Automatic Orchestration of Piano Scores for Wind Bands with User-Specified Instrumentation

Takuto Nabeoka[1], Eita Nakamura[1] and Kazuyoshi Yoshii[1⋆]

Kyoto University
eita.nakamura@i.kyoto-u.ac.jp

Abstract. We present a deep learning method for generating wind band scores with user-specified instrumentation from piano scores. The difficulty in curating large-scale pair data with accurately aligned wind band and piano scores poses two major challenges: (i) efficient preparation of training data and (ii) effective learning of orchestration rules, particularly for infrequently used instruments. We propose using an automatic piano arrangement method to generate pair data from existing wind band scores. Our method utilizes U-Net to assign notes in an input piano score to individual instrument parts, and we propose refined network architectures for efficient learning of characteristics of instrument parts in the wind band scores. We show that the method can generate partially playable scores that capture voicing rules and mutual relationships among instrument parts.

Keywords: symbolic music processing; automatic arrangement; orchestration for wind band; deep learning; U-Net.

1 Introduction

Wind band is a popular form of musical performance for amateur musicians; numerous schools and communities own wind bands. These bands often have only limited kinds of musical instruments, and the instrumentation may vary from year to year depending on the members' circumstances. Consequently, the repertoire for amateur wind bands is limited because wind band scores in the market tend to be expensive and may be difficult to perform due to discrepancies in the instrumentation of a particular band. This study aims to expand the available repertoire for wind bands by studying automatic orchestration of piano scores, which are relatively easy to obtain, for wind bands with user-specified instrumentation.

Orchestration is a challenging task even for human experts. It requires a high degree of expertise because it must take into account the simultaneous and temporal relationships among dozens of instrument parts, in addition to their pitch ranges and characteristics [1,2]. A previous study developed a method for converting a large wind band score

⋆ We thank Moyu Terao for cooperation and Hitomi Kaneko for useful discussions. This work was supported by JST PRESTO No. JPMJPR20CB, JSPS KAKENHI Nos. 19H04137, 21H03572, 21K02846, 21K12187, 22H03661, and JST FOREST Program No. JPMJPR226X.

to a smaller one, by implementing manually constructed criteria for phrase segmentation, instrument group extraction, and instrument assignment [3]. This approach cannot be easily generalized for orchestration of piano scores and it is difficult to manually set up constraints to incorporate all of the aforementioned aspects of expert knowledge. A viable approach is then to use machine learning and infer such constraints from data. Another study explored a spectral-based approach for orchestration [4]. However, it is inappropriate for orchestrating piano scores since spectral features cannot directly capture relevant musical structures such as melody and bass lines.

Recent studies have explored the potential of deep neural networks (DNNs) and statistical models for automatic music generation and arrangement (e.g. [5, 6]). A study attempted automatic orchestration of piano scores using a restricted Boltzmann machine [7]. It was shown that curation of pair data with accurately aligned orchestra and piano scores requires high cost [8]. A more recent study used a Transformer to generate symphonic music using a larger dataset [9]. In these studies, how to control the instrumentation and assure the playability of the output was not focused on. The problem in data curation is even more severe when we allow arbitrary instrumentations because some instruments are much less frequently used in wind band scores than others. Therefore, to train DNNs for converting piano scores into wind band scores, we need to solve two problems: (i) efficient preparation of pair data and (ii) effective learning of orchestration rules, particularly for infrequently used instruments.

To address these problems, we attempt to create pair data by generating piano scores from existing wind band scores using an automatic piano arrangement method [10]. Then, using the U-Net [11], we estimate a mask that determines whether or not to assign each note of the piano score to an instrument part. To improve the quality of infrequently used instrument parts, we propose refined network architectures to effectively use instrumentation information during training and inference. The results are evaluated quantitatively and analyzed in terms of the ability to reproduce the co-occurrence and exclusion relations among instrument parts.

