Creative Improvised Interaction with Generative Musical Systems

Shlomo Dubnov, Gérard Assayag, Vignesh Gokul

Creative Improvised Interaction with Generative Musical Systems

Shlomo Dubnov
UC San Diego
9500 Gilman Dr., La Jolla, CA 92093
sdubnov@ucsd.edu

Gerard Assayag
IRCAM
1 Place Igor Stravinsky, Paris
gerard.assayag@ircam.fr

Vignesh Gokul
UC San Diego
9500 Gilman Dr., La Jolla, CA 92093
vgokul@eng.ucsd.edu

Abstract

In this paper we survey the methods for control and creative interaction with pre-trained generative models for audio and music. By using reduced (lossy) encoding and symbolization steps we are able to examine the level of information that is passing between the environment (the musician) and the agent (machine improvisation). We further use the concept of music information dynamics to find an optimal symbolization in terms of predictive information measure. Methods and strategies for generative models are surveyed in this paper and their implications for creative interaction with the machine are discussed in the musical improvisation framework.

1. Introduction

In our previous studies we developed SOTA generative music models, with focus on interactive machine improvisation that can learn musical style from live or off-line examples and then produce 'more of the same" [3]. This "same" was interesting in improvisation settings, since the variations maintained resemblance to the immediate expressions of the musician on stage, but were distinct enough to create interest and inspire new interaction. The problem in this setting was that it was the human who found interest in the machine imitation and changed his playing strategies by being inspired by the new machine generated materials, while the artificial agent's generation was oblivious to the musician. In order to allow for more interactivity, special tools were introduced, such as query-based improvisation that biased the choices of the artificial music generator towards materials that had more coherence with the human musician (finding hot spots in the model memory, query-matching, a-priori scenarios and more, see [28]). Nonetheless, these modifications to the generation policy of the artificial agent that were hard-wired, are largely insufficient to capture the complexity or the subtle expressive inflections in joint multi-musician improvisations. Experiments with multiple artificial musical agents that are capable of listening or influencing each other showed that varying interaction regimes has an important effect on creating interest and prolonging the interaction into a meaningful musical form [22].

The research objective of this inter-disciplinary project is to model and enhance co-creativity as it arises in improvised musical interactions between human and artificial agents, in a spectrum of practices spanning from interacting with software agents to mixed reality involving instrumental physicality and embodiment. Such creative interaction strongly involves co-improvisation, as a mixture of more or less predictable events, reactive and planned behaviors, discovery and action phases, states of volition or idleness. Improvisation is thus at the core of this project and indeed a fundamental constituent of co-creative musicianship, as well as a fascinating anthropological lever to human interactions in general. The outline of the project unfolds as follows:

- Understanding, modelling, implementing music generative and improvised interaction as a general template for symbiotic interaction between humans and digital systems (cyber-human systems)

- Creating the scientific and technological conditions for mixed reality musical systems, based on the interrelation of creative agents and active control in physical systems.

- Achieving distributed co-creativity through complex temporal adaptation of creative agents in live cyber-human systems, articulated to field experiment in musical social sciences.

This project exploring co-creativity in many dimensions of interaction, learning and generativity is notably linked to the European REACH (Raising Co-creativity in Cyber-human Musicianship) project involving the authors.

2. On Co-Creativity

The psychologist Margaret Boden has given much attention to the many relations between creativity and machines [8] For her, creativity is the ability to find new, surprising and socially valuable ideas or artifacts, and can occur in three main ways: it can be combinatorial (new configurations of known materials), exploratory (discovering new paths in conceptual / mental spaces) or transformative (when the space itself is disrupted giving way to ideas that were properly inconceivable before). But what is the situation when part of the creativity is delegated to machines, when manifestations of co-creativity emerge from symbolic interactions between human and artificial agents?

In addition to the novelty / effectiveness criteria, cyber-human co-creativity is strongly felt when two features of improvisation linked to emergence [9] and non-linear dynamics [27] are identified: (1) emergence of cohesive behaviors that are not reducible to, nor explainable by the mere individual processes of agents; (2) apparition of non-linear regimes of structure formation, leading to rich musical co-evolution of forms. In our work with jazz improvisers, Bernard Lubat mentions the machine seems, in his words, to "liberate" him, perhaps from specific habits or automatisms. In other words, our inner atlases can be roamed and even modified by creative thinking, in order for the "unthought" (or the yet unthinkable) to find its way.

