Evaluation Score Prediction for Japanese Songs Based on Melody Fitness to Lyrics

Sosuke Nishimura* and Eita Nakamura*†
* Kyushu University, Japan
E-mail: nishimura.sosuke.638@s.kyushu-u.ac.jp, nakamura@inf.kyushu-u.ac.jp
† PTNA Institute of Music Research, Japan

***Abstract*—This study investigates a method for predicting listener evaluation scores of melodies in relation to lyrics, with the aim of improving the quality of songs generated by automatic music composition systems. To address the limitations of supervised learning approaches that rely on large-scale human evaluation data, we propose a model that represents the implicit evaluation process involved in music creation. This enables the estimation of melody fitness to lyrics using only training data of melodies with aligned lyrics. As a specific focus, we study Japanese songs, in which the correspondence between musical elements and lyrical features, such as accent patterns and word boundaries, has long been discussed. By employing deep generative models of melodies conditioned on lyrical features, we analyze the relevance of various lyrical attributes to melody fitness and evaluate the effectiveness of the proposed approach. Experimental results show that the proposed method can predict listener evaluation scores with high accuracy, achieving a correlation coefficient of 0.653 with human ratings.**

I. Introduction

Automatic music generation has emerged as an active area of research, with a wide range of methods developed based on machine learning. Notable examples include techniques using generative adversarial networks (GANs) [1], [2], variational autoencoders (VAEs) [3], diffusion models [4], and Transformer architectures [5]. In addition to symbolic generation, methods that directly synthesize audio waveforms have also been proposed [6], [7]. While machine learning techniques have made it possible to generate music that closely imitates existing compositions, evaluating the artistic value of the generated output and capturing listeners' preferences remain major challenges. Musical preferences play a significant role particularly in generating novel musical content, as exemplified in studies based on genetic algorithms using user evaluations as the fitness function [8], [9]. Accordingly, addressing the problem of automatic music evaluation is crucial for the development of systems capable of generating high-quality music across a broad range of styles.

Recent studies have explored the construction of machine learning models that predict human evaluation scores based on large-scale subjective data [10]. However, collecting such evaluation data is costly, posing significant limitations on the applicability of fully data-driven approaches across diverse musical genres and styles. Furthermore, musical evaluation is influenced not only by musical elements such as pitch, rhythm, and harmony, but also by extra contextual factors, including lyrics, accompanying visuals, and choreography [11]–[13]. This interplay of multiple modalities makes it particularly difficult to disentangle the specific components factors that affect listener preferences.

The purpose of this study is to develop a method for predicting human evaluation scores for singing melodies, with a particular focus on their dependence on lyrics. We develop a method for predicting evaluation scores that does not rely on a large amount of human evaluation data. Our approach is grounded in the assumption that published works have undergone an implicit internal evaluation by their creators. Building on this assumption, we propose an artwork selection model that represents this underlying music creation process. We then formulate machine learning methods for estimating the fitness of melodies to given lyrics using observed musical data, and for predicting average listener evaluation scores based on these fitness estimates.

As a concrete focus, we investigate Japanese songs, in which the correspondence between lyrical features and melodic elements is considered to influence evaluation scores through their effects on listenability and singability. Given that the Japanese language exhibits pitch accent, much work has been made to investigate the relationship between melodic pitch and the accent patterns of lyrics [14]–[16]. Melodic elements include rhythm as well as pitch, and potential lyrical features that may influence them include not only accent patterns but also word boundaries and semantic content. To capture the complex interplay among these components, we develop machine learning techniques to model their relationships and to analyze how various lyrical features contribute to the evaluation scores of melodies.