2 Method

2.1 Problem setup

The input of the proposed method is a piano score consisting of two parts for both hands, and the output is a wind band score with an instrumentation specified by the user. We assume that the user specifies the instrumentation by selecting any number of parts from the maximum instrumentation. Based on several sources of information (e.g. [2]), we define the maximum instrumentation to be consisting of $N = 43$ parts for 28 commonly used instruments (e.g. clarinet in B♭ has three parts), excluding percussion instruments with no pitch. Abbreviated labels for these 43 instrument parts will be listed in Fig. 4C. Thus, the user-specified instrumentation $I = (I_n)_{n=1}^N$ is represented by an N-dimensional binary vector ($I_n = 1$ indicates that instrument part n is used).

Each of the two hand parts, $A_L = (A_L^{\mathrm{o}}, A_L^{\mathrm{a}})$ and $A_R = (A_R^{\mathrm{o}}, A_R^{\mathrm{a}})$, in the piano score and each instrument part $B_n = (B_n^{\mathrm{o}}, B_n^{\mathrm{a}})$ ($n = 1, \ldots, N$) in the wind band score are represented by a pair of binary matrices, $M^{\mathrm{o}} = [M^{\mathrm{o}}(q,t)]$ and $M^{\mathrm{a}} = [M^{\mathrm{a}}(q,t)]$,

representing the onset times and activations for individual pitches, respectively; the number of rows is $Q = 128$, same as the number of pitches in the MIDI format, and the number of columns is the length of the piece with 1/3 of a 16th note as the unit. For example, $B_n^{\mathrm{o}}(q,t) = 1$ indicates that instrument part n has an onset of pitch q at time t, and $B_n^{\mathrm{a}}(q,t) = 1$ indicates that part n is playing pitch q at time t. Thus, the activation matrix B_n^{a} represents the piano roll when graphically visualized, and correspondingly, the onset matrix B_n^{o} the onset positions. The latter is necessary to represent repeated notes of the same pitch without gaps. For the input and output of the U-Net described below, these matrices are segmented by a time length of $T = 192$ corresponding to four measures in 4/4 time (zero padding is applied for fractional segments).

2.2 Preparation of pair data

First, we collected wind band scores in the MusicXML format from a public website (musescore.com). We extracted from the obtained scores only the 43 parts in the maximum instrumentation and used them for the following analysis.

Next, an automatic piano arrangement method [10] was applied to convert these wind band scores to piano scores, thus obtaining pair data with accurately aligned wind band and piano scores that can be used as output and input data for training a DNN. Since this method generates a piano score by selecting some of the notes in an input ensemble score, the notes in the obtained piano score are included in the wind band score. This is a desired property for our method, which generates a wind band score by assigning the notes of the piano score to individual instrument parts.

2.3 Network architecture

We formulate the problem of converting a piano score $A = (A_L, A_R)$ to a wind band score $B = (B_n)_{n=1}^N$ as the estimation of a mask indicating whether or not to assign the notes of the piano score to each instrument part [7]. We use U-Net [11], which has been successfully applied to mask estimation problems such as singing voice separation [12] and piano reduction [13]. U-Net is an encoder-decoder model that performs feature extraction at multiple levels by a stack of convolution and deconvolution layers (Fig. 1). At each level, the features extracted in the encoder side are concatenated to the decoder side. This is expected to enable processing that captures properties at multiple resolutions in the pitch and time directions.

