CHORD JAZZIFICATION: LEARNING JAZZ INTERPRETATIONS OF CHORD SYMBOLS

Tsung-Ping Chen[1] **Satoru Fukayama**[2] **Masataka Goto**[2] **Li Su**[1]
[1] Institute of Information Science, Academia Sinica, Taiwan
[2] National Institute of Advanced Industrial Science and Technology (AIST), Japan
{tearfulcanon, lisu}@iis.sinica.edu.tw
{s.fukayama, m.goto}@aist.go.jp

ABSTRACT

Chord symbols, typically notating the root note and the chord quality, are extensively used yet oversimplified representation of tonal harmony and chord progressions in popular music. In spite of its convenience, the chord symbol notation only provides basic information about the chordal configuration, and leaves much room for interpretation. With such limitations, an algorithm generating merely chord symbols is usually insufficient for a wide range of music genres such as jazz. To solve this problem, we propose *chord jazzification*, a process to generate realistic chord configurations in jazz style. With deep learning approaches, we decompose chord jazzification into *coloring* and *voicing*. Coloring concerns the choice of color tones, while voicing concerns the configurations of chords. We also create a new dataset featuring interpretations of chord symbols in pop-jazz compositions. By conducting experiments on the new dataset, we show that 1) the two-stage process outperforms an end-to-end generation approach in modeling chord configurations, and 2) attention-based models are better at capturing the structure of chord sequences in comparison with recurrent neural networks.

1. INTRODUCTION

Harmony and chords are the central topic in the study of tonal music. To facilitate the study, researchers in the fields of music information retrieval (MIR) and computational musicology have developed various techniques, such as automatic chord recognition [1–5], chord similarity and tonal distance [6–9], harmonic analysis [10–13], and chord generation [14–16].

Most of these aforementioned techniques process the chord data using *chord symbol* representation, i.e., a symbolic notation system indicating the root note, quality, and other additional information of the chord. For example, the chord symbol `C:7` stands for the C dominant seventh chord in root position, whose theoretical configuration is depicted in Figure 1a. This symbol, however, does not explicitly point out the actual configuration, e.g., every specific note performed by a musician. Hence, the notation system is limited in describing the nuances in real-world performance. Learning to interpret the chord symbols is challenging in two aspects. First, expert musicians often *color* a chord by adding notes to, or omitting notes from the chord according to the musical context, although the chord symbol itself does not specify such alterations. Second, chords composed of the same set of pitch classes can be *voiced* differently by spacing and doubling the chord tones. Take the *comping* technique in jazz music as an instance.[1] Instead of sticking to the typical configuration of a chord, a jazz pianist may play the C dominant seventh chord with the configuration shown in Figure 1b, in which the 5th of the root, `G4`, is omitted, and the 9th of the root, `D5`, is added; while another jazz pianist may arrange these notes in a totally different way as demonstrated in Figure 1c, where `Bb` is doubled and `D` is spaced a register lower. In fact, the way how musicians interpret chord symbols in a given context involves not only their thorough understanding of various musical styles, but also their personal tastes. Therefore, to generate the realization of chord symbols through MIR approaches is a challenging yet valuable task, despite this topic is rarely discussed possibly because of the lack of data.

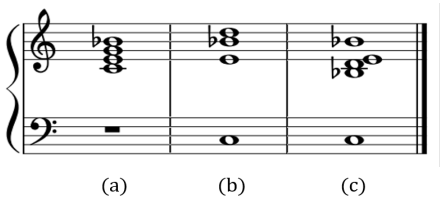
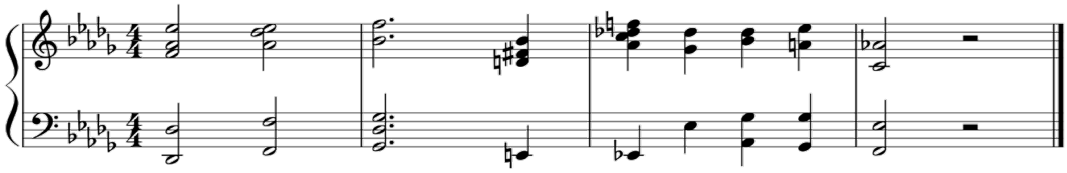


Figure 1: Three configurations of the C dominant seventh chord. In music theory, the C dominant seventh chord in root position is configured as (a). (b) and (c) are two alterations of the chord.

In this paper, we propose *chord jazzification*, a process to realize chord symbols with jazz harmony through deep learning approaches. Based on the two aforementioned aspects of interpreting the chord symbols, the chord jazz-

[1] Comping means accompanying or complementing a soloist by playing the chords to fill the harmonic and rhythmical vacancies in the music.