II. Data Collection and Analysis

A. Background: Japanese Lyrics and Melody

The Japanese language exhibits a pitch accent system, in which each syllable is assigned either a high or low pitch. For example, the words "rain" and "candy" are both written as "a-me" in Japanese but are distinguished by their pitch patterns: the former is pronounced with a "high-low" accent pattern, while the latter with a "low-high" pattern. Given this linguistic characteristic, music theorists have proposed a principle of melody composition in which the direction of pitch transitions

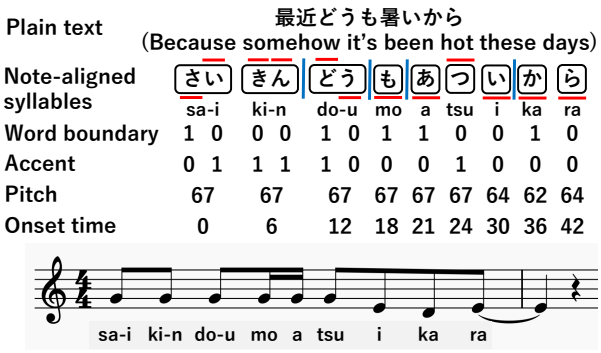

Fig. 1. Example of a melody phrase in the data.

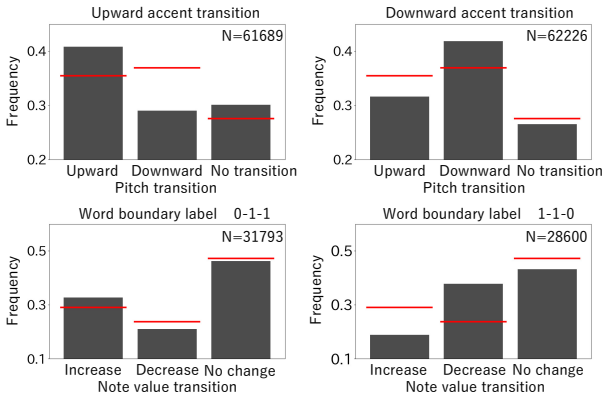

Fig. 2. (Top) Changes in the frequency distribution of melodic pitch transitions conditioned on accent transitions in the lyrics. The red line indicates the overall average frequency. (Bottom) Changes in the frequency distribution of note value transitions due to word boundary labels in the lyrics. The case "increase" indicates that the latter note is longer.

in the melody is aligned with the direction of accent transitions in the lyrics [14].

This principle has also been utilized for music information processing. The automatic song generation system Orpheus [15], imposes constraints on the pitch probabilities according to the accent patterns in the lyrics. The study showed that these constrained improved the user evaluation scores of generated melodies. In another study [16], a system for generating lyrics conditioned on melodies has been developed on the basis of similar constraints involving accents and punctuations of Japanese lyrics. Melody generation utilizing the correspondence between melody and lyrical features has also been studied for other languages such as Chinese and English [12], [17], [18].

B. Construction of Melody Data with Lyrics

To analyze the correspondence between lyrical features and melodic elements, we need melody data in which the lyrics are aligned at the musical note level. Additionally, lyrical features such as accent patterns and word boundaries must be extracted through linguistic analysis. Given that no large-scale dataset of Japanese songs with these conditions is publicly available, we newly constructed an original dataset. The data comprises musical information from the melody and linguistic information from the lyrics (Figure 1). Each note of a melody $(p_l, \tau_l)_{l=1}^{L}$ is represented by a pair consisting of a pitch $p_l \in \{0, \ldots, 127\}$, expressed as a MIDI note number, and an onset time $\tau_l \in \mathbb{N}$, measured in tick units, where one measure is divided into 48 ticks. The pitch is computed after transposing the melody into a natural key, either C major or A minor. The lyrics are expressed in two formats: the standard Japanese script ("plain text") and syllables aligned with each note ("note-aligned syllables"). The plain text is used for the extraction of lyrical features described later. Since 99.92% of the note-aligned syllables in the dataset contain no more than three syllables per note, we set the maximum number of syllables per note to three and excluded phrases with notes exceeding this threshold. The data are segmented into phrases, defined by line breaks in the plain text. Phrases containing English characters, numbers, or special symbols were excluded. In total, we collected $33,751$ phrases ($287,396$ notes) from 1988 Japanese popular songs released from 1960 to 2023.

