MELODY HARMONIZATION USING ORDERLESS NADE, CHORD BALANCING, AND BLOCKED GIBBS SAMPLING

Chung-En Sun, Yi-Wei Chen, Hung-Shin Lee, Yen-Hsing Chen, Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

Coherence and interestingness are two criteria for evaluating the performance of melody harmonization, which aims to generate a chord progression from a symbolic melody. In this study, we apply the concept of orderless NADE, which takes the melody and its partially masked chord sequence as the input of the BiLSTM-based networks to learn the masked ground truth, to the training process. In addition, the class weights are used to compensate for some reasonable chord labels that are rarely seen in the training set. Consistent with the stochasticity in training, blocked Gibbs sampling with proper numbers of masking/generating loops is used in the inference phase to progressively trade the coherence of the generated chord sequence off against its interestingness. The experiments were conducted on a dataset of 18,005 melody/chord pairs. Our proposed model outperforms the state-of-the-art system MTHarmonizer in five of six different objective metrics based on chord/melody harmonicity and chord progression. The subjective test results with more than 100 participants also show the superiority of our model.

Index Terms— Melody harmonization, orderless NADE, blocked Gibbs sampling, sample weighting.

1. INTRODUCTION

Automatic melody harmonization aims to build up a learning machine that can generate chord sequences to accompany a given melody [1, 2]. In music, a chord is any harmonic set composed of three or more notes that are heard as if sounding simultaneously [3]. However, melody harmonization is a highly subjective task that is difficult for machines to learn to imitate humans or fit human preferences. While some harmonization may be considered appropriate in some cases, it can be changed according to different contexts.

Many approaches have been proposed for automatic melody harmonization [3], such as hidden Markov models (HMMs) [4, 5, 6] and genetic algorithm (GA)-based methods [7]. Recently, with the prevalence of deep learning models, some deep learning methods have emerged to deal with the melody harmonization problem [8]. Lim *et al.* [9] first proposed a model composed of bidirectional long short-term memory (BiLSTM) layers [10]. The model predicts a chord for each bar from 24 major and minor chords. Based on the same neural structure, Yeh *et al.* extended the number of chords to 48 by considering major, minor, augment and diminish chords. In addition, they integrated information of chord functions [11] into the loss function to help chord label prediction, and predicted a chord every half bar [12]. Experiments showed that their model, called MTHarmonizer, could generate more creative and elaborated chord progressions than the previous model. Not only the expanded diversity of chord labels, the reason also lies in that MTHarmonizer can handle the transition between tonic (T), subdominant (S), and dominant (D) chords due to the use of chord functions.

However, there are several drawbacks in the above models. For example, Lim's model overuses common chords and has an incorrect phrasing problem [12]. Yeh's model tried to deal with these problems, but still could not appropriately assign altered chords through chord function prediction, because the chord function (TSD) easily loses its function when a secondary or borrowed chord event occurs.

In this study, we propose a BiLSTM-based model stemming from the structure of orderless NADE [13, 14], which takes the melody and its partially masked chord sequence as the input to learn the masked ground truth. Instead of chord functions, class weights are used to compensate for some reasonable chord labels that are rarely seen in the training set. In addition, we perform blocked Gibbs sampling during inference [15], similar to CoCoNet used in Google's Bach Doodle [16, 17]. As a result, our model can predict 96 chords, including major, minor, augment, diminish, suspend, major7, minor7, and dominant7. Through the chord balancing process in training, we finally achieve a generative model that can produce rich but still reasonable chord progressions. Such interesting progressions with sufficient tension have a close quality to that of human-composed progressions.

