Audio-Based Automatic Generation of a Piano Reduction Score by Considering the Musical Structure

Hirofumi Takamori[1(✉)], Takayuki Nakatsuka[1(✉)], Satoru Fukayama[2(✉)], Masataka Goto[2(✉)], and Shigeo Morishima[3(✉)]

[1] Department of Pure and Applied Physics, Waseda University, Tokyo, Japan
{tkmrkc1290,t59nakatsuka}@gmail.com
[2] National Institute of Advanced Industrial Science and Technology (AIST), Ibaraki, Japan
{s.fukayama,m.goto}@aist.go.jp
[3] Waseda Research Institute for Science and Engineering, Tokyo, Japan
shigeo@waseda.jp

Abstract. This study describes a method that automatically generates a piano reduction score from the audio recordings of popular music while considering the musical structure. The generated score comprises both right- and left-hand piano parts, which reflect the melodies, chords, and rhythms extracted from the original audio signals. Generating such a reduction score from an audio recording is challenging because automatic music transcription is still considered to be inefficient when the input contains sounds from various instruments. Reflecting the long-term correlation structure behind similar repetitive bars is also challenging; further, previous methods have independently generated each bar. Our approach addresses the aforementioned issues by integrating musical analysis, especially structural analysis, with music generation. Our method extracts rhythmic features as well as melodies and chords from the input audio recording and reflects them in the score. To consider the long-term correlation between bars, we use similarity matrices, created for several acoustical features, as constraints. We further conduct a multivariate regression analysis to determine the acoustical features that represent the most valuable constraints for generating a musical structure. We have generated piano scores using our method and have observed that we can produce scores that differently balance between the ability to achieve rhythmic characteristics and the ability to obtain musical structures.

Keywords: Piano reduction · Multivariate regression analysis
Musical structure · Acoustic feature · Self-similarity matrix

Supported by JST ACCEL, Japan (grant no. JPMJAC1602).

I. Kompatsiaris et al. (Eds.): MMM 2019, LNCS 11296, pp. 169–181, 2019.
https://doi.org/10.1007/978-3-030-05716-9_14

1 Introduction

One way to enjoy music is by playing an instrument yourself, which brings a different joy as compared to merely listening to music. The piano is the instrument that can simultaneously play multiple roles, including the melody line, harmony, and rhythm. For genres such as popular music, comprising a main vocal melody and an accompaniment played using various instruments, the piano is a suitable instrument for an individual to play his or her favorite songs. Herein, we focus on generating a piano reduction score for popular music.

Piano scores for several popular songs have been written by professional music arrangers. However, it is often necessary for a player to create a piano score from scratch because there is no guarantees that the desired song will be available as a piano score. Creating a piano reduction requires carefully working out how all parts of the original music can be expressed using a playable piano score. To address this issue, our method automatically generates a piano reduction score. It considers the audio signals from a pop song as the input and outputs a score suitable a piano.

The goal of this study is to automatically generate piano scores from audio signals. To achieve this goal, we adopt the approach proposed in paper [1]. The previous approach generates piano scores for each bar based on musical elements; melody, rhythm, chords, and number of notes. These elements are obtained from original scores. However, directly adopting this score-based method in audio-based applications is problematic since audio-based feature extraction is not always accurate, and it causes a lack of overall coherence. Hence, structural considerations are necessary for audio-based piano reduction.

In this study, we present a piano reduction method while considering the structure of the music. Our piano reduction method follows three stages: (a) analysis of the musical structure; (b) determination of the structure of the audio signal; and (c) score composition. Figure 1 shows the schematic of the proposed method. The main contribution of our study is twofold: First, we have generated piano scores that reflect both the rhythmic and structural features of the input audio signals. Second, we have determined self-similarity matrices (SSMs) for seven different acoustical features, which represent the structure of the piano performance. As the limitations of our method, we treat only popular songs in quadruple time, and the minimum resolution of generating the score is limited to a semiquaver note.