We consider the following lyrical features that may influence melodic elements: accent patterns, word boundaries, part-of-speech (POS) tags, semantic feature, and syllable labels. Accent patterns, word boundaries, and POS tags are extracted using the Japanese linguistic analyzer MeCab [19]. Following previous studies [15], [16], accent patterns are represented by binary sequences, where 0 indicates a low pitch and 1 indicates a high pitch. Word boundaries are also represented as binary labels, with 1 assigned to the first syllable of each word and 0 to the remaining syllables. POS tags are represented by categorical labels ranging from 1 to 16, corresponding to the 16 major categories of Japanese parts of speech. Semantic features are extracted using a Japanese Sentence-BERT model[1]. Each phrase is represented as a 768-dimensional embedding vector. To reduce the risk of overfitting due to high dimensionality, we apply principal component analysis (PCA) to obtain a lower-dimensional representation for use in the subsequent analysis.

C. Basic Data Analysis

We analyzed the dependence of melodic pitch on accent patterns in the lyrics using the constructed dataset. The upper part of Fig. 2 shows the frequency distribution of pitch transitions for each type of accent transition across consecutive notes. The results show that when the accent pattern exhibits an upward or downward transition, the corresponding upward or downward pitch transitions occur more frequently than the overall average. This suggests that the degree of alignment between accent transitions and the melodic pitch contour may serve as a meaningful indicator for predicting melody evaluation scores. We also analyzed the relationship between word boundaries and note values. The lower part of Fig. 2 shows the frequency distribution of note value transitions between consecutive notes for each pattern of word boundary labels. The results indicate that notes corresponding to single-syllable words tend to have longer durations than their neighboring notes. However, such clear correspondences are not easily observed for other lyrical

[1]https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

features and melodic elements. Accordingly, machine learning is considered an effective approach for modeling the complex relationships among these features and for predicting melody evaluation scores.

III. Method

A. Construction of Fitness Function Based on the Artwork Selection Model

To predict melody evaluation scores using machine learning, we propose a method in which the fitness of a melody with respect to the lyrics is first estimated, and the predicted fitness is then used to infer the melody's evaluation score. By employing the following artwork selection model, the fitness of a melody for a creator can be expressed as the ratio between the unconditional generation probability of the melody and its conditional probability given the lyrics. This function can be learned using melody data with lyrics.

In the artwork selection model we assume that a creator generates various candidate works and probabilistically selects which ones to publish as artworks, using their fitness as weights in the selection process. Formally, let $\widetilde{P}(X)$ denote the generation probability of a candidate melody X, and let $W(X;Y) \in \mathbb{R}_{>0}$ represent the fitness of melody X with respect to lyrics Y. Then, the conditional generation probability of X given Y, denoted as $P(X|Y)$, can be expressed as

$$P(X|Y) = \widetilde{P}(X)W(X;Y). \tag{1}$$

Here, the undetermined scale of the fitness function W is defined so that the above equation holds. Next, by applying the consistency assumption between training and generation processes the creators, which states that $\widetilde{P}(X)$ coincides with the marginal probability of a melody $P(X) = \sum_Y P(X|Y)P(Y)$, the fitness can be expressed as the ratio between the conditional probability given lyrics and the unconditional probability

$$W(X;Y) = P(X|Y)/P(X). \tag{2}$$

Furthermore, through the following transformation, we define the note-level cross-entropy (CE) difference $\Delta\mathrm{CE}(X;Y)$, which is equivalent to the fitness $W(X;Y)$:

$$\log_2 W(X;Y) = -\log_2 P(X) - [-\log_2 P(X|Y)] \tag{3}$$
$$= L(X)\Delta\mathrm{CE}(X;Y). \tag{4}$$

Here, $L(X)$ is the number of notes in melody X.

1) Construction of Generative Models: By constructing a generative model of melody, we compute $P(X|Y)$ and $P(X)$, which appear on the right-hand side of Eq. (2).