By producing emergent information structures as a result of cyber-human interaction, we might achieve an epistemological leap [4] beyond the difficulty of conceding creativity to artificial systems, and assess that creativity is not a state anyway, but rather a dynamical effect of interaction in a complex system, showing radical novelty as a marker of emergence [12]. By building on this epistemological boost, one would be able to model deep interactions that in turn will trigger co-creative behaviors.

3. Architecture of Improvising Musical Agents

The architecture of an agent in the improvisation system that we develop is shown in figure 1 The different elements of the system comprise of the following:

- Musical signal : stream of audio or multimedia content

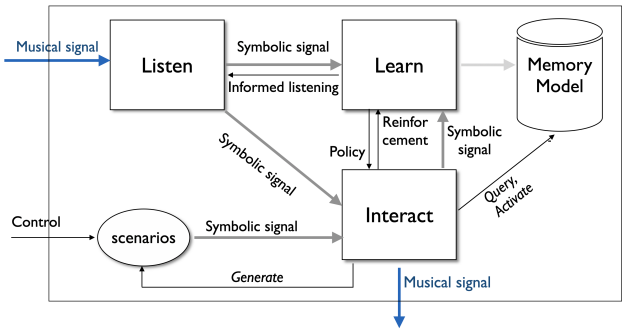

Figure 1. The flow of information between listen, learn and interact modules and the diverse data and control feed-backs channels in an agent

- Symbolic signal : stream of quantized units (audio descriptors, musical vocabulary, latent representations
- Informed listening : the more the structures are learned, the more powerful the predictions become to help machine listening recognize musical units
- Learn : statistical / deep modelling of musical structure and dynamics, reinforcement learning of interactive musical behaviour
- Memory model : variety of generative models for symbolic and audio signals, associated to activation states assessing the influence of the live environment on future predictions and their adequacy to musical input
- Interaction : the agent behaviour model; receive policies from the learning module ; queries the memory for generative content; assess influence from the external environment and weights on activation state of the memory; sends reinforcement signals to the learning module, follows or generate scenarios

This agent architecture can be replicated in a significant number of units in a multi-agent system, producing interaction between artificial agents as well as between agents and humans. Only musical and control signals are exchanged, with multiple cross-feedback loops (when agent A listens to agent B who listens to agent A etc.) that will promote and sustain emergence phases, such as in the Bayesian belief propagation scheme followed in [25]. Signal-symbol quantization [11], constitutes a critical part of the system as the symbolic signal constitute the main vector of information in the internal agent mechanism. Another important part is the capacity for an agent to not only "follow" the musical input from the context, but also to respect user defined scenario, and in more extreme cases, to incrementally generate scenarios by itself as in [10], where a LSTM with bottleneck

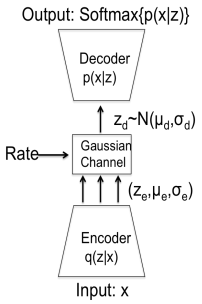

Figure 2. Noisy channel between encoder and decoder

encoder - decoder and teacher forcing algorithm is used to predict the next N chords of the harmony.

4 Reduced Representation

Studies of human cognition suggest Rate–Distortion as a way of extracting useful or meaningful information from noisy signals [30]. 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 [31]. 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[2],[14]. Accordingly, in order to achieve a better interaction between the human and the machine, we are seeking two types of data reduction:

- lossy representation of the signal (audio or midi) that effectively reduce the dimensionality of the latent representation and allow for better generalization
- symbolization of the lossy encoding to allow for better temporal representation by using language modeling with variable memory length

Learning music representation with auto-encoder, a schematic representation of the noise induced by bit-reduction is given by Figure 2. Performing finite bit-size encoding and transmission of the quantized latent values from encoder Z_e to decoder Z_d is not required, since we are interested in gating and biasing the original signal towards the prior distribution by encoding it at a limited bit-rate, which is given by the following optimal channel [7]

$$Q(z_d|z_e) = Normal(\mu_d, \sigma_d^2) \quad (1)$$

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

$$\sigma_d^2 = 2^{-4R}(2^{2R} - 1)\sigma_e^2 \quad (3)$$

Table 1. 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 left columns show the mutual information with different rates.

