Harmonic Recomposition using Conditional Autoregressive Modeling

Kyle Kastner [1] **Rithesh Kumar** [1] **Tim Cooijmans** [1] **Aaron Courville** [1]

Abstract

We demonstrate a conditional autoregressive pipeline for efficient music composition based on work presented in van den Oord et al. (2017). Applying these methods to the task of *harmonic recomposition*, we generate diverse and structured recompositions which can be dynamically conditioned on chord level annotations.

1. Introduction

Since the earliest days of computation, composers have explored methods of combining aleatoric music and algorithmic composition with the power of generic computing devices (Agon et al., 2003; Cope, 1989). Authors have taken a wide variety of data driven approaches to "creative generation" in various domains (Barbieri et al., 2012; Ha & Eck, 2017; Graves, 2013), with extensive application to music modeling (Briot et al., 2017; Roberts et al., 2018; Eck & Schmidhuber, 2002; Sturm et al., 2015; Hadjeres et al., 2016; Boulanger-Lewandowski et al., 2012; Bretan et al., 2017; Roberts et al., 2018).

In this paper, we focus on the task of *harmonic recomposition* (Casal & Casey, 2010), where *recomposition* indicates more intricate conditional input when compared to standard conditional generation. Melody generation and evaluation is a difficult task in itself, even in monophonic music (Jaques et al., 2016), so we use the term *harmonic* recomposition primarily to reference aspects of note agreement and structure between voices. Although melodic trends still emerge from the principles of voice leading, we do not focus directly on melody composition.

We find the best recomposition results utilize chord annotations which involve knowledge of voicing and inversion. This implies some level of compositional input or reuse of prior composition blocks (such as segments taken from existing works) is needed for generating quality output. Our pipeline is also potentially applicable to purely sequential and iterative generation with different conditional inputs, as has been shown in prior work (Huang et al., 2017).

1.1. Related Work

Autoregressive models have proven to be powerful distribution estimators for images and sequence data, showing excellent results in a wide array of generative settings (van den Oord et al., 2016). They have also performed well in related prior work for polyphonic music generation. Most related to the work described in this paper is CoCoNet (Huang et al., 2017; 2018), which also uses an autoregressive convolutional model over image-like structures for polyphonic music generation and was a direct inspiration for our approach. One key difference of our approach is our utilization of a two stage pipeline (first seen in the work of van den Oord et al. (2017)) which greatly improves training and generation speed and also creates an implicit separation between local voice agreement (largely handled in the first stage) and global consistency over measures (handled in the conditional generation of the second stage). We also apply chord function conditioning specifically for application to recomposition.

2. Implementation Details

In this section, we describe the data, model, and training details for our recomposition approach. An open source implementation of our setup (including audio samples) is available online[1].

2.1. Data

For our raw data source, we use a subset of the scores associated with the composer Josquin des Prez as compiled by the Josquin project [2]. We utilize only pieces with 4 parts, resulting in a dataset with 252 pieces comprising approximately 16000 measures. We hold a contiguous 1000 measures out for validation and for use as a source of harmonic chord sequences during conditional generation.

After extracting individual measures, we convert to a piano

[1]Montréal Institute for Learning Algorithms (MILA), Université de Montréal, Montréal, Québec, Canada. Correspondence to: Kyle Kastner <kyle.kastner@umontreal.ca>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

[1]https://github.com/kastnerkyle/harmonic_recomposition_workshop

[2]http://josquin.stanford.edu/

roll style multichannel image, with each measure having 48 quantized timesteps (regardless of time-signature) on the horizontal axis, and one of 88 possible tones on the vertical axis. We zero pad these 88 values to 96 and assign each voice its own channel in the image-like container of size $(N, 96, 48, 4)$ described in examples, height, width, and channels format (NHWC).