Although the melody data includes a variety of pitch ranges and score positions, we assume that the fitness of a melody is independent of these attributes. Therefore, to construct a generative model that is invariant to octave transposition of pitch and measure-level translation of onset time, we represent melody notes using extended pitch classes and metrical positions. The extended pitch class $q_l \in \{0,\dots,35\}$ is an extension of the conventional pitch class defined as $p_l\%12$ (where "%" denotes the modulo operation). This is a faithful representation of pitch intervals up to 18 semitones above and 17 semitones below, unlike the standard pitch classes can only faithfully represent 6 semitones above and 5 semitones below. This is defined as $q_1 = p_1\%12$ and by the following recurrence relation:

$$q_l = [p_{l-1}\%12 + \mathrm{Clip}_{-17}^{18}(p_l - p_{l-1})]\%36. \tag{5}$$

Here, $\mathrm{Clip}_a^b(x)$ is a function that maps the pitch x into the range $\{a,\dots,b\}$ by applying the minimal number of necessary octave transpositions. The metrical position $m_l \equiv \tau_l\%48 \in \{0,\dots,47\}$ represents the relative timing within a measure, faithfully capturing sequences of onset times with note values up to one measure in length. Accordingly, the melody is represented as a sequence of extended pitch classes and metrical positions $X = (q_l, m_l)_{l=1}^L$. Let y_l denote the lyrical features corresponding to the l-th note. Then, the conditional and unconditional generation probabilities are expressed as

$$P(X) = P(q_{1:L}, m_{1:L}), \tag{6}$$
$$P(X|Y) = P(q_{1:L}, m_{1:L}|y_{1:L}). \tag{7}$$

Here, $q_{k:l}$ denotes the set of variables $(q_k, q_{k+1},\dots,q_l)$.

As a simple example of a generative model, we first consider a Markov model. In the unconditional Markov model, the generation probabilities of the extended pitch class and beat position are assumed to be independent. Furthermore, the extended pitch class q_l of the l-th note is assumed to depend only on the preceding extended pitch class q_{l-1} (similarly for the metrical position). Under these assumptions, the generation probability is expressed by the following equation:

$$P(X) = P(q_1)\left[\prod_{l=2}^{L} P(q_l|q_{l-1})\right]P(m_1)\left[\prod_{l=2}^{L} P(m_l|m_{l-1})\right]. \tag{8}$$

In the lyric-conditioned Markov model, the generation probability is factorized as $P(X|Y) = P(q_{1:L}|y_{1:L})P(m_{1:L}|y_{1:L})$. Here, we consider only the accent label $\bar{a}_l \in \{0,1\}$ and word boundary label $\bar{b}_l \in \{0,1\}$ as lyrical features for the l-th note. These are defined from the accent labels $a_{ls} \in \{0,1\}$ and word boundary labels $b_{ls} \in \{0,1\}$ of each syllable s within the l-th note by the following equations:

$$\bar{a}_l = \max\{a_{l1}, a_{l2}, a_{l3}\}, \quad \bar{b}_l = \max\{b_{l1}, b_{l2}, b_{l3}\}. \tag{9}$$

When considering the dependency of accent and word boundary transitions with respect to the melody elements q_l, m_l of each note l, the generation probability is expressed as

$$P(q_{1:L}|y_{1:L}) = P(q_1|\bar{a}_1,\bar{b}_1)\prod_{l=2}^{L} P(q_l|q_{l-1},\bar{a}_l,\bar{a}_{l-1},\bar{b}_l,\bar{b}_{l-1}),$$
$$P(m_{1:L}|y_{1:L}) = P(m_1|\bar{a}_1,\bar{b}_1)\prod_{l=2}^{L} P(m_l|m_{l-1},\bar{a}_l,\bar{a}_{l-1},\bar{b}_l,\bar{b}_{l-1}).$$

Next, to construct a generative model that captures complex relationships between lyrical features and melody elements which cannot be represented by a Markov model, we consider an autoregressive deep generative model based

