

EMPIRICAL RESEARCH

Open Access



Generating chord progression from melody with flexible harmonic rhythm and controllable harmonic density

Shangda Wu^{1†}, Yue Yang^{1†}, Zhaowen Wang^{1†}, Xiaobing Li¹ and Maosong Sun^{1,2*}

Abstract

Melody harmonization, which involves generating a chord progression that complements a user-provided melody, continues to pose a significant challenge. A chord progression must not only be in harmony with the melody, but also interdependent on its rhythmic pattern. While previous neural network-based systems have been successful in producing chord progressions for given melodies, they have not adequately addressed controllable melody harmonization, nor have they focused on generating harmonic rhythms with flexibility in the rates or patterns of chord changes. This paper presents AutoHarmonizer, a novel system for harmonic density-controllable melody harmonization with such a flexible harmonic rhythm. AutoHarmonizer is equipped with an extensive vocabulary of 1462 chord types and can generate chord progressions that vary in harmonic density for a given melody. Experimental results indicate that the AutoHarmonizer-generated chord progressions exhibit a diverse range of harmonic rhythms and that the system's controllable harmonic density is effective.

Keywords Melody harmonization, Controllable music generation, Harmonic rhythm, Harmonic density

1 Introduction

In recent years, there has been considerable research effort devoted to developing the practical applications of neural networks in the field of music. Neural networks have enabled the implementation of automatic transcription [1, 2], which involves converting audio signals into symbolic musical representations. Researchers also used neural networks to classify musical pieces by genre [3, 4] and even generate original music [5–7]. This paper focuses on the task of melody harmonization, which involves creating a neural network-based system

that generates chord progressions to accompany a given melody, with the added ability to control the harmonic rhythm.

In music, a chord is a combination of multiple notes that produce a harmonious sound. The transition of chords within a musical composition is known as harmonic rhythm or harmonic tempo. Melody harmonization systems, as studied in [8–10], are designed to automatically generate suitable chords that accompany a given melody, essentially harmonizing the melody.

The process of harmonizing a melody involves selecting the appropriate chords that complement the melody's underlying tonality, structure, and rhythm. The harmonization system must analyze the melody's pitch, duration, and rhythmic patterns to identify the most appropriate chords to use at each point in the melody. The system's output should enhance the melody's expressiveness, while also maintaining a sense of coherence and musical logic.

[†]Shangda Wu, Yue Yang and Zhaowen Wang contributed equally to this work.

*Correspondence:

Maosong Sun
sms@tsinghua.edu.cn

¹ Department of Music AI and Information Technology, Central Conservatory of Music, Beijing, China

² Department of Computer Science and Technology, Tsinghua University, Beijing, China

Melody harmonization systems have a range of potential applications, including music composition, arranging, and production. These systems can be particularly useful for musicians who lack formal training in music theory or for those seeking inspiration for new musical compositions. Furthermore, these systems can facilitate the creation of harmonically complex and innovative musical arrangements by automating the process of writing chord progression.

Melody harmonization is a complex and subjective task that has received considerable attention from researchers in recent years. However, existing works in this area [11–14] have mainly focused on generating appropriate chords, while neglecting the equally important aspect of placing them within the proper musical context. Consequently, the aforementioned works suffer from limited flexibility in chord progression generation, as they tend to produce only a single chord for each bar or half-bar, leading to rigid harmonic rhythms that do not capture the subtleties of musical expression.

To address these challenges, it is necessary to develop more sophisticated models that can capture the full complexity of melody harmonization and generate musically satisfying and aesthetically pleasing results. Such models should take into account the wider musical context, including the relationships between different chords and their role in the overall harmonic structure. Furthermore, they should allow for greater flexibility in chord progression generation, allowing for variations in rhythmic and harmonic patterns that reflect the nuances of musical expression.

This study aims to develop a novel approach to achieve automatic melody harmonization with flexible

harmonic rhythm, where chord progressions rhythmically match a given melody. To achieve this objective, the proposed approach, AutoHarmonizer, as shown in Fig. 1, generates chords on a sixteenth note basis (referred to as a “frame” in this context) instead of bar-by-bar. This modeling strategy better represents the task at hand and allows for more accurate harmonization. Additionally, time signatures are encoded to establish rhythmic relationships between melodies and chords. The controllable harmonic density, which refers to the degree of richness or sparsity in the generated chord progressions, based on the work of [15], has been implemented to enable customized harmonizations according to user preferences.

Contributions of this paper are summarized as follows:

- The proposed model considers beat and key information, enabling it to handle any number of time signatures and key signatures in a piece, without being limited to specific notations such as C major and 4/4.
- The AutoHarmonizer predicts chords frame-by-frame, which enables the generation of flexible harmonic rhythms.
- The utilization of gamma sampling allows users to adjust the harmonic density of model-generated chord progressions.

2 Related work

2.1 Melody harmonization

Melody harmonization is a branch of algorithmic composition [8], which aims to generate a chord progression automatically for a given melody [12, 13, 16]. Some

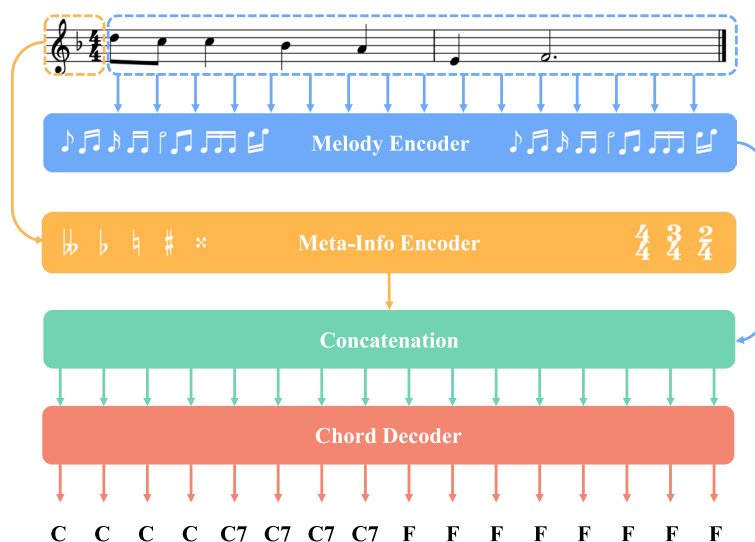


Fig. 1 The architecture of AutoHarmonizer, which predicts chord symbols frame-by-frame (sixteenth note)

of these studies have also focused on generating a four-part chorale to accompany a given melody [17–19]. The present paper specifically addresses the former approach.

Tsushima et al. [10] proposed a method for chord hierarchy representation based on a Probabilistic Context-Free Grammar (PCFG) of chords. They developed a metrical Markov model for controllable chord generation using this hierarchical representation. However, this approach relies on statistical learning, which tends to generate simpler and more basic chord sequences, resulting in fewer generated chords than bars.

Lim et al. [16] designed a model based on a Bi-directional Long and Short-Term Memory (Bi-LSTM) network that can generate a chord from 24 triads for each bar. However, this model has limitations such as disregarding note order, rhythm, and octave information within bars. It generates results with overuse of common chords and inappropriate cadences.