Scenario	R=10	R=100	R=1000	R=10000	Original
past-future	67.142	132.422	73.089	83.012	75.120
1st voice-2nd voice.	36.963	148.054	93.637	77.848	126.631
2nd voice-1st voice	61.643	91.893	104.920	91.821	82.037

To illustrate the effect of reduced data representation on predictive properties of music, we performed quantization at different rate for a monophonic encoding of music using a disentangled VAE training for a dataset of 14 two-part inventions composed by Johann Sebastian Bach. The MIDI files are collected from the *Complete Bach MIDI Index*[1]. Mutual information neural estimation (MINE) [6] was used to analyze the relations in time and across voices from their reduced latent representations.

From the results, we conclude that the mutual information values between conditional latent variables and predictive latent variables depend on the level of reduced representation. We find that reducing bit-allocation can effectively improve the mutual information between conditional latent variables and predictive latent variables for each scenario. For more details of this and other polyphonic and audio experiments we refer the readers to [17].

5. Musical Information Dynamics

Assuming the music signal $X = x[n]$ is encoded into a sequence of latent representations $Z = z[n]$, with n denoting discrete time step n. We would like to algorithmically discover the sequential structure of Z, and be able to present the structures quantitatively. Music Information Dynamics (MID)[1, 15, 29] provides a theoretical framework that utilizes mutual information between past and present observations to model the predictability of the signal. The advantage of adopting MID is that it optimizes or calculates an information theoretic measurements on the input sequence Z and is agnostic of specific sequence related applications, such as motifs discovery or structure segmentation. MID was shown to be important for understanding human perception of music in terms of anticipation and predictability [1, 15].

An efficient formal method for studying MID for sequence $Z[n]$ is the Information Rate (IR) that considers the relation between the present measurement $Z = z[n]$ and

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

it's past $\overleftarrow{Z} = z[1], z[2], \ldots, z[n], \ldots, z[N]$, formally defined as the maximum of mutual information over different quantized level of the sequence $S = Q(Z)$

$$IR(Z) = \max_{Q:S=Q(Z)} I(Q(Z), Q(\overleftarrow{Z})) \quad (4)$$

$$= H(S) - H(S|\overleftarrow{S}) \quad (5)$$

According to this measure, the maximal value of IR is obtained when the difference between the uncertainty of $H(S)$ and predictability $H(S|\overleftarrow{S})$ is at its greatest, meaning that there is a balance between variation and predictability. Quantization $Q(Z)$ is needed due to the need to detect inexact repetitions in the sequence Z, which in turn signifies the allowed level of similarities between observations in Z, or the amount of signal detail that is significant when comparing the present to the past.

6. Symbolization and Music Analysis using VMO

Variable Markov Oracle (VMO) [33] accepts a representation $Z = z[1], z[2], \ldots, z[N]$ and turns it into a symbolic sequence $S = s[1], s[2], \ldots, s[n], \ldots, s[N]$, with M states over a finite alphabet Σ. The labels are formed by finding suffixes in a graph structure constructed by the VMO algorithm. Due to space consideration, we leave out the VMO construction and refer the readers to [16, 34]. The essential step in symbolization is finding a threshold with the highest MID value. The threshold θ partitions the space of features into categories that capture and represent the different sound elements by determining if the incoming $z[n]$ is similar to one of the frames following previous instances in the sequence pointed to be a suffix link from $n-1$. VMO symbolization step assigns two frames $z[i]$ and $z[j]$ the same label $s[i] = s[j] \in \Sigma$ if $||z[i] - z[j]|| \leq \theta$. To find the optimal threshold θ, MID measure can be estimated by any predictive compression algorithm $C(\cdot)$. The compression gain over blocks of symbols is used to replace the the entropy term $H(\cdot)$ as our measure of complexity[26]

$$IR(Z) = \max_{\theta, s[n] \in \Sigma_\theta} [C(s[n]) - C(s[n]|\overleftarrow{S})]. \quad (6)$$

It should be noted that the alphabet out of the quantization is constructed dynamically, as new labels can be added when an input sample cannot be assigned to one of the existing clusters of samples already labeled by existing labels.