on a long short-term memory (LSTM) network. The unconditioned LSTM takes (q_{l-1}, m_{l-1}) as input at each step l, and output predicted probabilities $P(q_l|q_{1:(l-1)}, m_{1:(l-1)})$ and $P(m_l|q_{1:(l-1)}, m_{1:(l-1)})$. In the lyric-conditioned LSTM, the outputs remain the same, but the input includes lyrical features, using $(q_{l-1}, m_{l-1}, y_{(l-h):(l+h)})$. By including the lyrical features of surrounding notes within a half-width h in the input, it becomes possible to directly model their dependencies. As the lyrical feature y_l, we use a combination of the accent labels $a_{ls} \in \{0, 1\}$, the boundary labels $b_{ls} \in \{0, 1\}$ for each syllable s within note l, the syllable label c_{ls} (84 types), POS label d_{ls} (16 types), and semantic feature e_l (n-dimensional vector). The variables $q_{l-1}, m_{l-1}, a_{ls}, b_{ls}, c_{ls}, d_{ls}$ are each represented as one-hot vector of dimensions 36, 48, 2, 2, 84 and 16. If the corresponding information is missing, the vector is zero-padded. The semantic feature is provided as a phrase-level vector, which is input to all notes uniformly (this component remains n-dimensional regardless of h). Input dimension is $612h + n + 390$. In the experiments described in Section IV, comparisons are also made using only a subset of the lyrical features as input. In such cases, the input dimensions, except for the semantic feature, are kept fixed, and the unused feature portions are zero-padded.

To estimate the fitness using the above Markov model and LSTM network, each model's parameters are trained using the melody with lyrics data. During evaluation, the trained parameters are used to calculate the generation probabilities $P(X)$ and $P(X|Y)$ for the test data. Hereafter, the fitness estimated by this method is denoted as $W_{\mathrm{Markov}}(X; Y)$ and $W_{\mathrm{LSTM}}(X; Y)$ for the Markov model and LSTM, respectively.

2) Conditional Probability-Based Estimation: As a method for calculating the fitness of a melody with respect to lyrics using a generative model, directly using $P(X|Y)$ can also be considered. In this case, the fitness estimated using the probability $P_{\mathrm{LSTM}}(X|Y)$ from the lyric-conditioned LSTM can be expressed as

$$W_{\mathrm{cond}}(X; Y) = P_{\mathrm{LSTM}}(X|Y). \tag{10}$$

The fitness $W_{\mathrm{LSTM}}(X; Y)$ are represented by the equation $W_{\mathrm{LSTM}}(X; Y) = W_{\mathrm{cond}}(X; Y)/P_{\mathrm{LSTM}}(X)$. where, $P_{\mathrm{LSTM}}(X)$ is the generation probability from the lyric-unconditioned LSTM.

3) Rule-Based Fitness Estimation: It is also possible to estimate fitness using a rule-based method based on the principle of alignment between lyrical accent pattern and melodic contour. In this method, let N_{AC} be the number of syllables within a phrase where the accent goes upward or downward, and N_{match} be the number of times the direction of pitch transition in the melody matches at these accent transitions. The fitness is then expressed by the following equation:

$$W_{\mathrm{rule}}(X; Y) = \frac{N_{\mathrm{match}}}{N_{\mathrm{AC}} + \epsilon} + \epsilon. \tag{11}$$

The constant ϵ is introduced to properly define the fitness when $N_{\mathrm{AC}} = 0$ or $N_{\mathrm{match}} = 0$, and is specifically set to $\epsilon = 10^{-3}$.

B. Prediction of Evaluation Scores

We treat the average evaluation scores of melodies obtained from listening experiments as ground-truth data. Specifically, given a lyric Y and two melodies X_1 and X_2, we define the evaluation score of X_1 as the proportion $p(X_1; X_2, Y)$ of times X_1 is preferred over X_2 by the listeners.

As a method to predict this evaluation score from the fitness $W(X; Y)$ formulated in the previous section, we consider using logistic regression. First, we assume that the fitness of melody X to lyrics Y from the listener's perspective, denoted as $W'(X; Y)$, can be expressed as a function of fitness $W(X; Y)$ from the creator's perspective. Specifically, for fitness values based on rule-based methods or conditional probabilities, we introduce a coefficient α and express the listener's fitness as $W'(X; Y) = W(X; Y)^{\alpha}$. The coefficient α serves to adjust the fitness score, generally reflecting effects such as listeners being less sensitive to differences in musical attributes compared to creators. In the case of fitness based on the probability ratio, $W'(X; Y)$ is expressed as

$$W'(X; Y) = \begin{cases} W(X; Y)^{\alpha_+}, & W(X; Y) \geq 1; \\ W(X; Y)^{\alpha_-}, & W(X; Y) < 1. \end{cases} \tag{12}$$

Here, the coefficients α_+ and α_- are introduced to incorporate an asymmetric correction effect with respect to whether the probability ratio is greater than or less than 1.