To address these limitations, Yeh et al. [13] extended Lim's model, called MTHarmonizer, to predict a chord from 48 triads for each half-bar. They also included some extra information such as tonic and dominant to improve the model's performance.

In [12], Sun et al. applied orderless sampling and class weighting to the Bi-LSTM model. They expanded the types of chords to 96, and subjective experiments demonstrated that the generated results were comparable to those produced by human composers.

Chen et al. [14] proposed SurpriseNet, a model based on a Conditional Variational Auto-Encoder (CVAE) and Bi-LSTM. This model enables user-controllable melody harmonization.

Yang et al. [20] utilized two LSTM models. One model focused on the relationship between notes in the melody and their corresponding chords, while the other model focused on the rules of chord transfer.

Majidi et al. [21] combined genetic algorithms with LSTMs to generate and optimize melodies and chords.

A recent deep learning approach [22] by Rhyu et al. leverages a Transformer architecture, combined with a VAE framework, to generate structured chord sequences from melodies.

It should be noted that all of the models mentioned above, except for Tsushima et al. [10], cannot generate flexible harmonic rhythms.

2.2 Controllable music generation

Controllable music generation systems refer to computer programs or algorithms that can generate music based on specific requirements set by the user. These systems rely on the representation of various

properties of music, which may be subjective, such as emotion and style, or objective, such as tonality and beat. The generation of music by these systems can be customized to meet the specific needs of the user and can be tailored to a particular application.

Controllable music generation has been an active research area, and several models have been proposed to achieve this goal. Roberts et al. proposed a model based on recurrent Variational Auto-Encoders (VAEs) [23]. This model enables controllable generation through hierarchical decoders, allowing for control over various musical features such as harmony, melody, and rhythm.

Luo et al. proposed a model based on VAEs with Gaussian mixture latent distributions [24]. This model enables the learning of decoupled representations of timbres and pitches, facilitating the control of these two musical features separately. Zhang et al. proposed BUTTER [25], a representation learning model based on VAE, which can learn latent representations and cross-modal representations of music. This model allows for searching or generating corresponding music by inputting text, providing users with more convenient control over the generation process.

Chen et al. proposed Music SketchNet [26], which uses VAE to decouple rhythmic and pitch contours, allowing for guided generation based on user-specified rhythms and pitches. Wang et al. proposed PianoTree VAE [27], which uses a Gated Recurrent Unit (GRU) to encode notes played simultaneously and map them to a latent space. This model achieves controllable generation of polyphonic music based on a tree structure.

Finally, Di et al. proposed the Controllable Music Transformer [28], which achieves rhythmic consistency between video and background music. This model allows for the local control of rhythm while globally controlling the music genre and instruments.

It should be noted that these aforementioned models are designed for controllable music generation, and integrating their techniques into models without this capability might be challenging.

3 Methodology

3.1 Data representation

Our research has yielded a novel data representation that includes crucial meta-information obtained from sheet music, namely the time signature and key signature. The data representation is illustrated in Fig. 2, and it involves encoding each lead sheet into four sequences, with equal lengths. This approach enables us to accurately capture the temporal and harmonic structure of the music. By including both time signature and key signature in the data representation, we can capture the rhythmic and harmonic patterns of the music comprehensively.

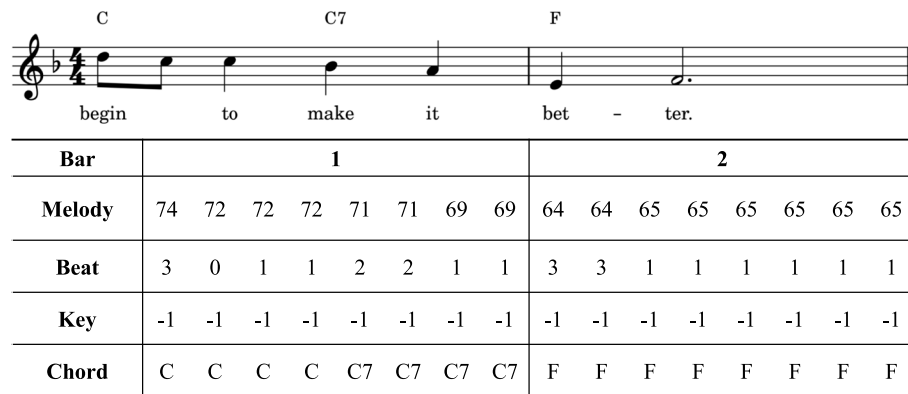


Fig. 2 A two-bar sample of a melody, beat, key, and chord representation (at a time resolution of eighth notes)

Consequently, this enhances the accuracy and completeness of our music representation.

- *Melody Sequence*: we adopt a 128-dimensional one-hot vector encoding scheme to represent musical frames. Specifically, each frame is represented as a one-hot vector with 128 dimensions, where the time resolution is set at the level of sixteenth notes. The first dimension of the vector is reserved for representing rests, while the remaining 127 dimensions correspond to the unique pitches in the MIDI standard (excluding pitch 0).
- *Beat Sequence*: a sequence of 4-dimensional vectors based on time signatures. It represents the beat strength of each frame in the melody sequence. Its values range from 0 to 3, corresponding to non-beat, weak, medium-weight, and strong beats. This sequence provides important information on the rhythmic structure of the melody.
- *Key Sequence*: the encoding of keys is based on the number of sharps or flats associated with each key. Specifically, flats are assigned a numerical value ranging from -7 to -1 , while sharps are assigned a numerical value ranging from 1 to 7. Keys with no sharps or flats are assigned a value of 0. In total, there are 15 possible types of key encoding based on this system.
- *Chord Sequence*: there are 1461 unique chord symbols were identified in our dataset. The first dimension is reserved for rests, leading to a one-hot vector representation of 1462 dimensions for each chord.

3.2 Network architecture

In musical composition, a crucial aspect is the consideration of individual notes that comprise a given melody segment to effectively match it with a suitable chord

progression. Generally, chords that incorporate notes already present in the melody, namely chord tones, are preferred. Nevertheless, there may be situations where several chords align with the current set of notes, necessitating the selection of the subsequent chord based on the upcoming notes of the melody. Thus, the selection of chords in musical composition involves a balance between the notes that are currently being played and those that are yet to come. This process is essential in creating a harmonious and coherent musical composition.

AutoHarmonizer is a model developed to capture music information bidirectionally. It employs a Bi-LSTM backbone network and an encoder-decoder architecture, as shown in Fig. 1. The model consists of two encoders, the melody encoder, and the meta-info encoder. The melody encoder takes a melody sequence as input, while the meta-info encoder takes a concatenated sequence of beat and key sequences. Both encoders have two stacked blocks, each comprising a Bi-LSTM layer with 256 units and a time-distributed layer with 128 units. The last hidden states of the encoders are concatenated, and the resulting vector is used as input to the decoder. The decoder is made up of three stacked layers, and its output layer has 1462 units, which represent chord types. The chord symbols are generated autoregressively, frame-by-frame (sixteenth note) in the decoder. During training, the model used a dropout rate of 0.2, a batch size of 512, and early stopping with a patience of 20 epochs, as determined by empirical evaluation.