2 Related Work

Several studies have attempted to generate piano scores from original scores involving the usage of multiple instruments. Fujita et al. [2] generated a piano score from an ensemble score by extracting the melody and bass part and using them to develop the piano reduction. Chiu et al. [3] have considered five roles of the piano in music, which are the lead, foundation, rhythm, pad, and fill that were originally proposed by Owsinski [4]. By analyzing an original score, Chiu et al.

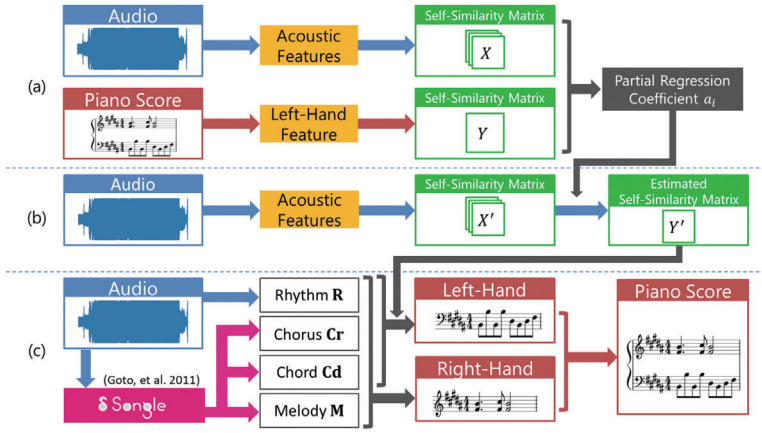

Fig. 1. Overview of our proposed method. (a) Using multivariate regression analysis, we analyze the correlation between the SSMs of the acoustical features and the SSM of the left-hand part of a manually arranged score. (b) We determine the SSM for a sample song using partial regression coefficients obtained in (a). (c) We generate a piano score with both the right- and left-hand parts using the extracted musical elements and structural feature estimated as SSM in (b).

associated each phrase in the score with a weighted importance value. They proposed a phrase-selection algorithm that maximized the importance value while considering the score's playability. Nakamura et al. [5] generated a piano reduction from an ensemble score using a fingering model. They focused both on the preservation of the sounds and on playability as constraints, where playability can be separately controlled by the respective difficulty parameters observed in case of the right- and left-hand parts. A common thread in these previous studies has been the reduction and selection of notes from an original score using either the original notes directly or through octave shifts. These are valid approaches that preserve the original impression of the music without generating any dissonance.

Methods exist that do not completely transcribe a score from an audio signal, but that can extract musical elements from an audio signal to generate an arrangement. For example, Percival et al. [6] have presented Song2Quartet, a system for generating string-quartet versions of popular music from audio recordings without requiring pitch determinations for all parts. This method can extract musical elements from an original piece, including the melodies, rhythms, chords, and number of notes. We emphasize the consideration of the constraints involved in piano reduction.

3 Piano Reduction of Popular Music

Melody, harmony, rhythm, timbre, and texture are deemed essential elements of music [7], and play important roles in musical expression. To preserve the original

song's impression, focusing on these elements is essential. Since expressing the timbre of different instruments using only the piano is difficult, we focus on preserving the remaining elements, including melody, harmony, rhythm, and texture, in the piano reduction. In particular, we note the following points:

- The melody is always the highest pitch.
- Each chord in the output score matches with the original one.
- The output score represents the original rhythm.
- The output score exhibits a contrast between the verse and the chorus.

A previous study [1] established the value of the aforementioned requirements. Along with the aforementioned requirements, we consider a long-term correlation structure that is observed especially in the piano scores of popular music. Popular music generally contains structures, such as a verse, a chorus, and a bridge, and some of these musical sections are repeated in a single song [8]. Hence, it is important to reflect these repetitive structures in the generated piano scores. Thus, we impose an additional requirement to express the structural features of popular music as follows:

- The left-hand part should exhibit similar accompaniment patterns within the same section.

In this study, we perform piano reduction by considering the five aforementioned requirements.