Next, in a pairwise comparison of two melodies X_1 and X_2 for a given lyrics Y, if we assume that the selection probabilities for each melody are determined based on the relative ratio of their perceived fitness values $W'(X_1; Y)$ and $W'(X_2; Y)$ by the listener, then the selection probability of X_1 can be expressed as

$$R(X_1; X_2, Y) = \frac{W'(X_1; Y)}{W'(X_1; Y) + W'(X_2; Y)}. \tag{13}$$

Furthermore, in actual experiments, the listener's selection of a melody can be influenced by noise and unknown biasing factors. If this effect is modeled as a linear function of $R(X_1; X_2, Y)$, then the final selection probability of X_1 can be expressed as

$$p(X_1; X_2, Y) = \sigma(\beta_1 R(X_1; X_2, Y) + \beta_0), \tag{14}$$

where $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function. By Eq. (14), this method can be interpreted as a logistic regression in which $R(X_1; X_2, Y)$ serves as the explanatory variable. The parameters $\alpha, \alpha_+, \alpha_-, \beta_0, \beta_1$ are optimized using logistic regression to minimize the squared prediction error.

By extending the above method, it is also possible to predict evaluation scores by combining multiple fitness measures described in Section III-A. For example, using both rule-based fitness W_{rule} and LSTM-based fitness W_{LSTM}, the predicted selection probability $p(X_1; X_2, Y)$ is expressed as

$$\sigma(\beta_2 R_{\mathrm{LSTM}}(X_1; X_2, Y) + \beta_1 R_{\mathrm{match}}(X_1; X_2, Y) + \beta_0). \tag{15}$$

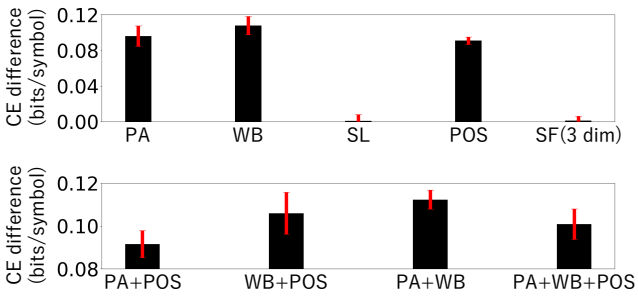

Fig. 3. (Above) Average CE difference for a single lyric feature. (Below) Average CE difference for multiple lyric features. (PA: pitch accent, WB: word boundary, SL: syllable label, POS: part of speech, SF: semantic feature)

IV. Experimental Validation

We split the dataset described in Section II-B into training, validation, and test sets in an 8:1:1 ratio. Both LSTM and Markov models were trained to estimate fitness scores. For the former, we used a 3-layer network with 512 LSTM units, with a linear layer mapping the input to 512 dimensions at the input side, and a linear layer followed by a softmax function at the output side. In preliminary experiments, the optimal values for the CE difference were found to be $n = 3$ for the dimension of the reduced semantic feature, and $h = 3$ for the half-width of the input window for other lyrical features. Accordingly, these values are used hereafter.

A. Relevance of Individual Lyric Features to Fitness

To investigate the contribution of each lyrical feature to the estimation of fitness based on the probability ratio, the average CE difference for each lyric feature calculated using the LSTM network is shown in Fig. 3(top). The average CE differences for accent, word boundary, and POS features were approximately 0.1 bits, which is sufficient to contribute meaningfully to the fitness estimation. In contrast, the average CE differences for syllable labels and semantic feature were small, indicating that their contribution to the fitness estimation is negligible. Regarding the syllable labels, overfitting may have occurred due to the high dimensionality of the input.