3.3 Controllable harmonic density

In [15], Wu et al. proposed the use of gamma sampling to control the language models based on the assumption that certain attributes of the generated text have a close correlation with the number of occurrences of specific tokens. Gamma sampling provides a means to generate controllable text by scaling the probability of

the token associated with the attribute during the generation process:

$$\begin{aligned} p_{\mathcal{A}_{out}} &= p_{\mathcal{A}_{in}}^{\tan\left(\frac{\pi\Gamma}{2}\right)}, \\ p_{a_{out}} &= p_{a_{in}} \cdot \frac{p_{\mathcal{A}_{out}}}{p_{\mathcal{A}_{in}}}, \quad \forall a \in \mathcal{A}, \\ p_{n_{out}} &= p_{n_{in}} \cdot \left(1 + \frac{p_{\mathcal{A}_{in}} - p_{\mathcal{A}_{out}}}{p_{\mathcal{A}_{in}}}\right), \quad \forall n \notin \mathcal{A}, \end{aligned} \quad (1)$$

where $\Gamma \in [0, 1]$ is the user-controllable control strength, \mathcal{A} is the set of attribute-related tokens ($\setminus\mathcal{A}$ is its complement), $p_{a_{in/out}}$ is the input/output probability of an attribute-related token a , and the same goes for every non-attribute-related token n . When $\Gamma = 0.5$, there is no change in the probability distribution, while when $\Gamma < 0.5$, the probabilities of the attribute-related tokens increase and vice versa.

AutoHarmonizer uses a strategy to achieve controllable harmonic density whereby the previously generated chord token c_{t-1} is selected as the attribute-related token for generating the chord c_t at time step t . Specifically, when the value of a parameter called Γ exceeds 0.5, the model is more inclined to generate chords that differ from c_{t-1} , resulting in a greater frequency of chord switching and an increase in harmonic density. Conversely, when Γ is less than 0.5, the likelihood of generating denser chord progressions is reduced. This approach enables AutoHarmonizer to produce musical pieces with controllable harmonic complexity, allowing for flexibility in the generation of diverse music styles.

Musical composition typically involves the frequent usage of essential chords, notably the tonic, dominant, and subdominant chords. While it is true that a model can loop between these chords, our model, with the introduction of high values of Γ , tends to diversify its chord choices, resulting in a more even distribution of chord types. As demonstrated in Fig. 4, a higher Γ leads to a more balanced distribution of scale degrees in the generated chord progressions, indicating the inclusion of less common chords. Thus, by adjusting Γ , composers can influence the harmonic density and the distribution between essential and non-essential chords, offering them more control over the musical texture and composition.

4 Experiments

4.1 Setups

4.1.1 Dataset

In our study, we utilized Wikifonia.org's lead sheet dataset, consisting of 6675 compositions predominantly from Western genres such as rock, pop, country, jazz, folk, R&B, and children's songs. In order to improve the

quality of the dataset, we removed lead sheets that lacked chord symbols or did not switch chords within 4 bars, ultimately resulting in a subset of 5204 lead sheets. Subsequently, we divided the subset into a training set comprising 90% of the data and a validation set containing the remaining 10%.

4.1.2 Baselines

In our research, we chose two previous melody harmonization systems as our baselines. The first is a traditional melody harmonization system proposed by Lim et al. [16], known as Chord Generation from Symbolic Melody (CGSM). This system is based on a Bi-LSTM architecture and was trained and validated using the Wikifonia dataset. The other two baseline models are STHarm and VTHarm [22]. Both of these models were trained on the Chord Melody Dataset (CMD)¹. STHarm directly translates melodies into chords by mapping individual melody notes to chord progressions. On the other hand, VTHarm, featuring a key-aware variational Transformer architecture, not only generates chords from melodies but also captures the broader musical context and structure.

It is important to note that the strategy adopted by STHarm and VTHarm, which generates two chords per bar, has its limitations; they are only applicable to pieces in 4/4 time. Specifically, out of the 515 valid pieces in the validation set, only 311 are in 4/4 time. This implies that their comparison with other models might not necessarily be apple-to-apple.

We evaluated our system in various settings to determine the effectiveness of controllable harmonic density in melody harmonization. The proposed system, referred to as AH- Γ , consists of AutoHarmonizer set at different Γ values ranging from 0.5 to 0.9. Through our analysis, we sought to establish a deeper understanding of the relationship between harmonic density and melody harmonization.

4.2 Metrics

Our study evaluated the performance of AutoHarmonizer using a variety of metrics. These metrics included Accuracy (ACC), which measured the proportion of matching frames between the generated and true chord progressions. We also utilized six metrics proposed in a previous study [13] that have become widely used in the literature [12, 22] for assessing chord progression and melody/chord harmonicity. By utilizing these metrics, we were able to thoroughly evaluate the performance of AutoHarmonizer.

¹ <https://github.com/shiehn/chord-melody-dataset>.

- *Chord Coverage (CC)*: the number of chord types in a piece of music. This value serves as an indicator of the richness and variety of chord progressions in the music, with higher CC values indicating a greater number of distinct chord types being used.
- *Chord Histogram Entropy (CHE)*: creates a histogram of chord occurrences based on a chord sequence:

$$CHE = - \sum_{k=1}^{CC} p_k \cdot \log p_k, \quad (2)$$

where p_k is the frequency of the k th chord occurrence. The higher the value of CHE, the greater the uncertainty and the variety of chords.

- *Chord Tonal Distance (CTD)*: the average value of the tonal distance [29] computed between every pair of adjacent chords in a given chord sequence. It involves three steps: (1) the Pitch Class Profile (PCP) features of both chords are computed; (2) these features are then projected onto a six-dimensional tonal space; (3) the Euclidean distance between the two six-dimensional feature vectors is calculated, resulting in a tonal distance value. The lower the value of CTD, the smoother the chord progression.
- *Chord Tone to non-Chord Tone Ratio (CTnCTR)*: calculates the ratio of the number of the chord tones (n_c) and proper non-chord tones (n_p), to the number of the non-chord tones (n_n):

$$CTnCTR = \frac{n_c + n_p}{n_c + n_n}, \quad (3)$$

The concept of chord tones refers to melody notes whose pitch class belongs to the current chord, specifically, one of the three pitch classes that constitute a triad for the corresponding half bar. Melody notes that do not fall into this category are considered non-chord tones. Among the non-chord tones, a subset of notes that are two semitones away from the notes immediately following them is referred to as proper non-chord tones. CTnCTR equals one when there are no non-chord tones at all, or all non-chord tones are proper.

- *Pitch Consonance Score (PCS)*: based on the musical interval between the pitch of the melody note and the chord notes, assuming that the pitch of the melody notes is always higher, which is the case in our system. Specifically, it assigns a score of 1 to consonant intervals, including unison, major/minor 3rd, perfect 5th, and major/minor 6th. A perfect 4th receives a score of 0, while other inter-

vals are considered dissonant and receive a score of -1 . To compute PCS for a pair of melody and chord sequences, these consonance scores are averaged across 16th-note windows, excluding rest periods.