The output of the U-Net is a set of matrices, $\widetilde{B}_n^{\mathrm{o}}$ and $\widetilde{B}_n^{\mathrm{a}}$ ($n = 1, \ldots, N$), each of which corresponds to a binary matrix representing the wind band score. More specifically, for example, the element $\widetilde{B}_n^{\mathrm{o}}(q,t)$ represents the probability that the corresponding element $B_n^{\mathrm{o}}(q,t)$ of the wind band score have a value 1. The following cross-entropy loss function is used for training:

$$\mathcal{L} = -\sum_{n=1}^{N}\sum_{q=1}^{Q}\sum_{t=1}^{T}\Big\{ wB_n^{\mathrm{o}}(q,t)\log\widetilde{B}_n^{\mathrm{o}}(q,t) + [1-B_n^{\mathrm{o}}(q,t)]\log[1-\widetilde{B}_n^{\mathrm{o}}(q,t)] + wB_n^{\mathrm{a}}(q,t)\log\widetilde{B}_n^{\mathrm{a}}(q,t) + [1-B_n^{\mathrm{a}}(q,t)]\log[1-\widetilde{B}_n^{\mathrm{a}}(q,t)]\Big\}.$$

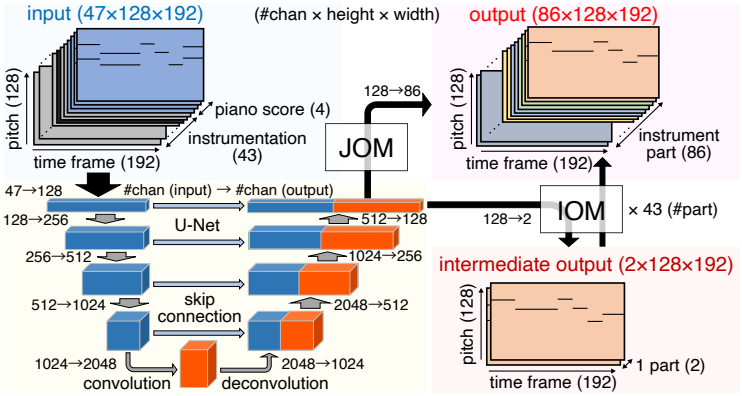

Fig. 1. Proposed network architecture. The output from the U-Net is differently processed in the joint output method (JOM) and individual output method (IOM).

Here, we introduced a weight w for positive samples since the onset and activation matrices are generally sparse in our data ($w = 100$ in our analysis). As we explain below, in the inference step the matrices $\widetilde{B}_n^{\mathrm{o}}$ are subjected to thresholding and other post-processes to obtain a wind band score. We consider the following three network architectures with different formats of input and output for the U-Net.

First, in the simple method (SM), Only the piano score A (4 channels) is used as input, and the maximum instrumentation wind band score B_{all} (86 channels) is obtained as output. During training, the loss function is computed using all instrument parts including those not used in each piece. During inference, only the instrument parts used in the specified instrumentation I are extracted. This method cannot adaptively change the output depending on the specified instrumentation.

Second, in the joint output method (JOM), we add to the input 43 channels of matrices $C_n = [C_n(q,t)]$ representing the instrumentation I (Fig. 1). All elements of matrix C_n are set to one, i.e. $C_n(q,t) = 1$ for all q and t, if instrument part n is used and $C_n(q,t) = 0$ if it is not used. During training, we set $C_n(q,t) = 1$ at all time frames in a piece if instrument part n plays at least one note in the piece. In this way, the network is trained to learn note assignment including rest intervals. The output form and loss function are the same as those of the SM. This method is expected to be more robust to unbalanced frequencies of use of instrument parts in the training data and to learn the dependence on instrumentation, such as balance among instrument parts.

Third, in the individual output method (IOM), the output is the score B_n (2 channels) of each instrument part n, and a single U-Net is used to process all instrument parts. As in the JOM, 43 matrices C_n representing instrumentation I are added to the input, but here all the matrices except for the instrument part to be processed are filled with zeros. During training, a loss function is computed for each instrument part in each piece. During inference, the output B_n for each instrument part n used in the specified instrumentation is combined to generate a wind band score. With this method, it is difficult to adjust the balance among instrument parts according to the instrumentation I, but even more efficient learning of infrequently used instruments is expected.