Chord Symbol	Db:M	Db:M/F	Gb:M7	D:M/E	Bb:m7/Eb	Eb:m7	Ab:7	G:dim	F:m7
Time: Onset	55	57	59	62	63	64	65	66	67
Time: Duration	2	2	3	1	1	1	1	1	2
Voicing	(Db2,Db3,F4,Ab4,Eb5) (F2,F3,Ab4,Db5,Eb5)		(Gb2,Db3,Gb3,Bb4,F5) (E2,D4,F#4,Bb4)		(Eb2,Ab4,C5,Db5,F5) (Eb3,Gb4,Db5)		(Ab2,Gb3,Bb4,Db5) (Gb2,Gb3,A4,Eb5)		(F2,Eb3,C4,Ab4)
Coloring	(2)	(2)	-	(o5,b6)	(o1,9)	(o5)		(o5,d7) (o3,o5,9,11)	-
Roman Numeral Analysis	Db: I	VI^{6}	IV^{7}	bIII$_{7}$	ii$^{7}_{4}$ – $_{3}$		V$_{7}$	vii$^{o}_{4}$ $_{3}$	iii^{7}
Structure: Phrase	B1	B1	B1	B1	B1	B1	B1	B1	B1
Structure: Measure	15	15	16	16	17	17	17	17	18
Structure: Metrical Position	0	2	0	3	0	1	2	3	0

Figure 2: Annotations of the proposed dataset. *Chord Symbol* specifies the root and the quality of each chord (a bass note is explicitly notated with a slash when it is not the root note), e.g., `Db:M/F` stands for a Db major triad with the bass note `F`. *Time* indicates the *onset* and the *duration* (measured in beats) of each chord. *Voicing* represents the configuration of each chord with a set of scientific pitch notations. *Coloring* indicates the chord degrees which appear in the voicing but are not specified by the chord symbol, and vice versa, e.g., `(o1,9)` indicates that the root is omitted and the 9th is added. Note that `7` is explicitly qualified by {`M`, `m`, `d`} (major, minor, diminished) to disambiguate its interval. *Roman Numeral Analysis* denotes the scale degree on which a chord is built, as well as the quality and the inversion information, e.g., `V7` stands for a dominant seventh chord in root position. *Structure* includes the annotations of *phrase*, *measure* and *metrical position*. To specify the phrase a chord belongs to, each chord is labeled with a letter plus a number in a way similar to musical form analysis. Metrical position shows the position of a chord with respect to the metric grid (starting with 0).

ification task is formulated as two subtasks, namely the *chord coloring* and the *chord voicing*. The chord coloring task decides which pitch classes are to be played for elaborating a chord symbol, while the chord voicing task deals with the spacing and the doubling of the pitch classes to be played. Although the jazzification of chord progressions is not limited to coloring and voicing, we focus on the two aspects for the primary study. To facilitate the research on chord jazzification, we also compile a new dataset consisting of chord symbols and corresponding chord configurations in pop-jazz compositions. In comparison with other jazz-related datasets, such as the Charlie Parker's Omnibook data [17], the Jazz Audio-Aligned Harmony (JAAH) dataset [18], the JazzCorpus [19], and the Weimar Jazz Database [20], our dataset includes more detailed information of chords, especially the chordal configuration. With the newly compiled dataset, we conduct several experiments to verify the two-stage framework for chord jazzification. Experiment results indicate that it is effective to decompose chord jazzification into coloring and voicing, rather than to adopt an end-to-end approach.

The current work is different from the accompaniment or harmonization tasks [21,22], in terms of that such works require melodies as prior knowledge and regard chords as appendages to melodies. Besides, our work concerns both the voicings of a chord sequence and the interpretation of each single chord, thus is distinct from the voice leading generation task [23, 24]. The jazzification of a chord progression is part of the music composition process to expand a tonal structure, i.e., the *prolongation* described in Heinrich Schenker's music theory [25]. Chord jazzification can advance other MIR-related tasks such as style analysis and chord generation [26–29]. We hope that our work can draw more attention to the musical knowledge regarding the implicit relations between the notated chord symbols and the actual harmony being performed.

In summary, our contribution is threefold. First, we address the issue of interpreting chord symbols by chord jazzification. Second, we compile a new dataset for the generation of jazzy harmony and for the study of chord embellishments. Finally, a deep learning framework is proposed to generate jazz-style chord progressions. In the following, we will first present the new dataset (Section 2), and then formulate the framework of chord jazzification (Section 3); based on the framework, several experiments are introduced thereon (Section 4).

2. THE CHORD JAZZIFICATION DATASET

The corpus is composed of 50 musical pieces selected from published Japanese pop-jazz piano solos, in which chord symbols are explicitly specified. [2] To obtain the corresponding voicing of each chord symbol, we manually perform harmonic reduction for each piece. Concretely, notes within the region of a chord symbol are selected to build the configuration of the chord symbol. With the chord

[2] The dataset is available at `https://github.com/Tsung-Ping/Chord-Jazzification`.

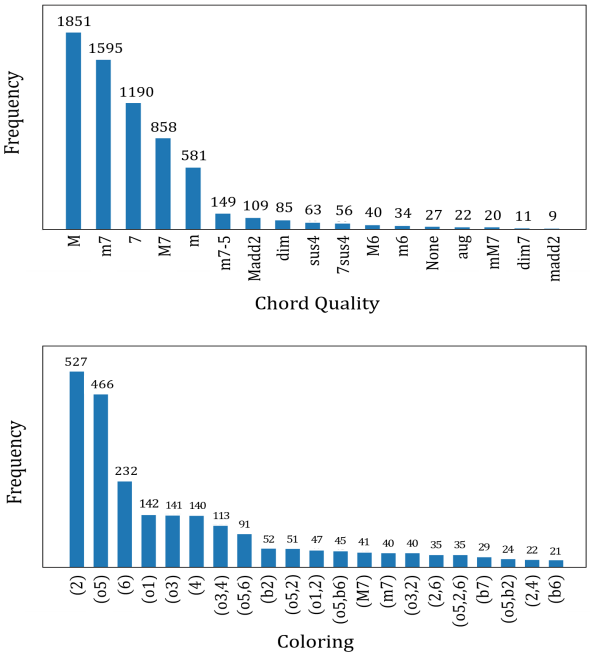


Figure 3: Chord qualities and chord colorings in the proposed dataset (the long tail of the coloring distribution is left out). Chord degrees of compound intervals are merged with simple ones in the coloring figure for simplicity.