- *Melody-Chord Tonal Distance (MCTD)*: represents a melody note using a PCP feature vector, which is essentially a one-hot vector. Next, it compares the PCP of this vector against the PCP of a chord label in a 6-D tonal space [29]. The resulting measure provides an estimate of the closeness between the melody note and the chord label. To obtain a comprehensive measure of the tonal distance between a melody sequence and its corresponding chord labels, it calculates the average of the tonal distance between every melody note and the corresponding chord label across the melody sequence. It also weights each distance by the duration of the corresponding melody note.

In light of the fact that the previously mentioned metrics do not consider the measurement of harmonic rhythm, we developed an additional set of three metrics, as outlined below.

- *Harmonic Rhythm Coverage (HRC)*: similar to CC, but it is computed specifically for harmonic rhythm types. A higher HRC value indicates the use of a greater number of unique harmonic rhythm types.
- *Harmonic Rhythm Histogram Entropy (HR-HE)*: same as CHE, but calculates the histogram of harmonic rhythm:

$$HRHE = - \sum_{u=1}^{HRC} p_u \cdot \log p_u, \quad (4)$$

where p_u is the frequency of the u th harmonic rhythm. The HRHE value reflects the uncertainty and variety of harmonic rhythms in the music. A higher HRHE value indicates greater variation and uncertainty in the harmonic rhythms.

- *Chord Beat Strength (CBS)*: chord placements can be assessed by their average beat strength, which is scored on a scale ranging from 0 (non-beat) to 3 (strong beat), based on the beat sequence. The beat sequence determines the recurring pulse of a musical composition and is used to evaluate the placement of chords in relation to the underlying beat. A smaller value of the CBS metric indicates that more chords are positioned on non-strong beats, while a higher CBS value indicates the opposite. Therefore, CBS provides a quantitative measure

of the degree to which chords are aligned with the underlying beat.

4.3 Quantitative evaluations

Table 1 presents the findings of this study, which can serve as a foundation for further analysis. It is crucial to note, however, that the metrics employed in this study are not meant to provide an all-encompassing evaluation of chord progression quality, as such assessments are inherently intricate and subjective [13]. Therefore, although the results presented in Table 1 can be used for comparison purposes, they should not be viewed as definitive indicators of chord progression quality. It is imperative to acknowledge these limitations when interpreting the findings of this study.

The results of ACC demonstrate that AutoHarmonizer consistently surpassed CGSM, VTHarm, and STHarm across all values of Γ . One might initially conclude that without incorporating less frequent chords, an over-reliance on prevalent triads could fail to effectively encapsulate the subtleties of human-composed musical compositions. However, it's crucial to note that the comparatively lower ACC of VTHarm and STHarm may stem from disparities in their training data compared to the other models. Furthermore, we found that an increase in Γ was associated with a decline in accuracy. This observation indicates that the increase in the frequency of chord transitions leads to a greater deviation from actual chord progressions. While our metrics are single-faceted and may not capture the entirety of what makes a composition human-like, they do emphasize the potential benefit of including a variety of chord progressions in generation.

The chord progression metrics, CC, CHE, and CTD, serve as essential tools for evaluating both ground truth and model-generated chord progressions. Ground truth show low CTD, high CC, and CHE, indicating inherent smoothness and diversity. Among model-generated

progressions, STHarm and VTHarm closely aligns with ground truth for CC and CHE. However, its considerably lower CTD is a result of frequently allocating identical chords within bars, leading to an “pseudo-smoothness.” Unlike VTHarm, the data representation of AH prevents consecutive identical chords, making AH-0.9's CTD closely resemble the ground truth. Meanwhile, CGSM offers the smoothest progressions with the lowest CTD but sacrifices diversity, evidenced by its low CC and CHE. This highlights the trade-off between progression smoothness and diversity.

The results of melody/chord harmonicities indicate that the actual musical compositions exhibit a greater prevalence of non-chord tones (the lowest CTnCTR), resulting in intervals that are more dissonant (the lowest PCS and highest MCTD) compared to those produced by the model. Notably, as the value of Γ increased, the AutoHarmonizer utilized more non-chord tones while also forming more consonant intervals with the melody notes, resulting in decreased MCTD and increased PCS, suggesting a tendency towards consonance. In addition, both STHarm and VTHarm showed low CTnCTR and PCS values, with high MCTD, indicating a more pronounced deviation from the ground truth. These findings indicate that there may be significant differences in harmonic structures between human-composed and model-generated musical compositions (see Fig. 6 for examples of melody harmonization by various models).

The harmonic rhythm metrics demonstrate a clear contrast between chord progressions generated by the CGSM and those produced by AutoHarmonizer and ground truth. Specifically, chord progressions generated by the CGSM, STHarm, and VTHarm exhibit a fixed harmonic rhythm, which is a shared limitation among all melody harmonization neural network systems [12, 13]. The results suggest that an increase in the parameter Γ leads to greater rhythmic diversity of chords, although the distribution is more concentrated, resulting in

Table 1 Quantitative evaluations on the validation set (515 tunes). The values closest to the ground truth are bolded

Methods	Accuracy	Chord progression			Melody/chord harmonicities			Harmonic rhythm		
		CC	CHE	CTD	CTnCTR	PCS	MCTD	HRC	HRHE	CBS
Ground truth	1.0000	12.0603	2.0582	1.0653	0.8476	0.4975	1.1525	5.7626	1.1877	2.6024
AH-0.5	0.3087	5.0914	1.3477	1.1864	0.9054	0.5590	1.1252	5.9027	1.3965	2.8904
AH-0.6	0.3142	5.7101	1.4353	1.1769	0.9030	0.5723	1.1076	5.9397	1.3494	2.8668
AH-0.7	0.3111	6.4163	1.5227	1.1584	0.8979	0.5775	1.1011	6.0233	1.2901	2.8310
AH-0.8	0.3088	7.3152	1.6267	1.1289	0.8916	0.5773	1.1016	6.1031	1.2451	2.7758
AH-0.9	0.2890	8.6420	1.7555	1.0911	0.8797	0.5712	1.1038	6.1770	1.1910	2.6572
CGSM	0.2822	5.9883	1.3619	0.7814	0.9077	0.6224	1.0901	1.2354	0.0459	2.9892
STHarm	0.0885	9.8051	1.5723	0.6331	0.6184	0.2604	1.3285	1.0000	0.0000	2.5000
VTHarm	0.0928	13.5588	2.2083	0.8367	0.6063	0.2686	1.3364	1.0000	0.0000	2.5000

higher values of HRC and lower HRHE. Furthermore, an increase in Γ leads to the placement of more chords on non-strong beats, as demonstrated by the lower values of CBS. These findings indicate that the parameter Γ plays a critical role in generating varied and complex chord progressions with a non-fixed harmonic rhythm.

