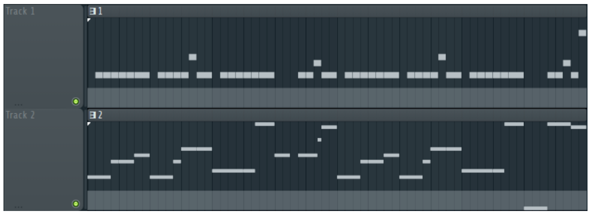


Figure 11. An example of generating a long music piece (the lower part) that has a similar structure to a given piece (the upper part). In generation, the latent code of the given piece is resampled but the MSASM is kept the same. The generated piece has a similar ABAC structure as the original one. However, some local structures (e.g., the similarity between the first 2 bars) are lost.

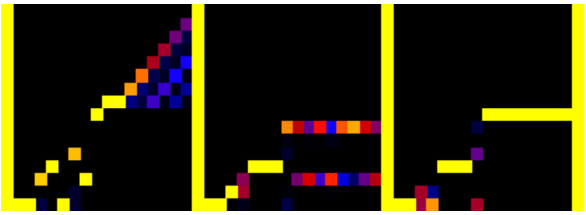


Figure 12. Parts (3 heads) of the MSASM values correspond to the song in Figure 11. Lighter colors in each matrix denote larger values and the vertical yellow bars are the separator for different matrices.

4.2.3. Experiment on Structure-Content Disentanglement

Sample length	**8 bars**	**16 bars**
HVAE	89.63	80.85
Proposed	93.82	86.27

Table 2. Reconstruction accuracy for the baseline model and the proposed model in section 3.3.

We train the model proposed in 3.3 with a latent dimension $z_{dim} = 512$, number of heads $H = 8$, number of components $C = 16$ and KL loss weight $(\beta_z, \beta_a) = (0.1, 2)$. We find the model hard to train as it tries to acquire a more compact music content representation $\mathbf{z}$. Still, we

Approach	**Structure Representation**	**Content Representation**	**Effectiveness on Structure Regularization**
Bar-Level Similarity	Implicit (Self-Attention)	Too Loose	Moderate
Structured Condition	Explicit (Harmonic Context)	Loose	Poor
Content-Structure Disentanglement	Implicit (Attention)	Compact	Good

Table 3. Comparison of the proposed methods

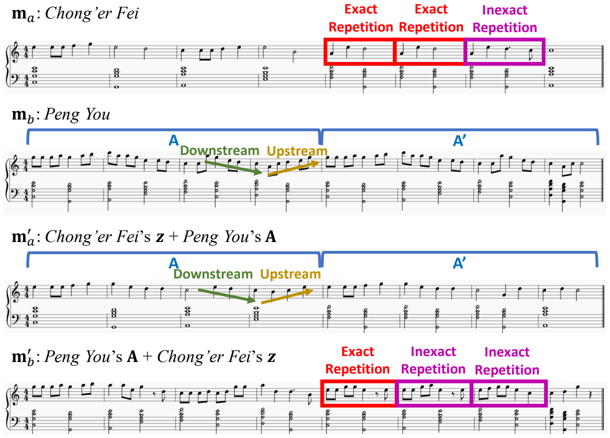

$\mathbf{m}_a$: *Chong'er Fei*
Exact Repetition
Exact Repetition
Inexact Repetition
$\mathbf{m}_b$: *Peng You*
A
Downstream
Upstream
A'
$\mathbf{m}'_a$: *Chong'er Fei*'s $\mathbf{z}$ + *Peng You*'s $\mathbf{A}$
A
Downstream
Upstream
A'
$\mathbf{m}'_b$: *Peng You*'s $\mathbf{A}$ + *Chong'er Fei*'s $\mathbf{z}$
Exact Repetition
Inexact Repetition
Inexact Repetition

Figure 13. A demonstration of the latent code exchange process on two Chinese pop songs *Chong'er Fei* and *Peng You*.[4]

find the variational attention helps by comparing the reconstruction accuracy to the pure hierarchical decoder by the original MusicVAE (Roberts et al., 2018) (the results are shown in Table 2). To validate the effectiveness of the content-structure disentanglement, we use latent-code swapping technique similar to the ones in (Natsume et al., 2018; Jetchev & Bergmann, 2017). Specifically, we are given two 8-bar melody lines $\mathbf{m}_a, \mathbf{m}_b$ from real music and calculate their latent representations:

$$(\mathbf{A}_a, \mathbf{z}_a) \leftarrow \text{Encode}(\mathbf{m}_a, \mathbf{c}_a) \quad (15)$$

$$(\mathbf{A}_b, \mathbf{z}_b) \leftarrow \text{Encode}(\mathbf{m}_b, \mathbf{c}_b) \quad (16)$$

where $\mathbf{c}_a$ ($\mathbf{c}_b$) denotes the corresponding chord condition to the melody $\mathbf{m}_a$ ($\mathbf{m}_b$). Then, we generate two new melodies by swapping one latent representation $\mathbf{A}_a, \mathbf{A}_b$:

$$\mathbf{m}'_a \leftarrow \text{Decode}(\mathbf{A}_b, \mathbf{z}_a, \mathbf{c}_a) \quad (17)$$

$$\mathbf{m}'_b \leftarrow \text{Decode}(\mathbf{A}_a, \mathbf{z}_b, \mathbf{c}_b) \quad (18)$$

Comparing the generated melody pieces to the original ones allow us to understand which features are controlled by $\mathbf{A}$ and $\mathbf{z}$, respectively. An example result is shown in Figure 13 . We experimented with 10 pop songs and observed some interesting results:

1. The generated melody $\mathbf{m}'_a$ shares a similar music structure as $\mathbf{m}_b$. The observed structure similarity is multi-level, including short-term and long-term repetitions.