Method	Octave augmentation	Precision	Recall	F-score
SM		29.2	30.1	28.8
SM	✓	32.3	21.1	25.1
JOM		22.0	2.3	3.8
JOM	✓	19.5	6.2	8.4
IOM		**33.9**	41.3	**36.8**
IOM	✓	31.9	**42.2**	35.9

Table 1. Average accuracies (%) for the simple method (SM), joint output method (JOM), and individual output method (IOM). The highest values are indicated in bold fonts.

In all of the above three methods, the following processes are applied in the inference step. After thresholding the probability estimates of the onset time matrix of each instrument part, the output score is obtained by selecting only the notes contained in the input piano score and imposing the instrument's pitch range and monophonic constraint. We use the pitch ranges written in standard books on orchestration. To impose the monophonic constraint, if more than one onset remain as candidates at a time frame, we choose the one with the largest probability. The duration of each note obtained is determined by referring to the input piano score.

Finally, to generate wind band scores that take advantage of the wide pitch range, it is desired to extend the method and utilize octave-shifted notes from the input piano score. This can be realized in the same mask estimation framework by adding the octave-shifted piano scores, A_L^+, A_R^+, A_L^-, and A_R^-, to the input. For example, $A_L^{\mathrm{o}+}(q,t) = A_L^{\mathrm{o}}(q-12,t)$ and $A_R^{\mathrm{a}-}(q,t) = A_R^{\mathrm{a}}(q+12,t)$. With this octave augmentation, the number of channels in the input increases by 8.

3 Result

From the pair data of 110 pieces obtained as in Sec. 2.2, we used randomly selected 80 pieces as training data and the remaining 30 pieces as test data. As evaluation metrics, we used the precisions, recalls, and F-scores for the output scores calculated individually for all instrumental parts with a criterion of exact match of pitch and onset time. The networks were trained by the AdamW optimizer with a learning rate of 10^{-6} for the SM and JOM and 10^{-7} for the IOM, batch size of 32, and dropout ($p = 0.5$) applied to the first two layers of the decoder. A threshold value of 0.5 was used for inference.

The results in Table 1 show that the IOM outperformed the SM in F-scores, confirming the effectiveness of the method using instrumentation information as input[1]. A comparison of the F-scores for individual instrument parts for the SM and the IOM shows that the latter method significantly improved the F-scores, especially for instrument parts that are used infrequently (Fig. 2). On the other hand, the JOM, which was expected to be the most effective, showed significantly lower accuracies, suggesting that the complex network structure may have reduced the learning efficiency. Therefore, for

[1] See also our demo webpage `https://nabeshinabe.github.io/PianoToBrassBand_nabeoka/demo.html`

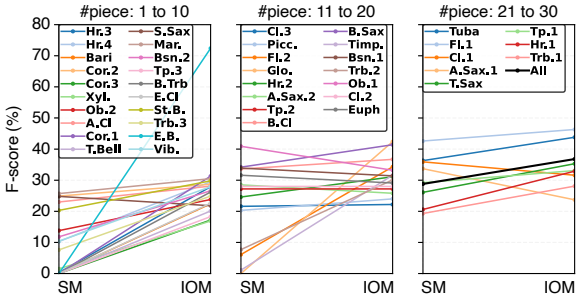

Fig. 2. Partwise F-scores for the SM and IOM (without octave augmentation), shown in three groups according to the number of pieces in the test data in which each instrument part is used.

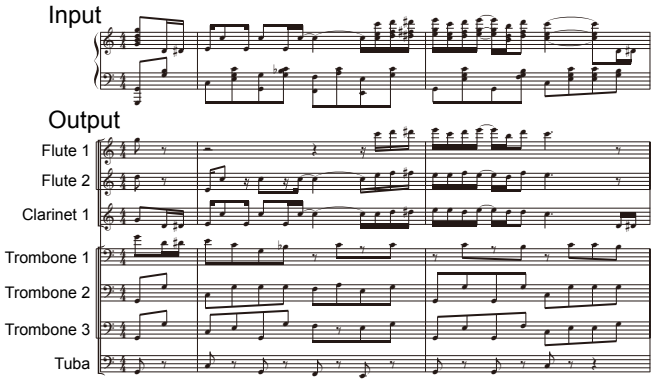

Fig. 3. Wind band score generated by the IOM (without octave augmentation) from Joplin's "The Entertainer." All instruments are notated in concert pitches.

the JOM, refinement of the learning method, for example, by improving the optimization method and increasing the amount of data, should further be investigated. As for the effect of octave augmentation, the metrics changed only slightly for the IOM, with an increased recall and decreased precision and F-score.