Figure 3 depicts the distribution of chord onsets on beat strengths. It is evident from the figure that most chord onsets occur on strong beats, while a lesser number of chord onsets occur on medium-weak beats, and a small number of chord onsets occur on weak beats. Non-beats are seldom used for chord placement. Notably, CGSM restricts chord progression variety by exclusively positioning chords on strong beats. In a similar vein, both STHarm and VTHarm distribute their chords evenly between strong and medium-weak beats, indicating a potential monotony in their transitions. In contrast,

by increasing the Γ parameter in AutoHarmonizer, there is a shift towards placing chord onsets on non-strong beats. Among these, the AH-0.9 model comes closest to mirroring the ground truth's chord onset distribution.

The distribution of scale degrees in chord progressions generated by AutoHarmonizer and baselines is depicted in Fig. 4. The results highlight a consistent pattern across all models, with the majority of chords being tonic, followed by subdominant and dominant chords. However, it is essential to note that some models, such as CGSM, tend to overuse tonic chords compared to the ground truth. In contrast, STHarm and VTHarm exhibit a distribution remarkably close to the ground truth. Notably, STHarm's utilization of scale degrees is even more evenly spread than that of the ground truth, indicating a diverse array of generated chord progressions. Interestingly, when we increase the parameter Γ , there is a noticeable

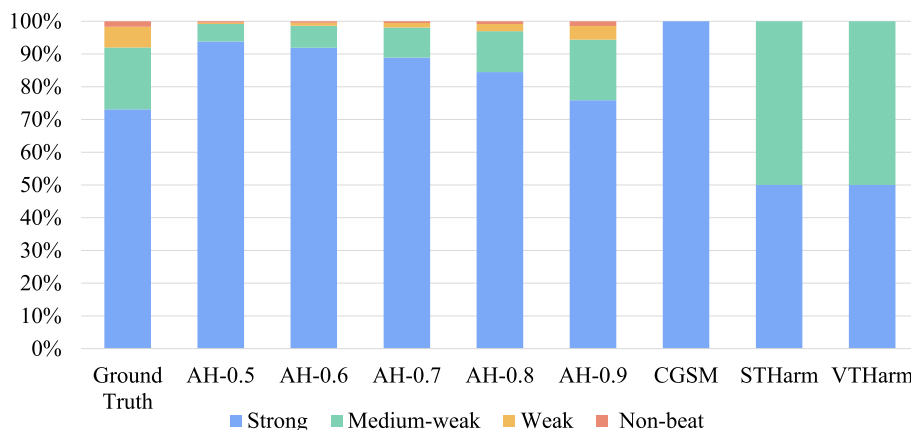


Fig. 3 Distribution of chord onsets on beat strengths, showing the proportion of chord onsets occurring on strong, medium-weak, and weak beats, and non-beats

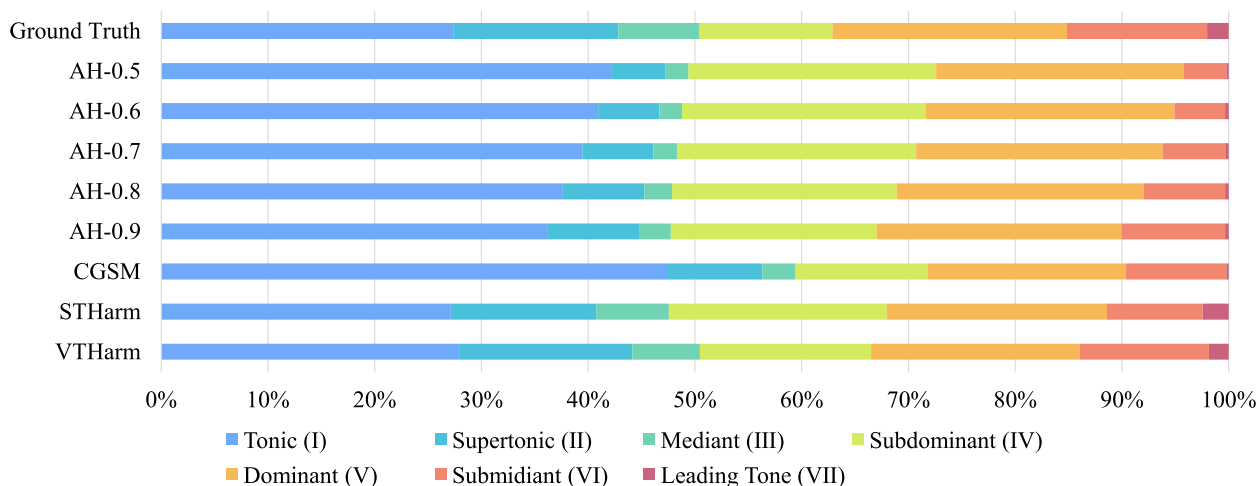


Fig. 4 Distribution of scale degrees in chord progressions generated by AutoHarmonizer and baselines, compared to chord progressions from ground truth

decrease in the usage of tonic chords, accompanied by a shift towards a more prominent presence of supertonic and submediant chords. This adjustment results in chord progressions that closely resemble the ground truth. By tuning the value of Γ , AutoHarmonizer can achieve a more balanced distribution across all chord degrees, thereby enhancing the diversity and richness of the generated chord progressions.

4.4 Discrimination test

This study recruited 83 participants with varying cultural and demographic backgrounds to engage in a discrimination task that aimed to differentiate between musical chords generated by humans and those created by machines.

The study categorized the subjects into three groups based on their level of music knowledge: (A) music faculty and students consisting of 33 subjects; (B) non-music majors with an understanding of harmony comprising 25 subjects; and (C) individuals with no knowledge of harmony but who listen to music often or occasionally, consisting of 25 subjects. The study selected a total of 20 tunes randomly from the validation set. Each tune had seven different versions, which included AH-0.5, AH-0.7, AH-0.9, CGSM, STHarm, VTHarm, and the ground truth.

It is important to note that the 20 tunes used for the listening test were randomly selected from the validation set. This choice was made because our dataset does not explicitly have a dedicated test set. We believe that this selection does not compromise the fairness of our experiments. Our primary focus is on comparing the performance of different models, rather than tuning the models.

Participants were presented with both melody-only versions and the ones with chords, and their task was to determine whether the chords were produced by humans or machines. The test consisted of questions that had an equal chance of containing chords generated by either humans (i.e., the ground truth) or machines. The primary objective of this study was to assess the capacity of the machine-generated chords to deceive human listeners into

believing they were created by humans, thereby determining the model's overall trustworthiness. The study design was also intended to provide insight into the ability of individuals from different backgrounds to distinguish between human and machine-generated musical content.