2. PROPOSED MODEL

2.1. Neural Structure

Similar to Lim's [9] and Yeh's [12] models, the neural structure of our model is mainly composed of two BiLSTM layers, followed by a fully connected layer, as shown in Fig. 1. In the training phase, dropout with a probability of 0.2 is applied

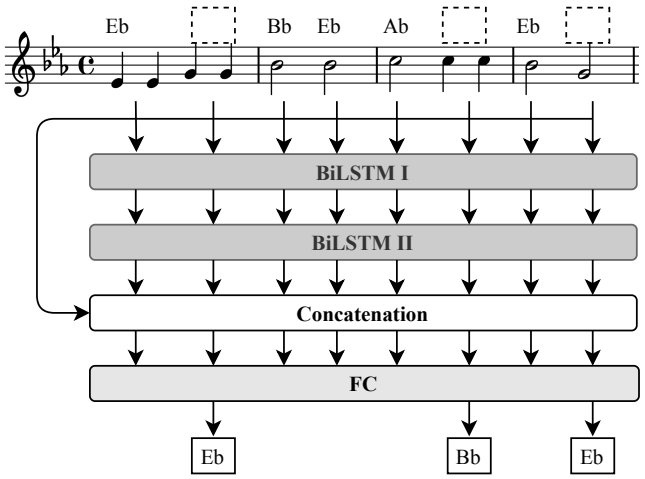


Fig. 1. Proposed neural structure for melody harmonization. The input of each hidden state of BiLSTM I is a melody sequence within half a bar, concatenated with a chord that may be masked (dashed boxes) and the mask itself. The output of BiLSTM II is concatenated with the input to form a residual-like network [21], and then fed into a fully-connected (FC) layer for final prediction.

to each hidden layer [18], and mini-batch gradient descent is used to minimize the categorical cross entropy [19]. We use Adam as the optimizer with a learning rate of 0.005 and early stopping at the 10-th epoch [20].

However, we employ the technique of orderless NADE to train our model in order to approximate the complex joint probability between melody and chords as well as chords and chords [13, 14]. Given any arbitrary context C, an orderless NADE model offers direct access to all distributions of the form $p_\theta(x_i|x_C), i \notin C$, where θ denotes the model parameters and x_i denotes the i-th variable that is not in the set of contextual variables x_C that are already known. We can use it to complete any arbitrarily partial chord progression.

The training procedure is as follows. First, there are three tensors at the input, including a one-hot encoding tensor $x_{\text{melody}} \in \{0,1\}^{T\times|P|}$, a one-hot encoding tensor $x_{\text{chord}} \in \{0,1\}^{T\times|S|}$, and a binary mask $x_{\text{mask}} \in \{0,1\}^{T\times 1}$. T denotes the number of time steps. The number of pitches $|P|$, i.e., the number of notes in the chromatic scale, is set to 12. The number of chords $|S|$ is set to 96. x_{mask} is used to mask some parts of the x_{chord} and provides the model with information about which parts are masked:

$$x_{\text{chord}_C} = x_{\text{chord}} \circ x_{\text{mask}}, \tag{1}$$

where $x_{\text{chord}_C} = \{x_{\text{chord}_t} | t \in C\}$ is a strict subset of x_{chord} based on context C, and $\circ$ denotes the Hadamard product.

Therefore, the model is only provided some part of x_{chord}. The positions of x_{chord} to be masked are randomly selected, and the masking ratio follows a uniform distribution between 0 and 1. The three tensors are further concatenated to form the model input x_{input} as follows:

$$x_{\text{input}} = \text{concat}(x_{\text{melody}}, x_{\text{chord}_C}, x_{\text{mask}}) \tag{2}$$

Finally, the task of our model is to reconstruct $x_{\text{chord}_{\neg C}}$, the complement of x_{chord_C}, from x_{input}.