The results when multiple elements (accent, word boundary, and POS features) were included as inputs are shown in Fig. 3(bottom). When both accent and word boundary features were used, the average CE difference slightly increased compared to when each was used alone. The slight increase is likely due to the nature of Japanese, where the accent often transitions upward at the second syllable, causing these features to be correlated. Adding the POS feature reduced the CE difference. Hereafter, the combination of accent and word boundary features, which yielded the maximum average CE difference, will be used as the lyrical features.

B. Listening Experiment

To collect real data on listener evaluation scores, we conducted an experiment in which participants were presented with two different melodies with the same lyrics and asked to choose the one they found more preferable [20]. In the experiment, we used 20 pairs of melodies corresponding to 20 different lyrics randomly selected from the test dataset. The synthesized singing voices were generated using NEUTRINO [21] and presented in a random order. The pairs of melodies for comparison were generated using the trained lyrics-conditioned LSTM network and lyrics-unconditioned LSTM network. To ensure the experiment to be informative enough to reveal differences through pairwise comparison, we selected melody pairs for which the CE difference was large and positive for the former and small and negative for the latter. For each melody pair i the proportion p_i (called selection rate) of times the melody generated by the lyrics-conditioned LSTM (the one with a larger CE difference) was selected was used as the evaluation score.

The average value of p_i obtained from the 1546 comparison results by 186 participants was 0.561, and the deviation from random selection ($p_i = 1/2$) was statistically significant ($p = 1.35 \times 10^{-6}$). Additionally, when melodies generated under the same conditions were presented with instrumental sounds, the average value of p_i was 0.452. This confirmed that the significant difference in the average evaluation scores from $1/2$ was attributable to the relationship with the lyrics. We examined the relationship between listers' attributes (age, experience of composition and performance, and daily music listening time) and their selections of preferred melodies, but no statistically significant dependence was observed.

TABLE I
Evaluation results for various fitness estimation methods.

Fitness estimation method	RMSE	Correlation	p-value
Rule-based (W_{rule})	0.149	0.542	0.0135
Conditional probability (W_{cond})	0.175	0.143	0.5467
Probability ratio W_{Markov}	0.171	0.264	0.2602
Probability ratio W_{LSTM}	0.159	0.443	0.0496
W_{LSTM} and W_{rule}	0.134	0.653	0.0018

C. Evaluation of Score Prediction

The results of the evaluation score prediction using the method in Section III-B are shown in Table I. The root mean squared error (RMSE) between the predicted and observed values of evaluation scores was used as the evaluation metric, and for reference, the correlation coefficient between the predicted and observed values, along with its p-value, is also provided. When using a single fitness, the rule-based method exhibited the smallest error. In the comparison using generative models, it was shown that using the probability ratio is more effective than using conditional probabilities, and that using the LSTM network was more effective than using the Markov model. Additionally, when combining the probability ratio from the LSTM network with the rule-based method, the prediction error was smallest. In Fig. 4, an example can be observed where the rule-based method yields the same prediction for certain data points. The prediction was improved by incorporating the fitness from the LSTM. This indicates that evaluation score prediction can incorporate relationships between lyrics and melody that cannot be captured by simple

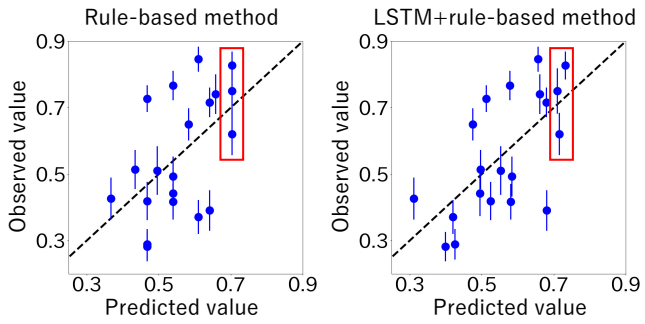

Fig. 4. Observed and predicted values of evaluation scores (error bars represent standard errors).

rules, through the use of deep generative models. Furthermore, the learned weights ($\alpha_+ = 0.001$, $\alpha_- = 0.933$) suggest that listener evaluation is more sensitive to low-fitness regions.