2.2. Conditional Information

We extract the chord function and voicing of all measures using the music21 software package (Cuthbert & Ariza, 2010). This chord information allows us to condition the generation not only on the particular realization of the measure, but also the harmonic function of this measure in the larger piece. We combine the function of both the previous and next measure with the function and full voicing of the current measure to generate. For example, a measure group of I, ii_2^4, V_5^6, I, chords would form triplet groupings of I, I, ii (we repeat chords to handle border issues), then I, ii_2^4, V, and finally ii, V_5^6, V. Each triplet grouping has a unique integer index, for use as a conditional input in our model.

2.3. Models

The model pipeline is a two-stage generative setup, as described by van den Oord et al. (2017), wherein an initial encode/decode stage (denoted VQ-VAE) is trained to compress large initial inputs to a spatially reduced, discrete representation (which we call Z), and uncompress back to a reconstructed image. Once the VQ-VAE stage is trained, we use it to generate a compressed Z for each element in the dataset, and train an autoregressive generative "prior model" on this representation. The prior model learns to generate components of Z (which takes the form of a $(12, 6)$ spatial map in this work), denoted $Z_{i,j}$, one at a time conditioning the next generation cell on all previous.

The prior model may also take conditioning as a vector (embedding indexed by triplet chord integer), a spatial map (a Z from some previous generation), or a combination of both during the generation process.

2.4. Experiment Details

The first stage VQ-VAE consists of 3 strided convolutional layers of kernel size $(4, 4)$ and a stride of 2 on both the horizontal and vertical axes, followed by an additional $(4, 4)$ convolution with stride 1. These layers have sizes 64, 128, 256, and 256 leading to a VQ codebook size of 256, which results in a latent Z dimension of $(12, 6)$ for size $(96, 48, 4)$ input. This procedure is inverted using transpose convolution for the decoder and combined with a binary cross-entropy loss, averaged over all channels and spatial dimensions. Training was performed over 50000 minibatches of size 64 with an Adam optimizer (settings are $\alpha = 0.0002$, $\beta 1 = 0.9$, $\beta 2 = 0.999$, $\epsilon = 1e - 08$).

In the second stage, the gated conditional PixelCNN is configured with 15 layers of 64 projection channels. The first layer has a kernel size of $(5, 5)$ and no residual connection. All layers after the first utilize residual connections and have a kernel size of $(3, 3)$. Training was performed over 100000 minibatches of size 50 with an Adam optimizer (configured as before) with categorical cross-entropy loss averaged over the output prediction dimension of 256 (size of the VQ codebook).

3. Results

We experiment with two types of conditioning combined with the aforementioned architecture. The higher level information contained in the chord sequences alone seems sufficient to produce directed, coherent trajectories. When spatial conditioning from the previous timestep is included, the resulting generations tend to be overly smooth, punctuated by dissonant intervals. Finding better ways to combine local note level information with chord annotation will be an important step to improving this pipeline.

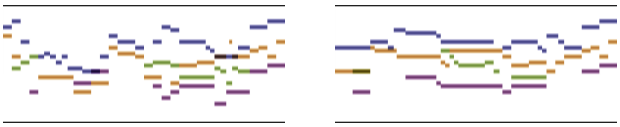

Figure 1. Two 8 bar sequences generated over conditioning sequence (III),v,i,v_6,v,bVI_{532},III_6,v,$bVII_6$,$(bVII)$, using different random seeds

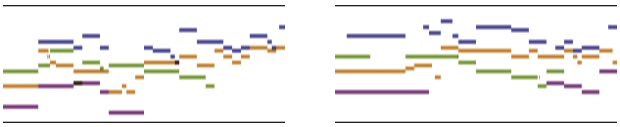

Figure 2. 8 bar sequences generated over conditioning sequence (ii_6),vi_5,$vi7$,ii,V,vi,iii,ii,vi,$(viio_4)$. *Left*: only chord triplet conditioning. *Right*: chord triplet conditioning along with previous generation as conditioning spatial map.

4. Conclusion

Chord-conditional generative models are an ideal fit for harmonic recomposition. We find that the two-stage pipeline of van den Oord et al. (2017) captures musical structure in a computationally efficient manner. Our work demonstrates diverse and interesting proposals for note-level realizations of given chordal sequences and provides an open-source implementation with examples.