2.2. Loss Function

The loss function with negative log-likelihood is expressed by

$$\mathcal{L}(x_{\text{melody}}, x_{\text{chord}}; C, \theta) = - \sum_{t\notin C} \sum_{s} w_s x_{\text{chord}_{t,s}} \log p_\theta(x_{\text{chord}_{t,s}} | x_{\text{melody}}, x_{\text{chord}_C}, C) \tag{3}$$

where w_s is the weight of chord s [22], $x_{\text{chord}_{t,s}}$ is the s-th element of the one-hot vector x_{chord_t}, C is the context given, and θ denotes the model parameters. Chord weighting can greatly solve the problem of excessive use of common chords in Lim's model [9] like Yeh's chord function [12]. However, our model can further manage to assign appropriate altered chords when the melody requires a secondary or borrowed chord to elaborate the whole piece, but Yeh's model [12] cannot. The weight matrix is derived by

$$w_s = |S| \times \frac{1/(1000 + n_s)}{\sum_{s'} 1/(1000 + n_{s'})} \tag{4}$$

where n_s is the number of times chord s appears in the training data, and n_s is added by 1000 for count smoothing. This process is also called inverse class frequency weighting.

The model is trained by minimizing

$$\mathbb{E}_{x_{\text{chord}}\sim p(x_{\text{chord}})} \mathbb{E}_{C\sim p(C)} \frac{1}{|\neg C|} \mathcal{L}(x_{\text{melody}}, x_{\text{chord}}; C, \theta). \tag{5}$$

The expectations are estimated by sampling x_{chord} from the training set and selecting a single context C per sample.

2.3. Inference

The original orderless NADE model uses ancestral sampling in the inference stage [13, 14], which samples one variable at a time in any order. Therefore, we can randomly choose an order, and then sample all the variables according to that order. However, this process may cause errors to propagate through that specific order. Furthermore, it is shown in [16] that some $p_\theta(x_{\text{chord}_t}|x_C, C), t \notin C$, may be poorly modeled, especially when C is small.

Therefore, we use blocked Gibbs sampling [15], a similar sampling procedure in [16, 17]. The blocked Gibbs sampling is an orderless sampling process that perfectly matches the training process of our model. There is no need to select an order first; instead, we randomly choose the blocks to be masked and ask the model to reconstruct them in each iteration. There is an annealing probability to determine the proportion of the variables that remain unchanged. At first, all the variables will be erased and sampled again. As the iteration goes on, more and more variables will be kept unchanged. Eventually, the output will converge and the joint

Algorithm 1 Blocked Gibbs Sampling

Input: A melody x_{melody}.
Output: A chord sequence y_{chord}.

1. $x_{\text{chord}} \leftarrow \mathbf{0}$ and $x_{\text{mask}} \leftarrow \mathbf{1}$
2. $x \leftarrow \text{concat}(x_{\text{melody}}, x_{\text{chord}}, x_{\text{mask}})$
3. $y_{\text{chord}} \leftarrow \text{model}(x)$
4. for $i = 0, 1 \dots n$,
   (a) $\alpha \leftarrow P_{\min} + \dfrac{(P_{\max} - P_{\min}) \times i}{n}$
   (b) $x_{\text{mask}} \leftarrow \text{create_random_mask}(\alpha)$
   (c) $y_{\text{chord}} \leftarrow y_{\text{chord}} \circ x_{\text{mask}}$
   (d) $x \leftarrow \text{concat}(x_{\text{melody}}, y_{\text{chord}}, x_{\text{mask}})$
   (e) $y_{\text{chord_new}} \leftarrow \text{model}(x)$
   (f) $y_{\text{chord}} \leftarrow \text{where}(x_{\text{mask}} \text{ equals } 1, y_{\text{chord_new}}, y_{\text{chord}})$

probability can be estimated elaborately. Algorithm 1 elaborates our sampling process in the inference stage. The number of iterations n is set to the average length of chord sequences, $P_{\max}$ is set to 1, and $P_{\min}$ is set to 0.05.