Fig. 3 shows an output score with seven instrument parts score obtained by the IOM with piece-level instrumentation information. The input was an existing piano score that was not included in our dataset. The three woodwind parts play the notes in the right hand part of the piano score, and the four brass parts mainly play the notes in the left hand part. The voicing of the chords follows the natural order of the parts within each instrument group. This suggests that the method enables orchestration that captures not only the pitch range of each instrument part, but also the characteristics of the instruments and the mutual relationships among the instrument parts. On the other hand, the IOM has limitations that it cannot adaptively change the roles of the instrument parts according to the specified instrumentation and it cannot assign notes from the piano score to each instrument part without omission. In addition, in the second measure of the 2nd Flute, only some notes of the melody are assigned, which is usually judged as inappropriate. Thus, a proper handling of sequential dependencies of notes, which is necessary for generating smoothly playable arrangements, needs to be improved.

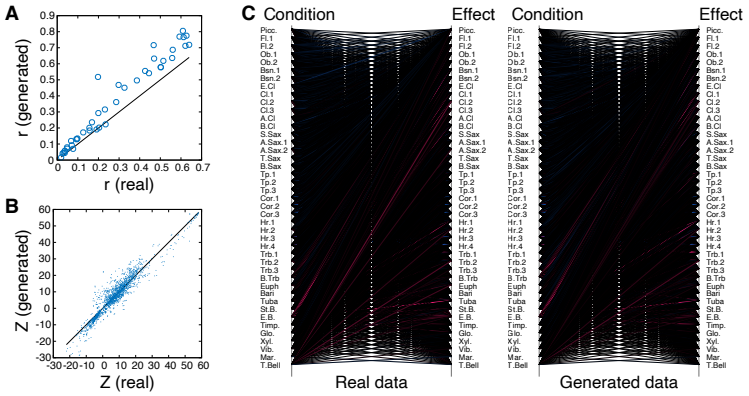

Fig. 4. Correlations of the sounding rates (A) and of the conditional significances (B) between real data and data generated by the IOM (with octave augmentation). C: Conditional significances between all pairs of instrument parts, with positive and negative significances indicated by blue and red lines, respectively (high significances are indicated in dark colors).

To examine the potential of the IOM for learning the interdependence between instrument parts in a larger time scale, we analyzed the sounding rates of individual instrument parts and their correlations. Let $h_{mn} \in \{0, 1\}$ represent whether part n plays at least one note in measure m ($h_{mn} = 1$) or not ($h_{mn} = 0$). We define the sounding rate r_n of part n as $r_n = \sum_m h_{mn}/M$, where M is the total number of measures analyzed. Similarly, we define the simultaneously sounding rate $r_{nn'}$ of parts n and n' as $r_{nn'} = \sum_m h_{mn}h_{mn'}/M$. Then, their correlation can be calculated as $\rho_{nn'} = r_{nn'} - r_n r_{n'}$, which measures the deviation from the independence hypothesis. The statistical significance of this quantity can be measured by the conditional significance $Z(n'|n) := (r_{n'n} - r_{n'}r_n)\sqrt{M}/\sqrt{r_{n'}r_n(1 - r_{n'})}$, where we assumed a binomial process for estimating the statistical error. A positive (negative) value of $Z(n'|n)$ indicates a co-occurrence (exclusion) of part n' conditioned on the presence of part n.