References

Agon, Carlos, Andreatta, Moreno, Assayag, Gérard, and Schaub, Stephan. Formal aspects of Iannis Xenakis' Symbolic Music: a computer-aided exploration of some compositional processes. *Journal of New Music Research*, -(0):–, 2003. cote interne IRCAM: Agon03a.

Barbieri, Gabriele, Pachet, François, Roy, Pierre, and Esposti, Mirko Degli. Markov constraints for generating lyrics with style. In *Proceedings of the 20th European Conference on Artificial Intelligence*, ECAI'12, pp. 115–120, Amsterdam, The Netherlands, The Netherlands, 2012. IOS Press. ISBN 978-1-61499-097-0. doi: 10.3233/978-1-61499-098-7-115.

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

Bretan, Mason, Oore, Sageev, Engel, Jesse, Eck, Douglas, and Heck, Larry P. Deep music: Towards musical dialogue. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pp. 5081–5082, 2017.

Briot, Jean-Pierre, Hadjeres, Gaëtan, and Pachet, François. Deep learning techniques for music generation - A survey. *CoRR*, abs/1709.01620, 2017.

Casal, David Plans and Casey, Michael. Decomposing autumn: A component-wise recomposition. In *ICMC*, 2010.

Cope, David. Experiments in musical intelligence (emi): Nonlinear linguisticbased composition. *Interface*, 18(1-2):117–139, 1989. doi: 10.1080/09298218908570541.

Cuthbert, Michael Scott and Ariza, Christopher. Music21: A toolkit for computer-aided musicology and symbolic music data. In *Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR 2010, Utrecht, Netherlands, August 9-13, 2010*, pp. 637–642, 2010.

Eck, Douglas and Schmidhuber, Juergen. Learning the long-term structure of the blues. In Dorronsoro, J. (ed.), *Artificial Neural Networks – ICANN 2002 (Proceedings)*, pp. 284–289, Berlin, 2002. Springer.

Graves, A. Generating sequences with recurrent neural networks. *arXiv:*`1308.0850 [cs.NE]`, August 2013.

Ha, David and Eck, Douglas. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

Hadjeres, Gaëtan, Pachet, François, and Nielsen, Frank. Deepbach: a steerable model for bach chorales generation. *arXiv preprint arXiv:1612.01010*, 2016.

Huang, Anna, Chen, Sherol, Nelson, Mark, and Eck, Douglas. Towards mixed-initiative generation of multichannel sequential structure. 2018.

Huang, Cheng-Zhi Anna, Cooijmans, Tim, Roberts, Adam, Courville, Aaron, and Eck, Douglas. Counterpoint by convolution. In *Proceedings of ISMIR 2017*, 2017.

Jaques, Natasha, Gu, Shixiang, Bahdanau, Dzmitry, Hernández-Lobato, José Miguel, Turner, Richard E, and Eck, Douglas. Sequence tutor: Conservative fine-tuning of sequence generation models with kl-control. *arXiv preprint arXiv:1611.02796*, 2016.

Raffel, Colin and Ellis, Daniel PW. Large-scale content-based matching of midi and audio files. In *ISMIR*, pp. 234–240, 2015.

Roberts, Adam, Engel, Jesse, Raffel, Colin, Hawthorne, Curtis, and Eck, Douglas. A hierarchical latent vector model for learning long-term structure in music. *CoRR*, abs/1803.05428, 2018.

Sturm, Bob, Santos, João Felipe, and Korshunova, Iryna. Folk music style modelling by recurrent neural networks with long short term memory units. In *16th International Society for Music Information Retrieval Conference, late-breaking demo session*, pp. 2, 2015.

van den Oord, Aaron, Dieleman, Sander, Zen, Heiga, Simonyan, Karen, Vinyals, Oriol, Graves, Alex, Kalchbrenner, Nal, Senior, Andrew, and Kavukcuoglu, Koray. Wavenet: A generative model for raw audio. 2016.

van den Oord, Aaron, Vinyals, Oriol, and kavukcuoglu, koray. Neural discrete representation learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 6306–6315. Curran Associates, Inc., 2017.