2. The generated melody $\mathbf{m}'_a$ shows a similar rhythmic pattern and local pitch transition with $\mathbf{m}_a$.

We conclude from these findings that in the new model, the original latent code $\mathbf{z}$ along with the conditions controls the local *content* of the music, while the variational attention $\mathbf{A}$ controls the *structure* on how these local content should be arranged to a long music piece. This agrees with human's perception of music structure.

5. Discussion and Conclusion

In this paper, we design and experiment with several novel methods trying to find a suitable way to represent music structures, and furthermore trying to disentangle it from the representation of music content. The comparison of different methods is shown in Table 3.

While we find the third method works best, it still has several remaining issues. First, the model still experiences significant accuracy drop when generalizing to longer music. Second, as the proposed method do not have extra regularization on the two latent variables, the disentanglement of structure and content is actually unsupervised. Therefore, the effectiveness of disentanglement is generally hard to guarantee theoretically. Third, the network structure is only a prototype and need to be further optimized.

References

Bahuleyan, H., Mou, L., Vechtomova, O., and Poupart, P. Variational attention for sequence-to-sequence models. *arXiv preprint arXiv:1712.08207*, 2017a.

[4] The audio demo can be accessed via `https://trinket.io/library/trinkets/9ab0e87a21`.

Bahuleyan, H., Mou, L., Vechtomova, O., and Poupart, P. Variational attention for sequence-to-sequence models. *arXiv preprint arXiv:1712.08207*, 2017b.

Böck, S., Krebs, F., and Widmer, G. Joint beat and downbeat tracking with recurrent neural networks. In *ISMIR*, pp. 255–261, 2016.

Boulanger-Lewandowski, N., Bengio, Y., and Vincent, P. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. *arXiv preprint arXiv:1206.6392*, 2012.

Conklin, D. and Witten, I. H. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73, 1995.

Dai, S., Zhang, Z., and Xia, G. Music style transfer issues: A position paper. *arXiv preprint arXiv:1803.06841*, 2018.

Hadjeres, G., Pachet, F., and Nielsen, F. Deepbach: a steerable model for bach chorales generation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1362–1371. JMLR. org, 2017.

Hamanaka, M., Hirata, K., and Tojo, S. *deepGTTM-III: Multi-task Learning with Grouping and Metrical Structures: 13th International Symposium, CMMR 2017, Matosinhos, Portugal, September 25-28, 2017, Revised Selected Papers*, pp. 238–251. 01 2018. ISBN 978-3-030-01691-3. doi: 10.1007/978-3-030-01692-0_17.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A., Hoffman, M., Dinculescu, M., and Eck, D. Music transformer: Generating music with long-term structure. 2019.

Jetchev, N. and Bergmann, U. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2287–2292, 2017.

Jiang, J., Chen, K., Li, W., and Xia, G. Mirex 2018 submission: A structural chord representation for automatic large-vocabulary chord transcription. 2010.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lo, M. and Lucas, S. M. Evolving musical sequences with n-gram based trainable fitness functions. In *2006 IEEE International Conference on Evolutionary Computation*, pp. 601–608. IEEE, 2006.

Mao, H. H., Shin, T., and Cottrell, G. Deepj: Style-specific music generation. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 377–382. IEEE, 2018.

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., and Dixon, S. Computer-aided melody note transcription using the tony software: Accuracy and efficiency. 2015.

Natsume, R., Yatagawa, T., and Morishima, S. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018.

Paulus, J., Müller, M., and Klapuri, A. State of the art report: Audio-based music structure analysis. In *ISMIR*, pp. 625–636. Utrecht, 2010.

Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., Ellis, D. P., and Raffel, C. C. mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

Roberts, A., Engel, J., Raffel, C., Hawthorne, C., and Eck, D. A hierarchical latent vector model for learning long-term structure in music. *arXiv preprint arXiv:1803.05428*, 2018.

Thornton, C. Hierarchical markov modeling for generative music. In *ICMC*, 2009.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Yan, X., Yang, J., Sohn, K., and Lee, H. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision*, pp. 776–791. Springer, 2016.

Yang, L.-C., Chou, S.-Y., and Yang, Y.-H. Midinet: A convolutional generative adversarial network for symbolic-domain music generation. *arXiv preprint arXiv:1703.10847*, 2017.

Yang, R., Wang, D., Wang, Z., Chen, T., Jiang, J., and Xia, G. Deep music analogy via latent representation disentanglement. In *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, 2019.