3. EXPERIMENTS AND EVALUATIONS

3.1. Dataset

The Hooktheory Lead Sheet Dataset (HLSD) [23] is used in this study. The dataset is collected through a community-based platform for users to upload lead sheets. It contains high-quality, human-arranged melodies with chord progressions and other information. After keeping song tracks containing melody/chord pairs, we constructed a dataset of 18,005 samples for this study. We divided the dataset into a training set containing 17,505 samples and a test set containing 500 samples. The dataset contains chords that are beyond our chord space, such as 9th, 11th, 13th, half-diminished, and slashed chords. We transferred these chords into one of the 96 chords according to their chord functions. The slashed chords were returned to their root positions. All the pieces were transferred into either C major or c minor. In addition, by sampling chords at the beginning and middle of a bar, we made the chords in the dataset last at least half a bar.

3.2. Metrics

For objective evaluation, we use six different objective metrics introduced in [12].

- **Chord histogram entropy (CHE):** The entropy of the histogram of chord space.
- **Chord coverage (CC):** The number of chords that are used in a chord sequence.
- **Chord tonal distance (CTD):** The tonal distance proposed in [24] for measuring the distance between two chords.
- **Chord tone to non-chord tone ratio (CTnCTR):** The ratio of the number of chord tones to non-chord tones.
- **Pitch consonance score (PCS):** The consonance score that is computed through every note in a given chord.
- **Melody-chord tonal distance (MCTD):** The tonal distance between a note and a chord in 6-D feature vectors.

The first three metrics measure the quality of the chord progression, and the others measure the coherence between the melody and chords. For details about the metrics, see [12].

For subjective evaluation, we randomly selected 50 melodies from our test set. Each human subject needs to evaluate the chord progressions of five melodies. For each melody, the subjects listen to the melody without chords first, and then three different harmonizations composed by humans, MTHarmonizer, and our model in random order.

We design four criteria to obtain feedback from subjects:

- **Coherence:** how well the chord progression matches with the melody in terms of harmonicity and phrasing.
- **Chord progression:** how reasonable or smooth the chord progression is on its own, independent from the melody.
- **Interestingness:** how surprising or innovative the chord progression is on its own, independent from the melody.
- **Overall:** how well the whole harmonization is in comprehensive consideration.

A total of 102 subjects participated in the test. The ratings of 9 subjects who always gave the same score regardless of different harmonization were discarded. After the cleaning step, 93 subjects remained. 57 of them know the definition of chord progression. In addition, if the overall score contradicted the ranking result, the ratings for a melody were discarded.

3.3. Results

3.3.1. Objective tests

The results are shown in Table 1. Before we get into the discussion, it should be noted that none of the objective metrics mentioned in Section 3.2 can directly reflect the quality of the composed chords. It is difficult to quantify which metrics are better for which purposes and how accurate these metrics are. The human ear is still the best criterion for all music pieces.

In terms of the melody/chord harmonicity metrics (i.e., CTnCTR, PCS, and MCTD), we can see that the human-composed pieces are worse than the automatically composed pieces. We think that this happens because human composers use many 7th, 9th, and 11th chords to produce more interesting chord progressions. However, these modern chords may degrade the performance in the melody/chord harmonicity metrics. As a result, we should not conclude that in terms of melody/chord harmonicity, the performance of human composers is worse than that of melody harmonization systems.

Comparing our model (cf. BGS w/o B, $|S| = 48$) with MTharmonizer, we find that the introduction of the orderless NADE training process and blocked Gibbs sampling can indeed improve the performance in all three melody/chord

Table 1. Objective evaluation scores of different models. BGS denotes blocked Gibbs sampling and B denotes chord balancing. Higher values in CTnCTR and PCS and lower values in MCTD indicate better melody/chord harmonicity. Higher values in CHE and CC and lower values in CTD implicitly mean better interestingness of chord progression.