The results of the discrimination test, depicted in Fig. 5, demonstrate that AutoHarmonizer surpassed CGSM and STHarm in all evaluated configurations. Professional subjects possessing musical backgrounds A and B consistently identified the chord progressions generated by CGSM as machine-generated and nearly all those by STHarm in the same manner. This observation aligns with the limitations of CGSM, which are characterized by a restricted vocabulary size (24 types of chords only) and a fixed harmonic rhythm. The fixed harmonic rhythm often leads to repetitive chord progressions, which, in turn, can be perceived as less convincing by listeners. Similarly, STHarm utilizes a constant harmonic rhythm, echoing these shortcomings. VTHarm preserved this harmonic rhythm setup, but incorporated a key-aware context encoder, enhancing its output. Conversely, chord progressions produced by AH-0.5 were viewed as most akin to human-composed ones, based on subject feedback. Intriguingly, half the subjects mistook the authentic human compositions as machine-generated, highlighting a deep-rooted skepticism towards human-composed chord progressions. Furthermore, we noted that, in contrast to professionals (i.e., subjects in group A and B), individuals in group C were more inclined to classify progressions with fixed harmonic rhythms (i.e., CGSM, VTHarm, and STHarm) as human-composed. We speculate that they might deem progressions with a constant harmonic rhythm as more familiar or acceptable, especially since many popular music pieces employ recurrent and predictable chord patterns.

5 Melody harmonization examples

Figure 6 showcases several instances of melody harmonization, which aims to exhibit the effectiveness of diverse techniques utilized in harmonizing an identical 8-bar

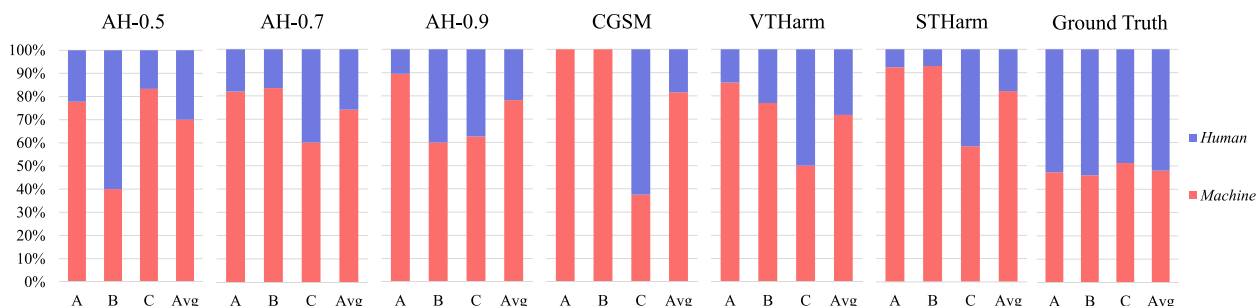


Fig. 5 Results of the discrimination test for each model and expertise level. Group A: Music faculty and students; Group B: Non-music majors with harmony knowledge; Group C: Individuals without harmony knowledge but frequent music listeners. The term “vote” refers to the participants’ tendency to categorize a given set of chords as either human-generated or machine-generated

melody that features a 4/4 time signature and is based in the key of F.

The progression that is manually created by humans shows a distinct periodicity, wherein each cycle is comprised of four chords, namely Dm-Bb-C-Dm7, with Dm functioning as the tonic chord. The progression conforms to a particular tonal structure, emphasizing the significant role played by the tonic chord in shaping the overall musical structure. The periodicity of the chord progression, with its repetition of chord progressions, provides a sense of predictability and familiarity that contributes to its aesthetic appeal. Furthermore, the use of the tonic chord creates a sense of resolution and stability, which is a common characteristic of tonal music. Thus, the chord progression crafted by humans showcases the creative potential of human composers in shaping musical structures through the careful selection and arrangement of chords.

The chord progression produced by AH-0.5 exhibited in this output is characterized by a significantly sparse harmonic rhythm, which is composed of only five chords in total. This chord progression follows a consistent pattern of F-C7-F in each cycle. The tonic chord is represented by F, indicating that the key of this chord progression is in F-major, which is the relative key of the ground truth. It is worth noting that for this melody, both

D-minor and F-major chord progressions are considered acceptable according to principles in music theory.

In music theory, chord progressions are often evaluated based on their harmonic function and tonal relationships within a key. While the ground truth chord progression is in D minor, which is the relative minor key of F major, the use of the F major chord progression generated by AH-0.5 is considered acceptable because it represents the relative major key of D minor. The relative major and minor keys share many harmonically related chords, making both progressions suitable for harmonizing the same melody.

It can be observed that increasing the parameter Γ to 0.9 resulted in an increase in both the density and variety of chord progressions generated by AutoHarmonizer. The generated chord progressions demonstrated two distinct types of cycles: F-C7-F, which consists of a tonic chord followed by a dominant chord and was also observed in the one generated by AH-0.5, and F-Dm-Bb-F, which includes a subdominant and supertonic chord, adding further movement to the progression. The increase in density and variety of chord progressions suggests that adjusting the Γ parameter has a significant impact on the output of AutoHarmonizer and can be a valuable tool for musicians and composers seeking to explore different chord progressions.

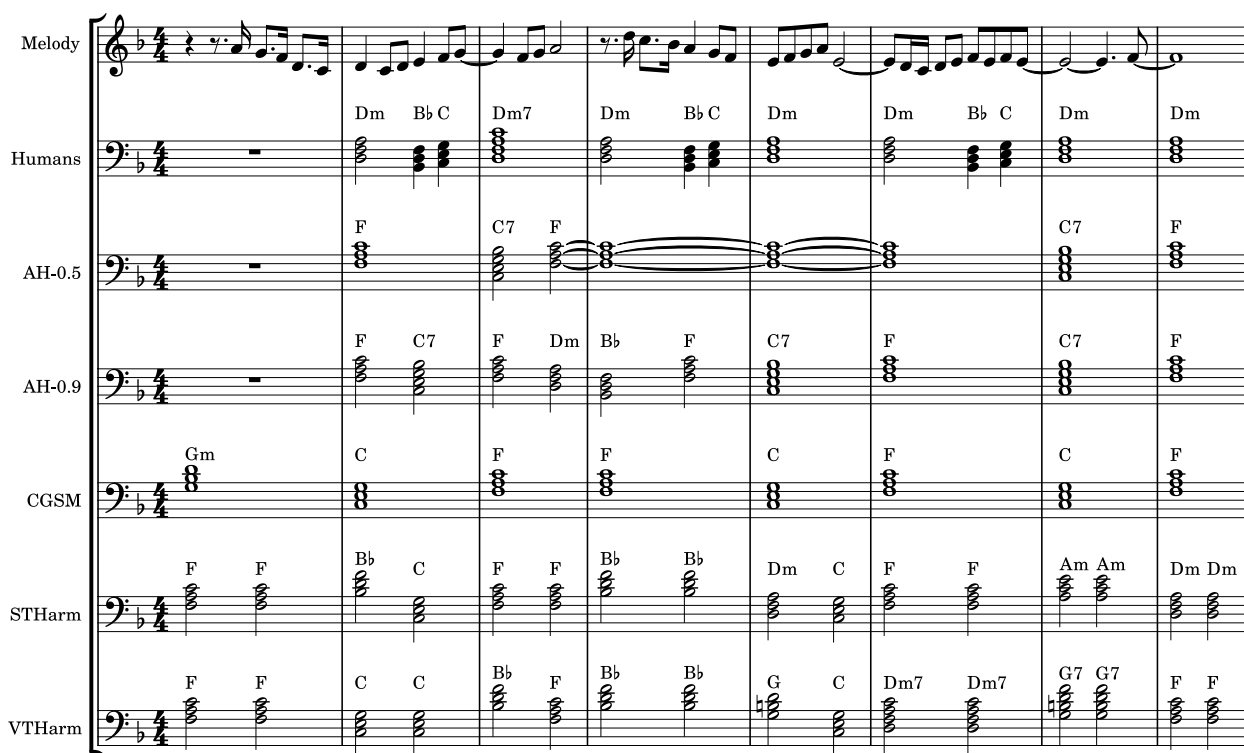


Fig. 6 Examples of melody harmonization. They are generated by various models from the same melody