4 Analysis of the Structure of the Music

In this section, we explain the analysis stage illustrated in Fig. 1(a) and outlined in Fig. 2. We have initially prepared a dataset containing 27 popular songs that include both the audio data and the corresponding piano scores. These audio data are acquired from the Internet, and these piano scores are manually produced.[1,2,3]

4.1 Feature Extraction

As acoustic features, we use chromagrams, Mel-frequency cepstrum coefficients (MFCCs), onsets, root-mean-square (RMS) energy, spectral centroid, spectral flatness, and zero-crossing rates (ZCRs). The audio signals are monaural and their sampling frequencies are 44.1 kHz. The window length during analysis is 1024, with an overlap of 256. We also set the number of channels of the Mel-scale filter bank at 20, and we use the 12 low dimensions. Especially for onsets detection, the methodology is inspired by Böck et al. [10].

[1] Bokaro Kamikyoku Daishugo Best 30, Depuro MP, Japan (2016).
[2] Jokyu Piano Grade Bokaro Meikyoku Piano Solo Concert, Depuro MP, Japan (2015).
[3] Print Score, https://www.print-gakufu.com/.

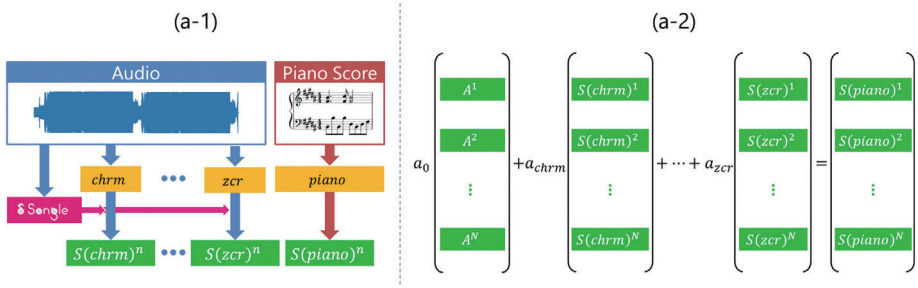

Fig. 2. Overview of the analysis of the musical structure. (a-1) Feature extraction from an audio recording and a score. We use *Songle* [9] to acquire the beats data. (a-2) Multivariate regression analysis of all songs in the dataset. The quantity, $\mathbf{A}$, represents a matrix whose complete list of entries are 1. The quantity, $S(\cdot)$, represents the SSM of acoustic features or of a piano feature, and a_i represents a partial regression coefficient.

In this study, we extract the aforementioned seven features for each bar. *Songle* [9], a web service for active music listening, is used to get the start time of each bar in a song. This allows us to work out how the times (frames) of the audio signal correspond to the bars. Each feature of the m^{th} bar can be represented as follows: $\mathbf{chr}_m \in \mathbb{R}^{12\times16}$ for the chromagram, $\mathbf{mfcc}_m \in \mathbb{R}^{12\times16}$ for the MFCCs, $\mathbf{onset}_m \in \mathbb{Z}^{1\times16}$ for the onsets, $\mathbf{rms}_m \in \mathbb{R}^{1\times16}$ for the RMS energy, $\mathbf{cent}_m \in \mathbb{R}^{1\times16}$ for the spectral centroid, $\mathbf{flat}_m \in \mathbb{R}^{1\times16}$ for the spectral flatness, and $\mathbf{zcr}_m \in \mathbb{R}^{1\times16}$ for the ZCR, respectively. Row-wise represents the time direction of a bar which is divided into segments with length of 16^{th} note. For example, the first column represents the one-dimensional feature for the first 16th note in a bar. Acoustic features, excluding the $\mathbf{onset}_m$, are projected onto each segment with length of a 16^{th} note by considering the mean between the current and subsequent beats. The j^{th} column value of $\mathbf{onset}_m$ is set to unity if there is a peak between the current and subsequent beats or is set to zero if there is none.

Considering the features of a piano score, we extract the positions and numbers of the notes from the left-hand part and described them for each bar. We denote the m^{th} bar's feature of a piano score by $\mathbf{piano}_m \in \mathbb{Z}^{1\times16}$. Row-wise again represents the time direction. The values of the j^{th} column of $\mathbf{piano}_m$ are the numbers of notes positioned at the j^{th} beat.