V. Conclusions

This study showed that, in predicting the evaluation scores of melodies with respect to Japanese lyrics, accent patterns and word boundaries in the lyrics contribute significantly, whereas syllable labels, part-of-speech tags, and semantic feature have relatively small contributions. It was found that fitness estimation based on the probability ratio from a melody generation model can predict the results of listening experiments with high accuracy when combined with a rule-based method derived from the principle of alignment between lyrical accent pattern and melodic pitch contour. The proposed method can be generalized as a framework for estimating the evaluation scores of artworks in a more specific domain by using the ratio between the generation probability of that data and that of data in a broader domain. Due to its generalizability, this approach could be applied to the automatic evaluation of vocal music in languages other than Japanese, instrumental music, and even artistic data beyond the domain of music.

Acknowledgements

This research was partially supported by JST FOREST No. JPMJPR226X and JSPS KAKENHI Nos. 25H01148 and 25H01169.

References

[1] L. C. Yang *et al.*, "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation," in *Proc. ISMIR*, 2017, pp. 324–331.

[2] H. W. Dong *et al.*, "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proc. AAAI*, 2018, pp. 34–41.

[3] A. Roberts *et al.*, "A hierarchical latent vector model for learning long-term structure in music," in *Proc. ICML*, 2018, pp. 4364–4373.

[4] G. Mittal *et al.*, "Symbolic music generation with diffusion models," in *Proc. ISMIR*, 2021, pp. 468–475.

[5] Y. -S. Huang and Y. -H. Yang, "Pop Music Transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proc. ACM Multimedia*, 2020, pp. 1180–1188.

[6] S. Mehri *et al.*, "SampleRNN: An unconditional end-to-end neural audio generation model," in *Proc. ICLR*, 2017.

[7] P. Dhariwal *et al.*, "Jukebox: A generative model for music," *preprint arXiv:2005.00341*, 2020.

[8] N. Tokui and H. Iba, "Music composition with interactive evolutionary computation," in *Proc. International Conference on Generative Art*, 2000, pp. 215–226.

[9] R. M. MacCallum *et al.*, "Evolution of music by public choice," *PNAS*, vol. 109, no. 30, pp. 12 081–12 086, 2012.

[10] G. Cideron *et al.*, "MusicRL: Aligning music generation to human preferences," in *Proc. ICML*, 2024, pp. 8968–8984.

[11] J. Copet *et al.*, "Simple and controllable music generation," in *Proc. NeurIPS*, 2023, pp. 47 704–47 720.

[12] C. Zhang *et al.*, "ReLyMe: Improving lyric-to-melody generation by incorporating lyric-melody relationships," in *Proc. ACM Multimedia*, 2022, pp. 1047–1056.

[13] L. Chai and D. Wang, "CSL-L2M: Controllable song-level lyric-to-melody generation based on conditional transformer with fine-grained lyric and musical controls," in *Proc. AAAI*, 2025, pp. 23 541–23 549.

[14] K. Yamada, "The accent of poetry viewed from the perspective of popular song composition (in japanese)," *Poetry and Music*, vol. 2, no. 2, 1923.

[15] S. Fukayama *et al.*, "Orpheus: Automatic composition system considering prosody of Japanese lyrics," in *Proc. ICEC*, 2009, pp. 309–310.

[16] K. Watanabe *et al.*, "A melody-conditioned lyrics language model," in *Proc. Conference of the North American Chapter of the ACL: Human Language Technologies*, vol. 1, 2018, pp. 163–172.

[17] Z. Sheng *et al.*, "SongMASS: Automatic song writing with pre-training and alignment constraint," in *Proc. AAAI*, vol. 35, 2021, pp. 13 798–13 805.

[18] Z. Ju *et al.*, "TeleMelody: Lyric-to-melody generation with a template-based two-stage method," in *Proc. EMNLP*, 2022, pp. 5426–5437.

[19] T. Kudo *et al.*, "Applying conditional random fields to Japanese morphological analysis," in *Proc. EMNLP*, 2004, pp. 230–237.

[20] Experiment page, https://ice.inf.kyushu-u.ac.jp/ExpNishi2 [online], 2024.

[21] NEUTRINO Diffusion, https://studio-neutrino.com/ [online], 2024.