The chord progression generated by CGSM differs from the others, as it employs only triads, given that this system's vocabulary does not include other chord types. Notably, the first bar is a pick-up bar (the bar before the first full bar), yet CGSM inserts a G_m chord in this position. This choice is unconventional since G_m does not correspond to a tonic, dominant, or subdominant chord. This is followed by $C-F$ repeatedly, suggesting a simple alternation between dominant and tonic. The absence of other chord variations limits the progression's expressiveness and dynamic movement.

On the other hand, chord progressions generated by STHarm display a more diverse set of chord choices, extending beyond basic triads. The presence of F , Bb , and C chords indicate a familiarity with the key of F major and its respective harmonic nuances. Furthermore, the progression's inclusion of the A_m chord adds an interesting color to the harmony, introducing a bit more diversity to the harmonic landscape. However, while its choices are more varied than those of CGSM, STHarm's progression sometimes lacks the structural cohesion and clarity seen in the human-composed and AH-generated harmonizations.

VTHarm, compared to STHarm, presents more transitions, specifically with the shift from F to C , then introducing Bb and then progressing to $Dm7$, eventually transitioning through $G7$ before resolving to F . These transitions indicate a more complex harmonic understanding than STHarm and certainly CGSM. However, its outputs occasionally diverge from traditional harmonic conventions, resulting in progressions that might be viewed as less cohesive by those familiar with music theory.

By showcasing these examples, we emphasize the nuances of each method, providing a comprehensive understanding of the strengths and weaknesses of different harmonization techniques. Additionally, we suggest that those interested can access and evaluate more generations of machine-generated chord progressions on GitHub².

6 Conclusions

AutoHarmonizer is a novel system for melody harmonization that aims to provide greater control over harmonic density and the flexibility of harmonic rhythm. The system has the ability to generate chord progressions with varying harmonic densities, thereby enabling users to create desirable chord progressions.

In order to evaluate the performance of AutoHarmonizer, a series of experiments were conducted. The results of these experiments demonstrate the effectiveness of the system in producing chord progressions

with varying harmonic rhythms. The system also enables users to control the harmonic rhythm, thereby providing greater flexibility in creating unique chord progressions.

Despite the promising results, the evaluation of AutoHarmonizer using both quantitative metrics and a discrimination test suggests that there is still room for improvement in the quality of chord progressions generated by the system. Further work is required to address these limitations and enhance the overall performance of the system.

Abbreviations

ACC	Accuracy
AH	AutoHarmonizer
CBS	Chord Beat Strength
CC	Chord Coverage
CGSM	Chord Generation from Symbolic Melody
CHE	Chord Histogram Entropy
CTD	Chord Tonal Distance
CTnCTR	Chord Tone to non-Chord Tone Ratio
GRU	Gated Recurrent Unit
HRC	Harmonic Rhythm Coverage
HRHE	Harmonic Rhythm Histogram Entropy
LSTM	Long and Short-Term Memory
MCTD	Melody-Chord Tonal Distance
PCFG	Probabilistic Context-Free Grammar
PCP	Pitch Class Profile
PCS	Pitch Consonance Score
VAE	Variational Auto-Encoder

Acknowledgements

Not applicable.

Authors' contributions

Shangda Wu initiated the research and wrote the paper, while Yue Yang and Zhaowen Wang contributed by investigating related work, conducting experiments, and writing corresponding parts. Xiaobing Li and Maosong Sun provided guidance and supervision throughout the project. All authors have reviewed and approved the final manuscript.

Authors' information

Not applicable.

Funding

This research work has been made possible by various funding sources, including the High-grade, Precision and Advanced Discipline Construction Project of Beijing Universities, the Major Projects of National Social Science Fund of China (Grant No. 21ZD19), and the Nation Culture and Tourism Technological Innovation Engineering Project.

Availability of data and materials

The dataset, source code, model weights, and model-generated samples used in this project are publicly available at <https://github.com/sander-wood/autoharmonizer>.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 9 May 2023 Accepted: 2 November 2023