4.2 Multivariate Regression Analysis of the Self-Similarity Matrices

SSMs are calculated for both acoustic features and the features of a piano score by the procedures described in Sec. 4.1. The SSM indicates the structural similarity between the bars included in a song. For the feature sequence of a song, $\mathbf{f} = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_M\}$, the SSM, $\mathcal{S}(\mathbf{f}) \in \mathbb{R}^{M\times M}$, can be defined as follows:

$$\mathcal{S}(\mathbf{f}) = [s_{ij}] = \begin{pmatrix} 1 & s_{12} & \dots & s_{1M} \\ s_{21} & 1 & \dots & s_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{M1} & s_{M2} & \dots & 1 \end{pmatrix} \tag{1}$$

where s_{ij} represents the similarity between a feature's i^{th} bar, $\mathbf{f}_i$, and the j^{th} bar, $\mathbf{f}_j$. The similarity, s_{ij}, can be given as follows:

$$s_{ij} = \frac{1}{1 + d_{ij}} \tag{2}$$

$$d_{ij} = ||\mathbf{f}_i - \mathbf{f}_j|| \tag{3}$$

where d_{ij} represents the distance between $\mathbf{f}_i$ and $\mathbf{f}_j$, which can be obtained by computing the Frobenius norm. The values of s_{ij} lie in the range $(0, 1]$, where $s_{ij} = 1$ indicates that the two features being compared are identical. Conversely, s_{ij} exhibits a low value if the two features being compared are unidentical.

We further perform multivariate regression analysis on the SSM. We assign the explanation variable, $\mathbf{X}$, and the objective variable, $\mathbf{Y}$, as follows:

$$\mathbf{X}_f = \{\mathcal{S}(\mathbf{f})^1, \mathcal{S}(\mathbf{f})^2, \cdots, \mathcal{S}(\mathbf{f})^N\}$$
$$\mathbf{Y}_{piano} = \{\mathcal{S}(\mathbf{piano})^1, \mathcal{S}(\mathbf{piano})^2, \cdots, \mathcal{S}(\mathbf{piano})^N\},$$

where $\mathbf{f}$ represents a sequence of each acoustic feature. The quantity, $\mathcal{S}(\cdot)^n$, represents the SSM of the dataset's n^{th} song. We denote the formula to perform multivariate regression analysis as follows:

$$a_0\mathbf{A} + \sum_{\gamma} a_{f_\gamma}\mathbf{X}_{f_\gamma} = \mathbf{Y}_{piano} \tag{4}$$

where a_0 represents the intercept and a_f denotes the acoustic features' partial regression coefficient. The quantity, $\mathbf{A}$, denotes a list of the matrices whose complete list of entries include 1, and it is introduced to match the matrix dimensions. The subscript, f_γ, represents one of the seven acoustic features used in this study. Eq. (4) is schematically depicted in Fig. 2(a-2). According to Eq. (4), we determine the intercept a_0 and the partial regression coefficients a_f.

5 Structural Segmentation

In this section, we explain the approach used to determine the structure of a piano score from the audio signals (Fig. 1(b)). The structure of a piano score is determined by segmenting SSM, which is estimated from acoustic features by using Eq. (4) with the partial regression coefficient obtained by multivariate regression described in Sect. 4.2. To segment a song into several musical sections, we adopt novelty detection [11]. We perform novelty detection by identifying the peaks of the novelty scores, obtained by multiplying the checkerboard kernel, $\mathbf{C}$, along the SSM's diagonal.

$$\mathbf{C} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \tag{5}$$

In our methodology, we calculate five kinds of novelty scores having checkerboard kernel sizes of (2×2), (4×4), (6×6), (8×8), and (10×10), respectively. We further consider the mean of the five novelty scores and normalize them to $[0, 1]$ range. Additionally, we perform peak detection for the novelty scores by introducing a second-order differential threshold, $th = 0.00, -0.05, -0.07, -0.10, -0.15$. The peak position is set to one at which the first-order differentiation turns from positive to negative and at which the second-order differentiation is less than th. We further obtain the musical structures' boundaries according to the peaks of the novelty scores. The musical structure is represented by lists of bars that mark the boundaries between various segments. Bars located between the acquired boundaries are considered to belong to the same segment.