symbols and the transcribed voicings, coloring and harmonic information are annotated. Specifically, there are six types of annotations in the dataset: *Chord Symbol*, *Time*, *Voicing*, *Coloring*, *Roman Numeral Analysis* [10, 30], and *Structure*. Figure 2 gives an example.

A brief introduction of the chord qualities and colorings used in the corpus is illustrated in Figure 3. Not surprisingly, the three seventh chords, major seventh (`M7`), minor seventh (`m7`), and dominant seventh (`7`), account for more than half of the chords, as jazz harmony is notable for the use of seventh chords. On the other hand, characteristic colorings of pop-jazz music can also be found in many chords, such as adding a major second or a compound major second (`2`) and omitting the perfect fifth (`o5`).

In summary, 796 musical phrases amounting to 6700 chord labels are included in the dataset. These annotations provide information concerning the relationship between the symbolic notation and the actual configuration of chords in human compositions. It therefore has the potential to be applied to many MIR-related research topics, such as corpus-based study of tonal harmony in music practice, generation of colorful chord progressions, and computer-aided composition, to name but a few. It has to be acknowledged that the dataset has some limitations in describing an actual interpretation of chord symbols, in the sense that the transcription of chords eliminates the harmonic and rhythmical variances within the region of each chord symbol. Nevertheless, the dataset can be a starting point for performers and composers to learn to elaborate and develop a tonal structure through the chord jazzification process, which is relatively easier compared to learning the elaboration directly from a complete musical piece.

3. CHORD JAZZIFICATION

The goal of chord jazzification is to endow plain chords (e.g., a sequence of triads) with jazz harmony. We tackle the chord jazzification task through two successive steps, that is, coloring and voicing. The coloring part functions as an intermediate state which specifies chords in terms of pitch classes, and the voicing part assigns pitch heights to the specified pitch classes.

3.1 Coloring

In this paper, 48 triads in root position ({major, minor, augmented, diminished} by 12 semitones) are considered for coloring. We define the *chord coloring* task as predicting the bass note and the pitch classes to render each triad in a given sequence.

Formally, the input of the coloring task is a triad sequence $\{\mathbf{x}_i\}_{i=1}^{T}$ and a duration sequence $\{\mathrm{d}_i\}_{i=1}^{T}$, where i denotes time steps, T is the length of the sequence, $\mathbf{x}_i \in \mathbb{R}^{12}$ is a chroma representation of the ith triad, and d_i is the duration of $\mathbf{x}_i$. The coloring task predicts the bass sequence $\{\mathbf{b}_i^c\}_{i=1}^{T}$ and the pitch class sequence $\{\mathbf{p}_i^c\}_{i=1}^{T}$ for the input sequence, where c stands for coloring, $\mathbf{b}_i^c \in \mathbb{R}^{12}$ is a softmax-activated chroma vector indicating the 12 pitch classes' probabilities to be the bass of colored $\mathbf{x}_i$, and $\mathbf{p}_i^c \in \mathbb{R}^{12}$ is a sigmoid-activated chroma vector indicating the 12 pitch classes' probabilities to be the constituent notes (except the bass) of colored $\mathbf{x}_i$. For example, $\mathbf{b}_i^c = [0.8, 0, 0, 0, 0.2, 0, 0, 0, 0, 0, 0, 0]$ indicates the pitch classes `C` and `E` respectively have a probability of 80% and 20% to be the bass note, and $\mathbf{p}_i^c = [0, 0.9, 0, 0, 0.9, 0, 0, 0, 0, 0, 0.9, 0]$ suggests that there is a 90% chance that `Db` , `E`, and `Bb` are activated to render $\mathbf{x}_i$.

We employ a basic sequential learning architecture for the coloring task. As shown in Figure 4a, the architecture is composed of three layers: 1) an input embedding layer, 2) a sequential modeling layer, and 3) an output layer. Specifically, the three layers are formulated as follows:

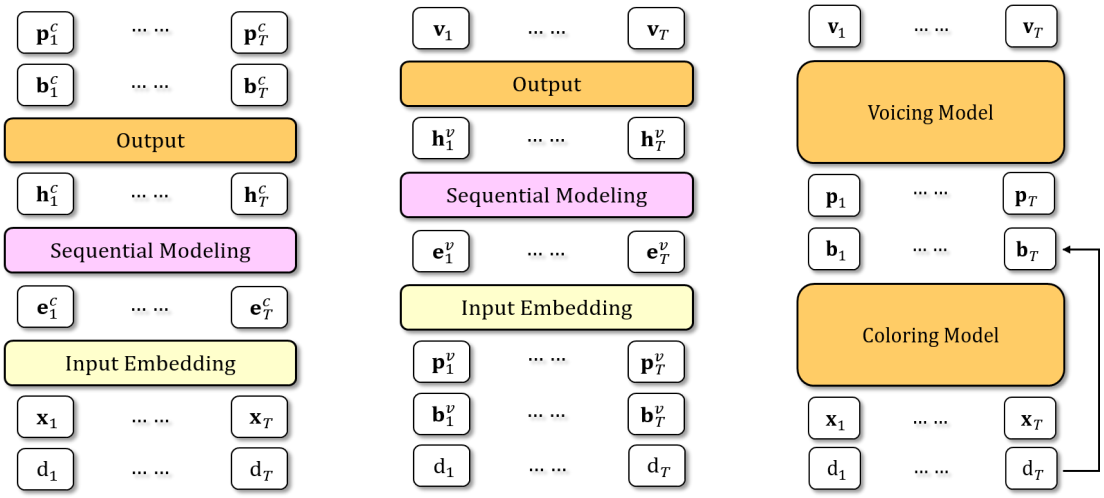
$$\begin{aligned}
\mathbf{e}_i^c &= \mathbf{W}^{e^c}(\mathrm{d}_i \mathbf{x}_i), \quad \text{(Input Embedding)} \\
\mathbf{h}_i^c &= f^c(\mathbf{e}_i^c \mid \mathbf{e}_{1:T}^c), \quad \text{(Sequential Modeling)} \\
\mathbf{p}_i^c &= \mathrm{sigmoid}(\mathbf{W}^{p^c} \mathbf{h}_i^c), \quad \text{(Output)} \\
\mathbf{b}_i^c &= \mathrm{softmax}(\mathbf{W}^{b^c} \mathbf{h}_i^c),
\end{aligned} \tag{1}$$

where $\mathbf{W}^{e^c} \in \mathbb{R}^{d \times 12}$, $\mathbf{W}^{p^c} \in \mathbb{R}^{12 \times d}$, and $\mathbf{W}^{b^c} \in \mathbb{R}^{12 \times d}$ are learnable parameters, $f^c : \mathbb{R}^d \to \mathbb{R}^d$ denotes a trainable neural network, and d is a hyperparameter indicating the dimensions of the embedding space. Two candidate networks are employed for the sequential modeling layer:

- Bi-directional Recurrent Neural Network with Long Short-Term Memory (BLSTM):

$$\begin{aligned}
\mathbf{h}_i^c &= \overrightarrow{\mathbf{h}}_i^c \oplus \overleftarrow{\mathbf{h}}_i^c, \\
\overrightarrow{\mathbf{h}}_i^c &= \mathrm{LSTM}(\mathbf{e}_i^c \mid \mathbf{e}_{1:i-1}^c), \\
\overleftarrow{\mathbf{h}}_i^c &= \mathrm{LSTM}(\mathbf{e}_i^c \mid \mathbf{e}_{i+1:T}^c),
\end{aligned} \tag{2}$$

  where $\oplus$ denotes vector concatenation.



(a) (b) (c)

Figure 4: (a) The coloring model. (b) The voicing model. (c) The two-stage chord jazzification model.

- Multihead Self-attention Network (MHSA):

$$
\begin{aligned}
\mathbf{h}_i^{c(l)} &= \mathbf{W}^{outer}\sigma(\mathbf{W}^{inner}\mathbf{u}_i + \mathbf{b}_1) + \mathbf{b}_2, \\
\mathbf{u}_i &= \mathbf{W}^{u}(\mathbf{u}'_{i1} \oplus \cdots \oplus \mathbf{u}'_{iJ}) \; + \; \mathbf{h}_i^{c(l-1)}, \\
\mathbf{u}'_{ij} &= \mathbf{V}_j \,\mathrm{softmax}\left(\frac{\mathbf{K}_j^{\top}\mathbf{q}_{ij}}{\sqrt{d}}\right), \\
\mathbf{q}_{ij} &= \mathbf{W}_j^{Q}\mathbf{h}_i^{c(l-1)}, \\
\mathbf{K}_j &= \mathbf{W}_j^{K}[\mathbf{h}_1^{c(l-1)}, \cdots, \mathbf{h}_T^{c(l-1)}], \\
\mathbf{V}_j &= \mathbf{W}_j^{V}[\mathbf{h}_1^{c(l-1)}, \cdots, \mathbf{h}_T^{c(l-1)}],
\end{aligned}
\tag{3}
$$

where l denotes the iteration step, and the initial value $\mathbf{h}_i^{c(0)} = \mathbf{e}_i^c$; σ represents the ReLU activation function; J is the number of heads; $\mathbf{W}_j^{outer} \in \mathbb{R}^{d\times 4d}$, $\mathbf{W}_j^{inner} \in \mathbb{R}^{4d\times d}$, $\mathbf{W}^{u} \in \mathbb{R}^{d\times d}$, and $\mathbf{W}_j^{Q}, \mathbf{W}_j^{K}, \mathbf{W}_j^{V} \in \mathbb{R}^{\frac{d}{J}\times d}$ are learnable parameters. This network is equivalent to the encoder of the Transformer [31], while we leave out layer normalization and position encoding terms for simplicity. In this paper, we set $d = 512$, $l = 2$, and $J = 8$.