Published online: 15 January 2024

² <https://sander-wood.github.io/autoharmonizer/test>

References

- A. Liu, L. Zhang, Y. Mei, B. Han, Z. Cai, Z. Zhu, J. Xiao, in *MMPT@ICMR2021: Proceedings of the 2021 Workshop on Multi-Modal Pre-Training for Multimedia Understanding, Taipei, Taiwan, August 21, 2021*, ed. by B. Liu, J. Fu, S. Chen, Q. Jin, A.G. Hauptmann, Y. Rui. Residual recurrent CRNN for end-to-end optical music recognition on monophonic scores (ACM, 2021), pp. 23–27. <https://doi.org/10.1145/3463945.3469056>
- J. Calvo-Zaragoza, D. Rizo, in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France, September 23-27, 2018*, ed. by E. Gómez, X. Hu, E. Humphrey, E. Benetos. Camera-primus: Neural end-to-end optical music recognition on realistic monophonic scores (2018), pp. 248–255. http://ismir2018.ircam.fr/doc/pdfs/33_Paper.pdf
- D. Ghosal, M.H. Kolekar, in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, ed. by B. Yegnanarayana. Music genre recognition using deep neural networks and transfer learning (ISCA, 2018), pp. 2087–2091. <https://doi.org/10.21437/Interspeech.2018-2045>
- E. Dervakos, N. Kotsani, G. Stamou, in *Artificial Intelligence in Music, Sound, Art and Design - 10th International Conference, EvoMUSART 2021, Held as Part of EvoStar 2021, Virtual Event, April 7-9, 2021, Proceedings, Lecture Notes in Computer Science*, vol. 12693, ed. by J. Romero, T. Martins, N. Rodríguez-Fernández. Genre recognition from symbolic music with cnns (Springer, 2021), pp. 98–114. https://doi.org/10.1007/978-3-030-72914-1_7
- J. Briot, G. Hadjeres, F. Pachet, Deep learning techniques for music generation - A survey. CoRR abs/1709.01620 (2017), accessed on November 27, 2023. [@article{DBLP:journals/corr/abs-1709-01620](https://arxiv.org/abs/1709.01620)
- L. Casini, M. Gustavo, M. Rocchetti. Some Reflections on the Potential and Limitations of Deep Learning for Automated Music Generation. In: Proceedings of the 29th IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) 2018, Bologna, Italy, September 9-12, 2018, pp. 27-31. [Online]. Available: <https://doi.org/10.1109/PIMRC.2018.8581038>
- D. Herremans, C. Chuan, E. Chew, A functional taxonomy of music generation systems. *ACM Comput. Surv.* **50**(5), 69:1–69:30 (2017)
- D. Makris, I. Kayrdis, S. Sioutas. Automatic melodic harmonization: An overview, challenges and future directions. Trends in music information seeking, behavior, and retrieval for creativity (2016), 146–165.
- W. Sun, J. Wu, S. Yuan, in *IEEE International Conference on Multimedia and Expo Workshops, ICME Workshops 2022, Taipei, Taiwan, July 18-22, 2022*. Melodic skeleton: A musical feature for automatic melody harmonization (IEEE, 2022), pp. 1–6. <https://doi.org/10.1109/ICMEW56448.2022.9859421>
- H. Tsushima, E. Nakamura, K. Itoyama, K. Yoshii. Function- and Rhythm-Aware Melody Harmonization Based on Tree-Structured Parsing and Split-Merge Sampling of Chord Sequences. In: Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR) 2017, Suzhou, China, October 23-27, 2017, pp. 502–508. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/2_Paper.pdf
- G. Brunner, Y. Wang, R. Wattenhofer, J. Wiesendanger, in *29th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2017, Boston, MA, USA, November 6-8, 2017*. Jambot: Music theory aware chord based generation of polyphonic music with lstms (IEEE Computer Society, 2017), pp. 519–526. <https://doi.org/10.1109/ICTAI.2017.00085>
- C. Sun, Y. Chen, H. Lee, Y. Chen, H. Wang, in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. Melody harmonization using orderless nade, chord balancing, and blocked gibbs sampling (IEEE, 2021), pp. 4145–4149. <https://doi.org/10.1109/ICASSP39728.2021.9414281>
- Y. Yin-Cheng, H. Wen-Yi, S. Fukayama, T. Kitahara, B. Genchel, H-M. Liu, H-W. Dong, Y. Chen, T. Leong, Y-H. Yang. Automatic Melody Harmonization with Triad Chords: A Comparative Study. CoRR, volume abs/2001.02360, 2020. [Online]. Available: <http://arxiv.org/abs/2001.02360>
- Y-W. Chen, H-S. Lee, Y-H. Chen, H-M. Wang. SurpriseNet: Melody Harmonization Conditioning on User-controlled Surprise Contours. In: Jin Ha Lee, Alexander Lerch, Zhiyao Duan, Juhan Nam, Preeti Rao, Peter van Kranenburg, Ajay Srinivasamurthy (Eds.), Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021, pp. 105-112. [Online]. Available: <https://archives.ismir.net/ismir2021/paper/000012.pdf>
- S. Wu, M. Sun. Efficient and Training-Free Control of Language Generation. 2023. [Online]. Available: <https://arxiv.org/abs/2205.06036>
- H. Lim, S. Rhyu, K. Lee. Chord Generation from Symbolic Melody Using BLSTM Networks. In: Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR) 2017, Suzhou, China, October 23-27, 2017, pp. 621-627. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/134_Paper.pdf
- F.T. Liang, M. Gotham, M. Johnson, J. Shotton. Automatic Stylistic Composition of Bach Chorales with Deep LSTM. In: Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR) 2017, Suzhou, China, October 23-27, 2017, pp. 449-456. [Online]. Available: https://ismir2017.smcnus.org/wp-content/uploads/2017/10/156_Paper.pdf
- G. Hadjeres, F. Pachet, F. Nielsen. DeepBach: a Steerable Model for Bach Chorales Generation. In: Proceedings of the 34th International Conference on Machine Learning (ICML) 2017, Sydney, NSW, Australia, August 6-11, 2017, pp. 1362-1371. [Online]. Available: <http://proceedings.mlr.press/v70/hadjeres17a.html>
- C.A. Huang, C. Hawthorne, A. Roberts, M. Dinculescu, J. Wexler, L. Hong, J. Howcroft, The bach doodle: Approachable music composition with machine learning at scale. CoRR abs/1907.06637 (2019). 1907.06637
- W. Yang, P. Sun, Y. Zhang and Y. Zhang, "CLSTMS: A Combination of Two LSTM Models to Generate Chords Accompaniment for Symbolic Melody," 2019 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS), Shenzhen, China, 2019, pp. 176-180. <https://doi.org/10.1109/HPBDIS.2019.8735487>
- M. Majidi, R.M. Toroghi, Music harmony generation, through deep learning and using a multi-objective evolutionary algorithm. CoRR abs/2102.07960 (2021). 2102.07960
- S. Rhyu, H. Choi, S. Kim, K. Lee, Translating melody to chord: Structured and flexible harmonization of melody with transformer. *IEEE Access* **10**, 28261–28273 (2022). <https://doi.org/10.1109/ACCESS.2022.3155467>
- A. Roberts, J. Engel, D. Eck. Hierarchical Variational Autoencoders for Music. In: Proceedings of [Conference Name], [Conference Date], [Conference Location]. [Online]. Available: <https://api.semanticscholar.org/CorpusID:4483033>
- Y. Luo, K. Agres, D. Herremans, in *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019, Delft, The Netherlands, November 4-8, 2019*, ed. by A. Flexer, G. Peeters, J. Urbano, A. Volk. Learning disentangled representations of timbre and pitch for musical instrument sounds using gaussian mixture variational autoencoders (2019). pp. 746–753. <http://archives.ismir.net/ismir2019/paper/000091.pdf>
- Y. Zhang, Z. Wang, D. Wang, G. Xia, (2020). BUTTER: A representation learning framework for bi-directional music-sentence retrieval and generation. In Proceedings of the 1st workshop on nlp for music and audio (nlp4musa) (pp. 54-58).
- K. Chen, C. Wang, T. Berg-Kirkpatrick, S. Dubnov, in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, ed. by J. Cumming, J.H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, T. de Reuse. Music sketchnet: Controllable music generation via factorized representations of pitch and rhythm (2020), pp. 77–84. <http://archives.ismir.net/ismir2020/paper/000146.pdf>
- Z. Wang, Y. Zhang, Y. Zhang, J. Jiang, R. Yang, G. Xia, J. Zhao, in *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, ed. by J. Cumming, J.H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, T. de Reuse. PIANOTREE VAE: structured representation learning for polyphonic music (2020), pp. 368–375. <http://archives.ismir.net/ismir2020/paper/000096.pdf>
- S. Di, Z. Jiang, S. Liu, Z. Wang, L. Zhu, Z. He, H. Liu, S. Yan, in *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, ed. by H.T. Shen, Y. Zhuang, J.R. Smith, Y. Yang, P. Cesar, F. Metzke, B.

Prabhakaran. Video background music generation with controllable music transformer (ACM, 2021), pp. 2037–2045. <https://doi.org/10.1145/3474085.3475195>

29. C. Harte, M. Sandler, M. Gasser, (2006, October). Detecting harmonic change in musical audio. In Proceedings of the 1st ACM workshop on Audio and music computing multimedia (pp. 21–26).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