6 Score Composition from Audio Signals

In this section, we explain the architecture for generating a piano score from the audio signals (Fig. 1(c)). First, we focus on the accompaniment database constructed beforehand. Further, we discuss the extraction of musical elements from the audio signals. We obtain the chorus, chord, and melody using *Songle* [9], while rhythm is obtained by detecting the onsets of spectral flux. Finally, we generate a piano score for both left- and right-hand parts based on the extracted elements.

6.1 Accompaniment Database

We construct the accompaniment database, **DB**, based on the existing piano scores [1]. The accompaniment database comprises accompaniment matrices. An accompaniment matrix represents a bar of the left-hand part as an 88×16 matrix, where 88 is the number of piano keys and 16 is set to match the length of a semiquaver. The matrix is generated after being transposed so that the root becomes C. In case of the matrix elements, the note value is stored in the places at which the note exists; zero is stored if there is no note. This allows the system to record the relative pitch transition and the rhythm of the original piano score. In this study, $\mathbf{DB}_n \in \mathbb{Z}^{88\times16}$ denotes the n^{th} accompaniment matrix; it contains the 16-dimensional vector, $\mathbf{DBR}_n$, that represents the rhythm. When a non-zero value is stored in the j^{th} column of $\mathbf{DB}_n$, the value 1 is stored in the j^{th} element of $\mathbf{DBR}_n$. If only zeros are stored in the j^{th} column of **DB**, the value zero is also stored in the j^{th} element of $\mathbf{DBR}_n$.

6.2 Extraction of Musical Elements from Audio Signals

We extract musical elements, including the melody **M**, chord **Cd**, chorus **Cr**, and rhythm **R**. We acquire **M**, **Cd**, and **Cr** from *Songle* [9], while we extract **R** by onset detection of the spectral flux, as described in Sect. 4.1. We obtain the following analysis results for each element from *Songle Widget* [4]:

[4] http://widget.songle.jp/.

Beat: index, start time, position
Chord: index, start time, duration, chord name
Melody: index, start time, duration, MIDI note number
Chorus: index, start time, duration

The index denotes the number of beats, chords, or notes of melody observed from the beginning of a song; start time represents the time at which each event starts; the position specifies the number of beats in each bar; duration denotes the length of the event; chord name shows the root note and the chord type; and the MIDI note number is the value that indicates the pitch. From these information, the melody and the chords can be described for each bar, and we also obtain the bar number, which is in the chorus section. For $\mathbf{M}$, $\mathbf{Cd}$, $\mathbf{Cr}$ and $\mathbf{R}$, the subscript, m, indicates the m^{th} bar of the score. The quantity, $\mathbf{M}_m \in \mathbb{Z}^{88\times 16}$, provides the pitch, note values, and position in the score for each bar. The number of rows corresponds to the number of keys on the keyboard, whereas the number of columns corresponds to the time resolution (16^{th} note). The chord notes are represented in $\mathbf{Cd}_m$ for each beat as a set of MIDI note numbers. For Cr_m, the value is set to unity if the m^{th} bar is in the chorus and is set to zero otherwise. The rhythm, $\mathbf{R} \in \mathbb{Z}^{1\times 16}$, is acquired in the same manner as the onset, as explained in Sect. 4.1.

6.3 Generation of the Right-Hand Part

We allocate the melody, $\mathbf{M}$, to the right-hand part, $\mathbf{RH}$, as follows:

$$\mathbf{RH}_m = \begin{cases} \mathbf{Add}(\mathbf{M}_m, \mathbf{R}_m, \mathbf{Cd}_m) & Cr_m = 1 \\ \mathbf{M}_m & \text{otherwise} \end{cases} \tag{6}$$

The quantity, $\mathbf{RH}_m \in \mathbb{Z}^{88\times 16}$, represents the m^{th} bar of $\mathbf{RH}$. If the m^{th} bar is in a chorus section, $\mathbf{Add}_m$ attaches additional chord notes to each note in the melody, $\mathbf{M}_m$, where a component of the rhythm, $\mathbf{R}_m$, is unity. The chord notes at each beat are obtained from $\mathbf{Cd}_m$ and are considered to be lower in pitch than the melody and more than four semitones away.