As an example of the effect of different levels of symbolization on discovery of the motif structure in different musical styles, we performed comparative analysis of several works for the flute [18] using human engineered (Chroma and MFCC) and machine learned representations (VAE). We provide a partial example of the finding in the figure 3. It can be seen that the Dongxiao music is characterized by much shorter motifs, which were found at much finer threshold value compared to Telemann that is characterized by longer motifs that required a coarse quantization. In a different work, a VMO based MID estimator was used to evaluate the performance of generative recurrent latent models for MIDI data. The results showed that Variational encoding, which added randomization into the latent representation of the generative model, resulted in an improved motif structure of musical generation output as it better resembled the motif statistics found in the original data compared to RNN methods that tend to overly reproduce repetitions having looping sub-sequences[19].

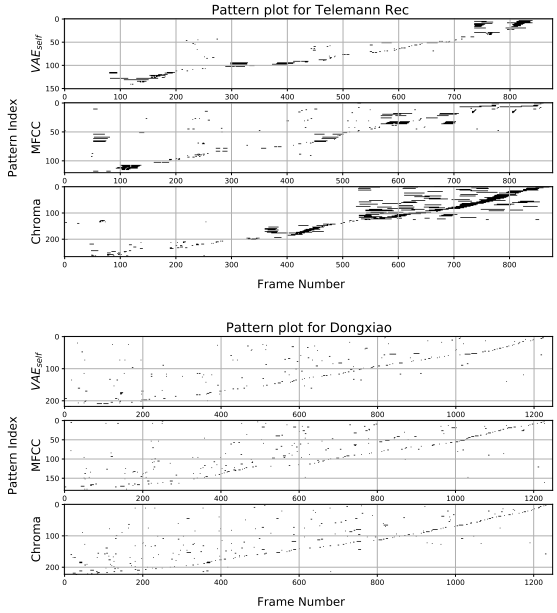

Figure 3. Motifs found in different flute pieces using the best VMO for VAE_{self}, Chroma features, and MFCCs.

7. RL tuning of musical generators

There have been a lot of research in modelling music generation and synthesis as a Maximum Likelihood Estimation (MLE) problem [35, 13, 20], where the objective is to predict the next best likely note/symbol. However, one of the key issues is exposure bias -the training and testing conditions might be completely different, due to the vast possibilities in music improvisation. Hence, we need a mechanism to facilitate zero-shot learning. Another issue is Greedy vs Dynamic Programming - in MLE, the approach

is greedy i.e the note/symbol which is best suited for the current time-step is chosen. This might cause issues in long-term music generation, as we need a solution to optimize for long term rewards (dynamic programming). We solve the above issues by treating music improvisation/generation as a reinforcement learning problem. An agent maximizes a long-term discounted reward function by exploring different states. In order to achieve this, some key challenges in applying RL to music need to be addressed:

- Representation of States: a trivial solutions could be representing a state as the sequence of notes played so far, or sequence of past spectral feature, symbolized into a discrete representation. This poses a problem as we need a state space independent of the previous observations, so as to allow the musical agent to explore.

- Reward function: Another key aspect of RL is finding a good reward function. The rewards are responsible for tuning the agent towards 'good behavior'. In an open-ended topic such as music improvisation, it can be very hard to quantify what 'good music' is across different styles and cultures.

Recent works [13] have used similar representations to model raw audio. They use MLE to synthesize audio sequentially, by fitting another model, say a transformer, on top of these learned representations. This is extremely time consuming and also computationally expensive. Since, our goal is to have real-time music improvisation, we need to parallelize this synthesis process.