The binary cross entropy (BCE) and the categorical cross entropy (CCE) are used to calculate the losses. Let $\mathbf{p}_i^{c*}$ and $\mathbf{b}_i^{c*}$ denote the ground truths of $\mathbf{p}_i^c$ and $\mathbf{b}_i^c$; the total loss of the coloring model $\mathcal{L}^c$ is defined as:

$$
\mathcal{L}^c = \sum_{i=1}^{T}\left[\mathrm{BCE}(\mathbf{p}_i^{c*}, \mathbf{p}_i^c) + \mathrm{CCE}(\mathbf{b}_i^{c*}, \mathbf{b}_i^c)\right]. \tag{4}
$$

3.2 Voicing

We define *chord voicing* as a task which predicts the voicings of a chord sequence. Formally, given a chord sequence of T time steps in terms of their basses $\{\mathbf{b}_i^v\}_{i=1}^T$, constituent pitch classes $\{\mathbf{p}_i^v\}_{i=1}^T$, and durations $\{\mathrm{d}_i\}_{i=1}^T$, the task predicts the voicings $\{\mathbf{v}_i\}_{i=1}^T$ for the chord sequence, where v stands for voicing, $\mathbf{b}_i^v \in \mathbb{R}^{12}$ is a one-hot chroma vector indicating the bass of the ith chord, $\mathbf{p}_i^v \in \mathbb{R}^{12}$ is a multi-hot chroma vector representing the pitch classes of the ith chord except the bass note, and $\mathbf{v}_i \in \mathbb{R}^{88}$ is a voicing vector indicating the 88 tones' probabilities to be played on the piano.

Similar to the coloring task, we employ a 3-layer architecture for the voicing task, as shown in Figure 4b. The three layers are formulated as follows:

$$
\begin{aligned}
\mathbf{e}_i^v &= \mathbf{W}^{e^v}(\mathrm{d}_i(\mathbf{p}_i^v \oplus \mathbf{b}_i^v)), \;\; \text{(Input Embedding)} \\
\mathbf{h}_i^v &= f^v(\mathbf{e}_i^v \mid \mathbf{e}_{1:T}^v), \;\; \text{(Sequential Modeling)} \\
\mathbf{v}_i &= \mathrm{sigmoid}(\mathbf{W}^v\mathbf{h}_i^v), \;\; \text{(Output)}
\end{aligned}
\tag{5}
$$

where $\mathbf{W}^{e^v} \in \mathbb{R}^{d\times 24}$ and $\mathbf{W}^v \in \mathbb{R}^{88\times d}$ are learnable parameters, and $f^v : \mathbb{R}^d \to \mathbb{R}^d$ is a neural network. Likewise, the BLSTM and the MHSA networks are two options for the sequential modeling layer.

Let $\mathbf{v}_i^* \in \mathbb{R}^{88}$ denote the target voicing of the ith chord; we define the loss as:

$$
\mathcal{L}^v = \sum_{i=1}^{T}\mathrm{BCE}(\mathbf{v}_i^*, \mathbf{v}_i). \tag{6}
$$

As the voicing task is to arrange the constituent notes of chords on an 88-key piano based on the given basses and the sets of pitch classes, the outcome of $\mathbf{v}_i^*$ can be known to a certain degree. More precisely, a note in $\mathbf{v}_i^*$ can be activated only if its pitch class is activated in $\mathbf{b}_i^v$ or $\mathbf{p}_i^v$. With this consideration, we design corresponding masks to modify the loss computation. Let $\mathbf{b}_i^{v\prime} \in \mathbb{R}^{88}$ and $\mathbf{p}_i^{v\prime} \in \mathbb{R}^{88}$ be the extensions of $\mathbf{b}_i^v$ and $\mathbf{p}_i^v$ to all octaves of the piano. Then, the loss constrained by the masks becomes:

$$
\begin{aligned}
\mathcal{L}^{v\prime} &= \sum_{i=1}^{T}\mathrm{BCE}(\mathbf{v}_i^*, \mathbf{m}_i \odot \mathbf{v}_i), \\
\mathbf{m}_i &= \mathbf{b}_i^{v\prime} \vee \mathbf{p}_i^{v\prime}, \;\; \text{(Mask)}
\end{aligned}
\tag{7}
$$

where $\odot$ stands for the Hadamard product, and $\vee$ denotes the logical OR operator.

3.3 Two-stage Chord Jazzification

We stack the chord voicing model on the top of the chord coloring model by setting $\mathbf{b}_i = \mathbf{b}_i^v = \text{onehot}(\arg\max \mathbf{b}_i^c)$ and $\mathbf{p}_i = \mathbf{p}_i^v = \text{round}(\mathbf{p}_i^c)$, as illustrated in Figure 4c. In other words, the outputs of the coloring model are first converted into binary vectors, and then taken as inputs by the voicing model. Such an integrated model jazzifies chord progressions in two stages: first, for a given sequence of triads $\{\mathbf{x}_i\}_{i=1}^T$ and the corresponding durations $\{\mathrm{d}_i\}_{i=1}^T$, the coloring model generates a colored sequence represented by $\{\mathbf{b}_i\}_{i=1}^T$ and $\{\mathbf{p}_i\}_{i=1}^T$; based on the colored sequence, the voicing model subsequently generates a voiced chord progression $\{\mathbf{v}\}_{i=1}^T$.