M/C Harmonicity	CTnCTR↑	PCS↑	MCTD↓
Humans	0.726	0.515	1.276
MTHarmonizer, $\|S\| = 48$	0.804	0.576	1.152
BGS w/o B, $\|S\| = 48$	0.850	0.652	1.087
BGS w/o B, $\|S\| = 96$	0.853	**0.657**	1.053
BGS w/ B, $\|S\| = 48$	0.876	0.653	1.068
BGS w/ B, $\|S\| = 96$	**0.887**	0.652	**1.052**
Chord Progression	CHE↑	CC↑	CTD↓
Humans	1.266	4.344	0.628
MTHarmonizer, $\|S\| = 48$	0.859	3.104	**0.512**
BGS w/o B, $\|S\| = 48$	0.755	2.926	0.521
BGS w/o B, $\|S\| = 96$	0.726	2.862	0.540
BGS w/ B, $\|S\| = 48$	1.209	4.698	0.690
BGS w/ B, $\|S\| = 96$	**1.280**	**4.900**	0.730

Table 2. Votes and rates of the best with respect to various methods. In each of the 5 melody/chord test sets, the subject was asked to choose the best from three chord progressions.

Method	**Votes**	**Rate of the Best %**
Humans	87	31.9
MTHarmonizer	66	24.2
Proposed	**86**	**31.5**
No preference	34	12.5

harmonicity metrics. This result shows that our model can compose the chords that are highly coherent with the given melody. In addition, introducing chord balancing (i.e., class weighting) and augmenting the chord space to 96 will not only not damage the harmonicity score, but will actually increase the performance (cf. BGS w/ B, $|S| = 96$). We think that this is because the more chords that can be selected and the more balanced the distribution of chord appearance, our model can better demonstrate its ability.

For the metrics of chord progression, we can find that humans compose very interesting chord progressions. Neither MTHarmonizer nor our model (cf. BGS w/o B, $|S| = 48$ or 96) can achieve such diversity. However, with the introduction of chord balancing, our model (cf. BGS w/ B, $|S| = 48$ or 96) can compose highly diverse chord progressions. It is worth noting that our model did not get a good CTD score. However, the CTD score can only represent the distance between two chords. It will be 0 if the chord does not change throughout the chord progression. In fact, the CTD scores of our model are closer to that of human-composed chords.

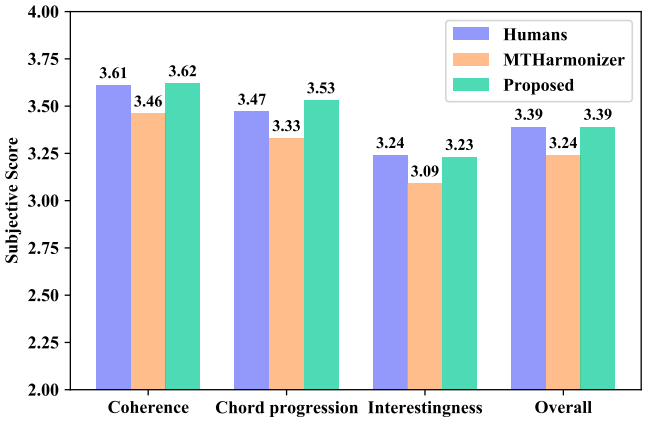


Fig. 2. Histogram of subjective evaluation results. Our model outperforms MTHarmonizer in all aspects and even gains higher scores than human composers in chord progression.

3.3.2. Subjective tests

Fig. 2 shows the results of the subjective test. Our model (BGS w/ B, $|S| = 96$) surpasses MTHarmonizer in all criteria and has close results to human composers. Our model has excellent performance in terms of coherence, which is consistent with the high melody/chord harmonicity in the objective evaluation. The high score also indicates that with blocked Gibbs sampling, our model can produce more reasonable and smoother chord progressions.

The subjects were also asked to choose the best from three chord progressions of a melody. Table 2 shows that the music pieces composed by humans are the most favorable, but the percentage of being selected as the best is only slightly higher than that of our model. We conclude that our model can arrange music pieces very well.

Several examples are available at `https://chord-generation.herokuapp.com/demo`. They are randomly rendered from all 500 test samples in HLSD.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a framework based on the BiLSTM structure and the orderless NADE training process for automatic melody harmonization. We applied blocked Gibbs sampling in inference. In addition, we applied chord balancing and augmented the chord space to 96 to increase the diversity of chord progressions. The objective and subjective test results demonstrate that our model is better than Yeh's model and is comparable in performance to human composers.