6.4 Generation of the Left-Hand Part

The left-hand part, $\mathbf{LH}$, of the piano score can be generated by selecting from the accompaniment database, $\mathbf{DB}$. First, we select an accompaniment for which the rhythm is similar to that of the audio signal, $\mathbf{R}$, for each bar. This accompaniment list is defined as $\mathbf{LH}'$, which represents the accompaniment list by considering only the rhythm. Further, we reflect the features of the musical structure, as determined in Sec. 5. We reduce the kinds of accompaniments that appeared in the same musical section so that they exhibit similar rhythmic patterns. The result of this reduction is designated as $\mathbf{LH}$. The m^{th} bar of $\mathbf{LH}'$ is denoted by $\mathbf{LH}'_m$. We define the parameter, λ, to be the number of kinds of

accompaniments in the same musical section. This process is described by the following formula:

$$\mathbf{LH} = FuncS(\mathbf{LH}', \lambda, th) \tag{7}$$

$$\mathbf{LH}'_m = \mathrm{a}rgmin\ \ CostR(\mathbf{DB}, \mathbf{R}_m) \tag{8}$$

$$CostR(\mathbf{DB}, \mathbf{R}_m) = \sum_n ||\mathbf{DBR}_n - \mathbf{R}_m|| \tag{9}$$

Here, $FuncS(\cdot)$ reduces the kinds of accompaniments by changing the accompaniment patterns with a low appearance frequency to form accompaniments with a high appearance frequency. This reduction continues until the number of kinds of accompaniments became less than λ for each musical section. The value, $\lambda = 1$, indicates that only one kind of accompaniment is selected in each musical section, and $\lambda = \infty$ indicates that the musical structure is not considered. $\mathbf{LH}'$ is a candidate accompaniment list selected on the basis of $CostR(\cdot)$, where $CostR$ is introduced to select accompaniments with high similarities in audio rhythms. Finally, to ensure that the sound of $\mathbf{LH}_m$ reflects $\mathbf{Cd}_m$ we shift $\mathbf{LH}_m$ to the nearest chord notes for each bar.

7 Results and Evaluation

As the result of the multivariate regression analysis described in Sect. 4, the partial regression coefficients and t-values of them are presented in Table 1.

Table 1. The result of multivariate regression analysis.

	a_0	a_{chr}	a_{mfcc}	a_{onset}	a_{cent}	a_{flat}	a_{rms}	a_{zcr}
coefs	−0.0373	0.4439	−0.0329	0.1403	0.2224	0.3153	0.0494	−0.1175
t-value	−8.974	74.50	−8.863	74.88	111.3	87.47	27.87	−34.95

Here, a_0 represents the intercept and a_f represents each acoustic feature's partial regression coefficient. The t-value in linear regression analysis is commonly considered to be a value that indicates the similarity of the explanation value with the objective value. The t-value can be derived by dividing a coefficient with its standard error. A large absolute t-value indicates that the explanation value exhibits large effectiveness in determining the objective. We also calculate p-values and an adjusted coefficient of determination R^2_{adj}. The p-value represents the probable significance of the explanation value. Generally, an explanation value is effective when the p-value is lower than 0.05. The results of all the p-values of the coefficients are lower than the order of 10^{-18}. R^2_{adj} is an indicator of multivariate regression analysis' accuracy, and can be defined as follows:

$$R^2_{adj} \equiv 1 - \frac{\sum_i (y_i - y'_i)^2/(N-p-1)}{\sum_i (y_i - \overline{y}_i)^2/(N-1)} \tag{10}$$

where y is the true data; $\overline{y}$ is the mean of y; y' is the predicted data; N is the sample size; and p is the number of explanation values. Unity is achieved when there is no residual relative to the predicted data, and unity decreases as the residual increases. The obtained R^2_{adj} in our model was 0.1590.

We output the piano score for each of the 27 popular songs included in the dataset. To verify the effectiveness of our method, we conduct leave-one-out cross-validation for the SSMs of all the songs in the dataset. Each SSM is calculated from a left-hand part of each score.