4. EXPERIMENTS

4.1 Chord Jazzification with Supervised Learning

With the formulations of chord jazzification as multi-class classification (for $\mathbf{b}_i^c$) and multi-label classification (for $\mathbf{p}_i^c$ and $\mathbf{v}_i$) problems, we train the coloring and voicing models using the new dataset, and perform 4-fold cross validation. For the coloring task, the input triad sequences and the input duration sequences are respectively derived from the *Chord Symbol* and the *Time* annotations of the dataset, while the ground truths of the bass sequences and the pitch class sequences are obtained from the *Voicing* labels. As for the voicing task, the duration sequences and the ground truth labels of the coloring task are taken as the inputs, while the *Voicing* labels are used as the ground truths of the output sequences. We augment the training set through transposing the data from 4 semitones down to 5 semitones up (within the valid range of the piano), leading to 10 times the training data. As a result, there are 5970 and 199 sequences for training and testing respectively.

Evaluation results are shown in Table 1. For both the coloring and voicing tasks, the employment of either the BLSTM or the MHSA as the sequential modeling layers yields comparable performance to the other, while the MHSA appears to surpass the BLSTM in cases of multi-label classification, i.e., the predictions of pitch classes and voicing. When the input embedding layers in the two subtasks are removed, all the performances decrease by from 3.19% to 4.69%. This indicates that the transformation to dense vectors benefits the learning process when the input data is sparsely represented. Moreover, the introduction of the input-related masks to the loss calculation in Eqn (7) also improves the modeling of voicing; precisely, the F1 score increases 2.66% if the masks are utilized. It is worth noting that the amount of training data is quite limited, and therefore the performance seems to be satisfactory in the current experimental setting.

4.2 End-to-end Chord Jazzification

To motivate the decomposition of chord jazzification into 2 stages, we train a chord jazzification model in an end-to-end manner for comparison. Technically, we replace the output module of the coloring model with that of the voicing model; and we employ a 2-layer BLSTM, rather than a 1-layer BLSTM as defined in Eqn (2), for the sequential modeling layer in order to make the number of parameters comparable to the two-stage chord jazzification model. The evaluation result is shown in Table 1.

Model	Coloring		Voicing
	Bass	Pitch Classes	
BLSTM	**81.87**	76.52	63.64
MHSA	80.78	**77.02**	**64.86**
BLSTM w/o E	77.18	73.33	60.12
BLSTM w/o M	-	-	60.98
End-to-End	-	-	37.87

Table 1: Results of the coloring and the voicing tasks. The lower part shows the ablation tests without the embedding layer (w/o E) and without masks (w/o M), as well as the result using end-to-end training. All the values indicate the average F1 scores (%) over 4 validations.

In this end-to-end architecture, the performance drops substantially to nearly half of the value. This result conversely validates the two-stage approach. Given the fact that the prediction of polyphony is often challenging, it turns out to be beneficial to generate an intermediate stage, that is, the chroma representations with respect to coloring, before the overall jazzification of chords.

4.3 Consistency of Chord Jazzification

Chord progressions often have a repetitive structure, therefore it is important for a model to preserve this property and generate chord sequences of self-consistency. To measure the consistency of a model's generations, we compute the self-similarity matrices (SSMs) of each generated voicing sequence and corresponding label sequence, and then calculate the difference between each generation-label SSM pair. Let $\overline{\mathbf{V}} = [\overline{\mathbf{v}}_1, \cdots, \overline{\mathbf{v}}_T]$ denote the normalized voicing sequence generated by a model, and $\overline{\mathbf{V}}^* = [\overline{\mathbf{v}}_1^*, \cdots, \overline{\mathbf{v}}_T^*]$ denote the normalized label sequence, where $\overline{\mathbf{v}}_i = \frac{\text{round}(\mathbf{v}_i)}{\|\text{round}(\mathbf{v}_i)\|}$ is a binarized and normalized voicing, and $\overline{\mathbf{v}}_i^* = \frac{\mathbf{v}_i^*}{\|\mathbf{v}_i^*\|}$ is a normalized target voicing; we define the consistency score (CS) of a generated sequence as follows:

$$\begin{aligned} &\text{CS} = 1 - \text{reduce\_mean}(\Delta\text{SSM}), \\ &\Delta\text{SSM} = |\text{SSM}_{pred} - \text{SSM}_{label}|, \\ &\text{SSM}_{pred} = \overline{\mathbf{V}}^\top \overline{\mathbf{V}}, \\ &\text{SSM}_{label} = \overline{\mathbf{V}}^{*\top} \overline{\mathbf{V}}^*. \end{aligned} \tag{8}$$

The more similar the structures of the two sequences are, the higher the CS score is.

Table 2 shows the average consistency score over 4 cross-validation folds. To provide a benchmark for the consistency measure, we also show the CS score computed from label sequences and randomly-generated sequences (denoted as *RANDOM*). Both the two models get higher scores than the random condition, indicating that they learn some structural information. Besides, the MHSA outperforms the BLSTM due to an essential difference between them: the BLSTM processes each time step of a

Model	BLSTM	MHSA	RANDOM
CS Score (%)	86.91	**87.68**	72.80

Table 2: Consistency measure of chord jazzification.