In our future work, we will try to make our neural model learn Roman numeral analysis, which is how humans learn how to compose chords, and one of the reasons why our proposed model cannot surpass human composers.

5. REFERENCES

[1] C.-H. Chuan and E. Chew, "A hybrid system for automatic generation of style-specific accompaniment," in *Proc. IJWCC*, 2007.

[2] I. Simon, D. Morris, and S. Basu, "MySong: Automatic accompaniment generation for vocal melodies," in *Proc. CHI*, 2008.

[3] D. Makris, I. Kayrdis, and S. Sioutas, "Automatic melodic harmonization: An overview, challenges and future directions," in *Trends in Music Information Seeking, Behavior, and Retrieval for Creativity*. IGI Global, 2016, pp. 146–165.

[4] J. F. Paiement, D. Eck, and S. Bengio, "Probabilistic melodic harmonization," in *Proc. Canadian AI*, 2006.

[5] H. Tsushima, E. Nakamura, K. Itoyama, and K. Yoshii, "Function- and rhythm-aware melody harmonization based on tree-structured parsing and split-merge sampling of chord sequences," in *Proc. ISMIR*, 2017.

[6] ——, "Generative statistical models with self-emergent grammar of chord sequences," *J. New Music Res.*, vol. 47, no. 3, pp. 226–248, 2018.

[7] T. Kitahara, S. Giraldo, and R. Ramírez, "JamSketch: Improvisation support system with GA-based melody creation from user's drawing," in *Proc. CMMR*, 2018.

[8] J.-P. Briot, G. Hadjeres, and F.-D. Pachet, "Deep learning techniques for music generation: A Survey," 2017. [Online]. Available: http://arxiv.org/abs/1709.01620

[9] H. Lim, S. Rhyu, and K. Lee, "Chord generation from symbolic melody using BLSTM networks," in *Proc. ISMIR*, 2017.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *J. Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] T.-P. Chen and L. Su, "Functional harmony recognition of symbolic music data with multi-task recurrent neural networks," in *Proc. ISMIR*, 2018.

[12] Y.-C. Yeh, W.-Y. Hsiao, S. Fukayama, T. Kitahara, B. Genchel, H.-M. Liu, H.-W. Dong, Y. Chen, T. Leong, and Y.-H. Yang, "Automatic melody harmonization with triad chords: A comparative study," 2020. [Online]. Available: https://arxiv.org/abs/2001.02360

[13] B. Uria, I. Murray, and H. Larochelle, "A deep and tractable density estimator," in *Proc. ICML*, 2014.

[14] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, "Neural autoregressive distribution estimation," *J. Mach. Learn. Res.*, vol. 17, pp. 1–37, 2016.

[15] L. Yao, S. Ozair, K. Cho, and Y. Bengio, "On the equivalence between deep NADE and generative stochastic networks," in *Proc. ECML*, 2014.

[16] C.-Z. A. Huang, T. Cooijmans, A. Roberts, A. Courville, and D. Eck, "Counterpoint by convolution," in *Proc. ISMIR*, 2017.

[17] C.-Z. A. Huang, C. Hawthorne, A. Roberts, M. Dinculescu, J. Wexler, L. Hong, and J. Howcroft, "The Bach doodle: Approachable music composition with machine learning at scale," in *Proc. ISMIR*, 2019.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[19] S. Ruder, "An overview of gradient descent optimization algorithms," 2016. [Online]. Available: http://arxiv.org/abs/1609.04747

[20] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. ICCV*, 2016.

[22] Y. Cui, M. Jia, T. Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proc. CVPR*, 2019.

[23] C. Anderson, D. Carlton, R. Miyakawa, and D. Schwachhofer, "Hooktheory." [Online]. Available: https://www.hooktheory.com

[24] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proc. AMCMM*, 2006.