We address the problem of representation by considering VMO symbolization and using ideas from VQ-VAE [32] and VQ-GAN [21]. We use raw audio signals as the training dataset to train a codebook model to obtain the vector-quantized embeddings (codified audio) as our representation. The representation of a state would be the combined sequences of vector-quantized embeddings of both the user input audio and the machine improvised audio observed so far. More formally, given a raw audio signal (user input) x and a VQ model E, we represent the signal as $E(x) = [h_1, h_2, h_3, h_4, ..., h_n]$, where h_1,h_2,h_n are indices corresponding to the embedding vectors in the codebook. Similarly, we represent the generated music signal as a sequence of codes $[m_1, m_2, ..., m_n]$. To represent a state s_t, we use a concatenated representation of $[s_{ht}, s_{mt}]$ where $s_{ht} = h_{0:t-1}$ and $s_{mt} = m_{0:t-1}$.

We design our rewards based on recent works on using RL for music synthesis [24, 23]. Given state s_t at timestep t, our reward agent assigns a score based on $p(m_{t-\Delta t} : m_{t+\Delta t}|s_t)$, where m_t denotes the codified audio generated by our synthesizer at timestep t and Δt is a parameter to control the window size. Note that this is different than MLE, as the RL agent tries to maximize the discounted reward over the entire sequence, which accounts for long-term structure and eliminates the problem of greedy selection at every timestep. The effect of discounted rewards is demonstrated in figure 4. It is evident that the top example is repetitive while RL creates more varied musical structure.

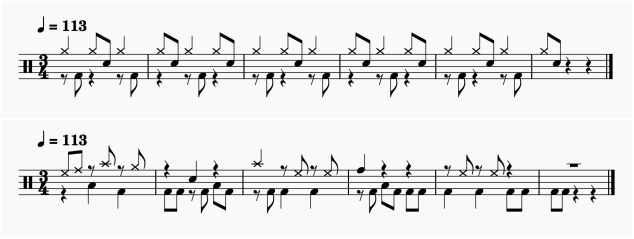

Figure 4. Percussion phrase generated with CNN trained on music notation, without (top) and with (bottom) RL refinement

For negative reinforcement, the reward agent penalizes compositions, where the pitch remains the same for more than n timesteps. Our framework also allows for customization where the rules of different synthesizers can be customized according to the needs of the user. The reward agent for the composer is based on the mutual information between the observed input signal and the current signal of the music agent.

For our synthesis component, we choose a RL module, where each action is replaced with skills (options)[5]. We can think about these skills as a series of temporally extended courses of actions. Our RL agent consists of two components: lower-level components (specializers) that are specialized in a skill, and a higher level component (composer) that synthesizes improvisations by using a combination of the lower level components. Each of the specializers are generators that output sequences of vector-quantized codes such that the decoded raw audio satisfies a particular rule, i.e optimizes for a reward function. The composer is responsible for selection of options or combining these options to compose improvisations to human-played compositions at real-time.

References

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

[2] A. A. Alemi. Variational predictive information bottleneck. In *AABI*, 2019.

[3] G. Assayag, M. C. Georges Bloch, A. Cont, and S. Dubnov. Omax brothers : a dynamic topology of agents for improvization learning. In *Workshop on Audio and Music Computing for Multimedia, ACM Multimedia*, 2006.

[4] G. Bachelard. *La formation de l'esprit scientifique*. (reprint. Paris PUF coll. « Quadrige », 2013), Paris, Alcan, 1934.

[5] A. Barreto, D. Borsa, S. Hou, G. Comanici, E. Aygün, P. Hamel, D. Toyama, S. Mourad, D. Silver, D. Precup, et al. The option keyboard: Combining skills in reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, R. D. Hjelm, and A. C. Courville. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning, (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 530–539. PMLR, 2018.

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

[8] M. Boden. Computers models of creativity. 30:23–34, 2009.

[9] C. Canonne and N. Garnier. A model for collective free improvisation. In *Mathematics and Computation in Music. Third International Conference MCM*, IRCAM, Paris, France, June 15-17 2011.

[10] T. Carsault, J. Nika, P. Esling, and G. Assayag. Combining real-time extraction and prediction of musical chord progressions for creative applications. *Electronics, MDPI*, 10(21):2634, 2021.