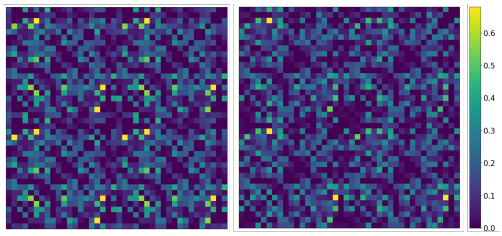


Figure 5: The difference of self-similarity matrices. Left: ΔSSM of the BLSTM. Right: ΔSSM of the MHSA. The origin of the matrices is at the lower left corner.

sequence recurrently, while the MHSA accesses the entire sequence simultaneously. As a result, the MHSA can capture more structural features than the BLSTM does, leading to more consistent generations. Two examples of the ΔSSM are demonstrated in Figure 5. Evidently, there are fewer bright regions in the ΔSSM of the MHSA, indicating that the generation by the MHSA is structurally closer to the ground truth than that by the BLSTM.

4.4 Generating Jazz Harmony

We train the two-stage chord jazzification model using the proposed dataset, and generate jazzified chord progressions with input triad sequences derived from the JAAH dataset. [3] In total, 2210 sequences with 23199 chords were generated. To examine the effect of jazzification, we quantitatively analyze the difference between each input triad and its jazzified counterpart. Particularly, we are interested in changes with respect to chord coloring: 1) what notes are added to or omitted from a triad? 2) Is a note other than the root being chosen as the bass note?

The result is represented in Figure 6. Around 40% of the input triads are embellished with a minor seventh (`m7`). And the addition of a major seventh (`M7`) also accounts for around 3.6%, ranked in the top fourth. These frequent colorings with a major or minor seventh reflect the characteristics of jazz music in which most triads that appear in lead sheets or fake books can have sevenths added to them. Moreover, extended chords and inverted chords can also be found. For instance, the coloring `(b2,m7)` for a major triad will lead to a dominant seventh flat ninth chord; and the coloring `None/3` indicates the first inversion of triads. It is worth mentioning that some generated slash chords are not inverted chords. An example is shown in the bottom of Figure 6. With the coloring `(2,M7)/2`, the last triad `Db:M` becomes `Db:M7/Eb`, which can be interpreted as an Eb dominant thirteenth chord—the dominant chord of the relative major mode (assuming `F` is the tonic). In other words, this coloring not only breaks the repetitive structure

[3] `https://github.com/MTG/JAAH`

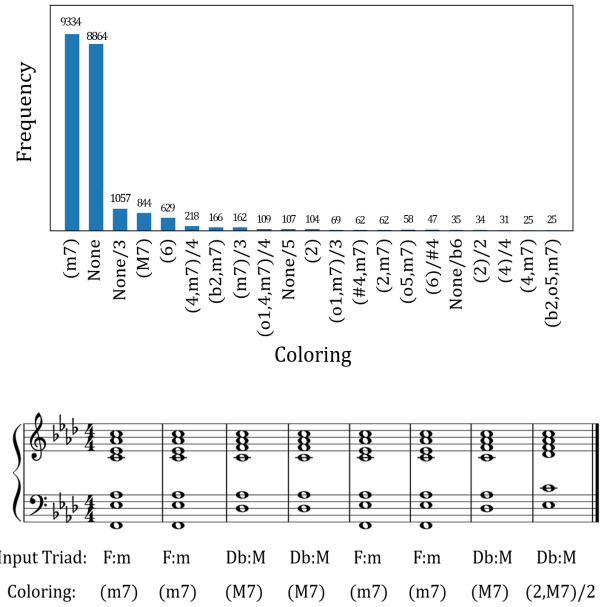


Figure 6: Top: the coloring distribution of the generated chords (part of the distribution is omitted). Bottom: a generated example. A number after the slash symbol indicates the degree of the bass note relative to the root note.

of the input triad sequence, but also implies a new tonality. In addition to coloring, it can be observed that the linear progressions of voices are quite smooth, showing that the model also learns the knowledge of voice leading.

5. CONCLUSION

To learn the interpretation of chord symbols from musical data, we proposed chord jazzification, a process of generating realistic jazz-style chord progressions through two musical techniques: chord coloring and chord voicing. Chord coloring decides a bass and a set of pitch classes for elaborating a triad, while chord voicing arranges the bass and the set of pitch classes on the piano. We correspondingly built a dataset which includes coloring and voicing annotations, and hence can be used as the training data of the chord jazzification task. By formulating the chord coloring and chord voicing tasks as classification problems, we experimentally showed that the two-stage framework is capable of generating plausible chord configurations from a sequence of chord symbols.

Chord jazzification has the potential to be applied to two different yet related practices in music: performing and composing. For music performing, it is practical and desirable to interpret the chord symbols on lead sheets through jazzification. For music composing, the generated sequence by the chord jazzification model can be regarded as an intermediate product with which a musician can further create a human-machine collaborative musical work. In future research, we are planning to include more rhythmical and structural information, such as metrical position, to better represent the harmonic features, and apply more advanced deep learning techniques to improve the chord jazzification task.