First, we select one song as test data and generate a piano score of this song using the partial regression coefficients derived from the remaining 26 songs. Finally, we obtain the residuals by calculating the Frobenius norm of the difference between the SSM of the generated and manually created piano scores. We calculate the residuals for each of the seven different values of λ, which denotes the number of types of accompaniments in one musical section, and for five different block thresholds, th. Table 2 presents the result of this cross-validation.

Table 2. The result of leave-one-out cross-validation. ($\times 10^{-3}$)

threshold th	The number of kinds of accompaniments λ						
	1	2	3	4	5	6	∞
0.00	3.550	3.630	3.674	3.685	3.714	3.730	3.760
−0.05	3.535	3.612	3.684	3.708	3.738	3.742	3.760
−0.07	3.277	**3.357**	3.452	3.522	3.591	3.647	3.760
−0.10	**3.257**	3.451	**3.413**	**3.430**	**3.460**	**3.506**	3.760
−0.15	3.311	3.503	3.538	3.535	3.564	3.560	3.760

All the values are normalized by dividing with the size of the matrix. We calculate the residuals for all songs in the dataset and estimate the mean value. The values represent how closely the generated scores' accompaniments resemble those of manually-produced scores in their structure. Figure 3 depicts the variation of the SSM structure with λ for the cases in which $th = -0.07, -0.10$. We selected three songs[5,6,7] where the SSMs of these songs clearly changed by trying different values for λ. We also test our method using one song from the RWC Music Database [12] (RWC-MDB-P-2001 No. 7).[8]

8 Discussion

Table 1 shows that the t-values of the spectral centroid, spectral flatness, onset, and chromagram exhibit comparably high values, indicating that these acoustic

[5] *Mosaic Roll* (DECO*27), https://www.nicovideo.jp/watch/sm11398357.
[6] *Ghost Rule* (DECO*27), https://www.nicovideo.jp/watch/sm27965309.
[7] *Irohauta* (Ginsaku), https://piapro.jp/t/0D18/20100223020519.
[8] The generated results is available at https://youtu.be/Yx9c0LnEyyE.

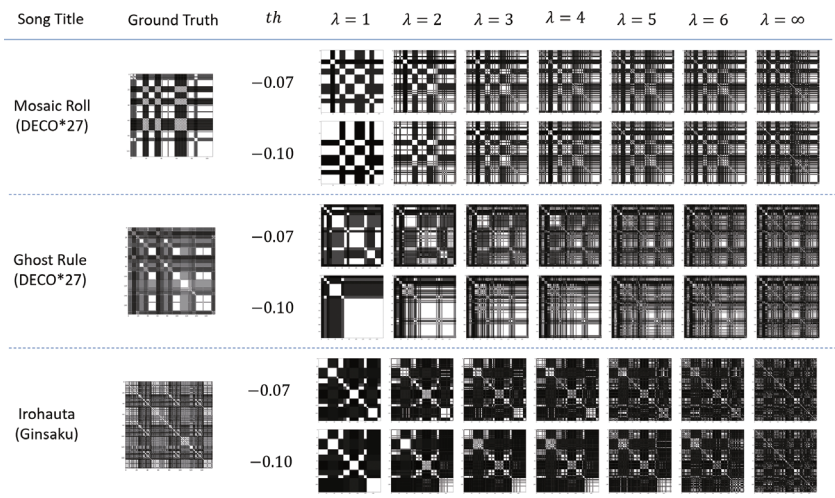

Fig. 3. Comparison of the SSMs with the ground truth about three songs.

features effectively determine the structure of the piano performance. The spectral centroid, spectral flatness, and chromagram are features related to the pitch of the sound and harmony. Hence, the registers, melodies, and chords are deemed important for structural determination. The onset is a rhythmic feature; hence, focusing on the sounds that correspond to beats or rhythm is also important.