[11] A. Chemla-Romeu-Santos, S. Ntalampiras, P. Esling, G. Haus, and G. Assayag. Cross-modal variational inference for bijective signal-symbol translation. In *Proceedings of the 22 nd International Conference on Digital Audio Effects (DAFx-19)*, 2019.

[12] P. A. Corning. The re-emergence of "emergence": A venerable concept in search of a theory. *Complexity*, 7(6):18–30, 2002.

[13] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020.

[14] Z. Dong, D. Oktay, B. Poole, and A. A. Alemi. On predictive information in rnns, 2020.

[15] S. Dubnov. Musical information dynamics as models of auditory anticipation. In W. Wang, editor, *Machine Audition: Principles, Algorithms and Systems.*, pages 371–397. IGI Global, 2011.

[16] S. Dubnov, G. Assayag, and A. Cont. Audio oracle analysis of musical information rate. In *IEEE International Conference on Semantic Computing (ICSC)*, pages 567–571. IEEE, 2011.

[17] S. Dubnov, K. Chen, and K. Huang. Deep music information dynamics: Novel framework for reduced neural-network music representation with applications to midi and audio analysis and improvisation. *Journal of Creative Musical Systems*, 2022.

[18] S. Dubnov, K. Huang, and C. Wang. Towards cross-cultural analysis using music information dynamics. *arXiv preprint arXiv:2111.12588*, 2021.

[19] D. W. ES Koh, S Dubnov. Rethinking recurrent latent variable model for music composition. In *IEEE 20th International Workshop on Multimedia*, 2018.

[20] C.-Z. A. Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.

[21] V. Iashin and E. Rahtu. Taming visually guided sound generation. *arXiv preprint arXiv:2110.08791*, 2021.

[22] G. A. I. in Creative Symbolic Interaction. *Mathematical Conversations : Mathematics and Computation in Music Performance and Composition*. World Scientific; Imperial College Press, 2016.

[23] N. Jaques, S. Gu, R. E. Turner, and D. Eck. Tuning recurrent neural networks with reinforcement learning. 2017.

[24] N. Jiang, S. Jin, Z. Duan, and C. Zhang. Rl-duet: Online music accompaniment generation using deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 710–718, 2020.

[25] G. A. Ken Déguernel, Emmanuel Vincent. Probabilistic factor oracles for multidimensional machine improvisation. *Computer Music Journal*, 42(2):52–66, 2018.

[26] A. Lefebvre and T. Lecroq. Compror: on-line lossless data compression with a factor oracle. *Information Processing Letters*, 83(1):1–6, 2002.

[27] P. Mouawad and S. Dubnov. On modeling affect in audio with non-linear symbolic dynamics. *Advances in Science, Technology and Engineering Systems Journal*, 2(3):1727–1740, 2017.

[28] J. Nika, K. Déguernel, A. Chemla-Romeu-Santos, E. Vincent, and G. Assayag. Dyci2 agents: merging the "free", "reactive", and "scenario-based" music generation paradigms. In *International Computer Music Conference*, Shangai, China, Oct 2017.

[29] M. T. Pearce. Statistical learning and probabilistic prediction in music cognition: mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, 1423(1):378, 2018.

[30] C. R.Sims. Rate–distortion theory and human perception. *Cognition*, 152:181–198, 2016.

[31] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. 2015.

[32] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[33] C. Wang and S. Dubnov. Guided music synthesis with variable markov oracle. In *The 3rd International Workshop on Musical Metacreation, 10th Artificial Intelligence and Interactive Digital Entertainment Conference*, 2014.

[34] C.-i. Wang, J. Hsu, and S. Dubnov. Music pattern discovery with variable markov oracle: A unified approach to symbolic and audio representations. In *International Society for Music Information Retrieval Conference*, pages 176–182, 2015.

[35] H. Zhu, Q. Liu, N. J. Yuan, C. Qin, J. Li, K. Zhang, G. Zhou, F. Wei, Y. Xu, and E. Chen. Xiaoice band: A melody and arrangement generation framework for pop music. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2837–2846, 2018.