Results in Fig. 4 show that both the sounding rates and simultaneous sounding rates were highly correlated between the real and generated data of wind band scores. This indicates that the U-Net trained by the the IOM learned the co-occurrence and exclusion relations between instrument parts. For example, Fig. 4C indicates a co-occurrence of Soprano Sax and Cornet parts, both of which are expected to be used in large bands but not in small bands, and an exclusion relationship between Bass Trombone and Tuba and between 2nd Bassoon and Electric Bass, which are likely to be a result of substitutability of these instrument parts. These properties of wind band scores were reproduced in the data generated by the IOM. We also conducted the same analysis for the SM but did not observed such clear correlations in the data generated by this method, showing the nontriviality of learning these statistical properties.

4 Discussion

In this paper, we showed the possibility of training DNNs for automatic orchestration of piano scores for wind bands, by generating pair data only from existing wind

band scores using a method for piano arrangement. The experimental results indicated the ability of the proposed U-Net-based method to learn voicing rules and co-occurrence/exclusion relations among instrument parts, and demonstrated the potential for generating partially playable wind band scores in user-specified instrumentations.

A number of challenges remain for the generation of wind band scores suitable for actual performance. Increasing training data and further refinements of network architectures should be attempted to successfully train the JOM or similar networks that can adaptively change the roles of instrument parts according to the specified instrumentation. To suppress note sequences with unnatural leap motions, rhythms, etc. in the outputs that are difficult to play, use of autoregressive networks, such as a long short-term memory (LSTM) network and Transformer, is expected to be effective. More thorough evaluations by arrangement experts and through actual performance tests of the output results should be conducted in the future.

References

1. Berlioz, H., Strauss, R.: Treatise on Instrumentation (transl. by T. Front). Dover Publications, New York (1991)
2. Newton, B.: Band Orchestration: Volume 1: Introduction and Orchestration. CreateSpace Independent Publishing Platform (2016)
3. Maekawa, H., et al.: On machine arrangement for smaller wind-orchestras based on scores for standard wind-orchestras. In: Proc. Int. Conf. on Music Perception and Cognition, pp. 278–283. Bononia University Press, Bononia (2006)
4. Cella, C.E.: Orchidea: a comprehensive framework for target-based computer-assisted dynamic orchestration. Journal of New Music Research 51(1), 40–68 (2022)
5. Roberts, A., et al.: A hierarchical latent vector model for learning long-term structure in music. In: Porc. Int. Conf. on Machine Learning, pp. 4364–4373. ICML, Stockholm (2018)
6. Huang, Y.S., Yang, Y.H.: Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In: Proc. ACM Int. Conf. on Multimedia, pp. 1180–1188. ACM, Seattle (2020)
7. Crestel, L., Esling, P.: Live Orchestral Piano, a system for real-time orchestral music generation. In: Proc. Int. Sound and Music Computing Conf., pp. 434–442. Aalto University, Espoo (2017)
8. Crestel, L., et al.: A database linking piano and orchestral MIDI scores with application to automatic projective orchestration. In: Proc. Int. Society for Music Information Retrieval Conf., pp. 592–598. ISMIR, Suzhou (2017)
9. Liu, J., et al.: Symphony generation with permutation invariant language model. In: Proc. Int. Society for Music Information Retrieval Conf., pp. 551–558. ISMIR, Bengaluru (2022)
10. Nakamura, E., Yoshii, K.: Statistical piano reduction controlling performance difficulty. APSIPA Transactions on Signal and Information Processing 7(e13), 1–12 (2018)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer, Switzerland (2015)
12. Jansson, A., et al.: Singing voice separation with deep U-Net convolutional networks. In: Proc. Int. Society for Music Information Retrieval Conf., pp. 745–751. ISMIR, Suzhou (2017)
13. Terao, M., et al.: Difficulty-Aware Neural Band-to-Piano Score Arrangement based on Note- and Statistic-Level Criteria. In: Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 196–200. IEEE, Singapore (2022)