6. ACKNOWLEDGMENTS

This work was supported in part by JST ACCEL Grant Number JPMJAC1602.

7. REFERENCES

[1] F. Korzeniowski and G. Widmer, "Improved chord recognition by combining duration and harmonic language models," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 10–17.

[2] F. Korzeniowski, D. R. W. Sears, and G. Widmer, "A large-scale study of language models for chord prediction," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 91–95.

[3] J. Deng and Y. Kwok, "Large vocabulary automatic chord estimation with an even chance training scheme," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 531–536.

[4] B. McFee and J. P. Bello, "Structured training for large-vocabulary chord recognition," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 188–194.

[5] F. Korzeniowski and G. Widmer, "A fully convolutional deep auditory model for musical chord recognition," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[6] W. B. de Haas, M. Robine, P. Hanna, R. C. Veltkamp, and F. Wiering, "Comparing approaches to the similarity of musical chord sequences," in *Exploring Music Contents - 7th International Symposium (CMMR)*, 2010, pp. 242–258.

[7] W. B. de Haas, R. C. Veltkamp, and F. Wiering, "Tonal pitch step distance: a similarity measure for chord progressions," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, 2008, pp. 51–56.

[8] C. Harte, M. Sandler, and M. Gasser, "Detecting harmonic change in musical audio," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, 2006, pp. 21–26.

[9] J. Paiement, D. Eck, and S. Bengio, "A probabilistic model for chord progressions," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, 2005, pp. 312–319.

[10] T. Chen and L. Su, "Functional harmony recognition of symbolic music data with multi-task recurrent neural networks," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 90–97.

[11] N. Condit-Schultz, Y. Ju, and I. Fujinaga, "A flexible approach to automated harmonic analysis: multiple annotations of chorales by Bach and Prætorius," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 66–73.

[12] P. B. Kirlin and P. E. Utgoff, "A framework for automated Schenkerian analysis," in *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR)*, 2008, pp. 363–368.

[13] T. Chen and L. Su, "Harmony Transformer: Incorporating chord segmentation into harmony recognition," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 259–267.

[14] H. Lim, S. Rhyu, and K. Lee, "Chord generation from symbolic melody using BLSTM networks," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 621–627.

[15] K. Choi, G. Fazekas, and M. Sandler, "Text-based LSTM networks for automatic music composition," in *the 1st Conference on Computer Simulation of Musical Creativity*, 2016.

[16] K. Kosta, M. Marchini, and H. Purwins, "Unsupervised chord-sequence generation from an audio example," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, 2012, pp. 481–486.

[17] K. Déguernel, E. Vincent, and G. Assayag, "Using multidimensional sequences for improvisation in the OMax paradigm," in *Proceedings of the 13th Sound and Music Computing Conference (SMC)*, 2016, pp. 481–486.

[18] V. Eremenko, E. Demirel, B. Bozkurt, and X. Serra, "Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research," in *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 483–490.

[19] M. Granroth-Wilding, "Harmonic analysis of music using combinatory categorial grammar," Ph.D. dissertation, University of Edinburgh, 2013.

[20] M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach, and B. Burkhart, Eds., *Inside the Jazzomat - New Perspectives for Jazz Research*. Schott Campus, 2017.

[21] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: a steerable model for Bach chorales generation," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1362–1371.

[22] T. Kitahara, M. Katsura, H. Katayose, and N. Nagata, "Computational model for automatic chord voicing based on Bayesian network," in *Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC)*, 2008, pp. 395–398.

[23] D. Hörnel, "ChordNet: Learning and producing voice leading with neural networks and dynamic programming," *Journal of New Music Research*, vol. 33, no. 4, pp. 387–397, 2004.

[24] P. M. C. Harrison and M. T. Pearce, "A computational cognitive model for the analysis and generation of voice leadings," *Music Perception: An Interdisciplinary Journal*, vol. 37, pp. 208–224, 2020.

[25] O. Jonas, *Introduction to the Theory of Heinrich Schenker*, 2nd ed. Musicalia Press, 2005.

[26] D. Conklin, M. Gasser, and S. Oertl, "Creative chord sequence generation for electronic dance music," *Journal of Applied Sciences*, vol. 8, no. 9, p. 1704, 2018.

[27] D. Scanteianu, E. Jackson, and R. M. Keller, "A fluid chord voicing generator," in *Proceedings of the International Computer Music Conference (ICMC)*, 2016, pp. 171–175.

[28] R. Dias, C. Guedes, and T. Marques, "A computer-mediated interface for jazz piano comping," in *Music Technology meets Philosophy - From Digital Echos to Virtual Ethos: Joint Proceedings of the 40th International Computer Music Conference (ICMC), and the 11th Sound and Music Computing Conference (SMC)*, 2014, pp. 558–564.

[29] I. Simon, D. Morris, and S. Basu, "MySong: automatic accompaniment generation for vocal melodies," in *Proceedings of the 2008 Conference on Human Factors in Computing Systems (CHI)*, 2008, pp. 725–734.

[30] S. Kostka, D. Payne, and B. Almen, *Tonal Harmony*, 3rd ed. McGraw-Hill, 1995.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," in *Neural Information Processing Systems (NIPS)*, 2017, pp. 5998–6008.