Conversely, the t-values for MFCC, RMS, and ZCR exhibit comparably low values, indicating that the features related to timbre and texture are not as effective as the aforementioned harmonic and rhythmic features. Obviously, timbre and texture change during a song. However, repetitive structures of these features are not as well-correlated with the structure of the piano performance. In this study, the left-hand part of the piano score represents the structure of the piano score. We conclude that the timbre and texture of the sounds are expressed in the piano score by other elements such as musical symbols and the number of notes. Hence, the structure of the piano score generated by our method is not sufficient to express the structure of timbre and texture, resulting in low t-values for MFCC, RMS, and ZCR.

R^2_{adj} is below unity, which is the maximum value of R^2_{adj}. Several possible variations of piano scores exist for a given song depending on the arranger. Therefore, R^2_{adj} inevitably exhibits low values owing to these fluctuations. However, a determination of the musical structure from the acoustic features is possible to some extent because R^2_{adj} exhibits a positive value.

Table 2 shows that the SSMs of the generated piano scores exhibit low residuals when λ is small and when th is -0.07 and -0.10. This indicates that segmenting the music with moderate roughness and selecting fewer kinds of accompaniments in the segmented sections produces results closer to the ground truth. However, in a piano score by an arranger, several kinds of accompaniments are

often included in the same section. For instance, several kinds of accompaniments may be alternatively observed, or distinguishing accompaniment appears just before the next phrase. Therefore, it is important to consider short-term structure and to focus on musical transitions, in order to ensure that the generated result is close to the original piano score.

9 Conclusions and Future Work

Herein, we have proposed method of a piano reduction from audio signals by considering the structure of the music. We output several patterns of piano scores and calculated the SSMs to verify the relation between the musical structures of the audio signals and the piano scores. The results of the multivariate regression analysis show that spectral centroid, spectral flatness, onset, and chromagram were valuable features for determining the musical structure. The results also show that giving consideration to music structure ensures that the generated piano structure will be close to one written by an arranger. In future studies, we aim to augment the music database and reselect acoustic features that are more valuable for generating a piano reduction score.

References

1. Takamori, H., Sato, H., Nakatsuka, T., Morishima, S.: Automatic arranging musical score for piano using important musical elements. In: Proceedings of the 14th Sound and Music Computing Conference, Aalto, Finland, pp. 35–41 (2017)
2. Fujita, K., Oono, H., Inazumi, H.: A proposal for piano score generation that considers proficiency from multiple part. In: IPSJ SIG Technical reports, pp. 47–52 (2008)
3. Chiu, S., Shan, M., Huang, J.: Automatic system for the arrangement of piano reductions. In: Proceedings of the 11th IEEE International Symposium on Multimedia, pp. 459–464 (2009)
4. Owsinski, B.: The Mixing Engineer's Handbook. Thomson Course Technology (1999)
5. Nakamura, E., Sagayama, S.: Automatic piano reduction from ensemble scores based on merged-output Hidden Markov model. In: Proceedings of the 41st International Computer Music Conference (ICMC), pp. 298–305 (2015)
6. Percival, G., Fukayama, S., Goto, M.: Song2Quartet: a system for generating string quartet cover songs from polyphonic audio of popular music. In: Proceedings of the International Symposium Music Information Retrieval, pp. 114–120 (2015)
7. Schmidt-Jones, C.: The Basic Elements of Music (2014). Lulu.com
8. Doll, C.: Rockin' out: expressive modulation in verse-chorus form. J. Soc. Music Theory **17**(3), 1–10 (2011)

9. Goto, M., Yoshii, K., Fujihara, H., Mauch, M., Nakano, T.: Songle: a web service for active music listening improved by user contributions. In: Proceedings of the International Symposium on Music Information Retrieval, pp. 311–316 (2011)
10. Böck, S., Florian, K., Markus, S.: Evaluating the online capabilities of onset detection methods. In: Proceedings of the International Symposium on Music Information Retrieval, pp. 49–54 (2012)
11. Jonathan, F.: Automatic audio segmentation using a measure of audio novelty. In: Proceedings of the 2000 IEEE International Conference on Multimedia and Expo, vol. 1, pp. 452–455 (2000)
12. Goto, M.: Development of the RWC music database. In: Proceedings of the 18th International Congress on Acoustics (ICA 2004), vol. 1, pp. 553–556 (2004)