Controllable Automatic Melody Composition Model across Pitch/Stress-accent Languages

Takuya Takahashi[1], Shigeki Sagayama[1] and Toru Nakashika[1] *

The University of Electro-Communications, Tokyo, Japan
{takahashi,sagayama,nakashika}@uec.ac.jp

Abstract. This study proposes a model for automatically composing linguistically and musically natural song melodies reflecting the linguistic characteristics of both pitch-accent (e.g., Japanese) and stress-accent (e.g., English) languages as well as user's intentions. We have designed and provided publically, for more than 10 years, an automatic composition system (called "Orpheus") for Japanese lyrics. Extending the principle for lyrics written in stress-accent languages, a new compositional model was constructed by introducing a melodic rhythm generator formulated by a probabilistic model considering the relationship between stress of lyrics and rhythm intensity (linguistic naturalness and music theory) and the rhythm style chosen by the user (controllability). The parameters of the proposed model can be learnt from domain knowledge without large amounts of data. In our experimental evaluation, the proposed system achieved ratings equal to or better than state-of-the-art deep learning approaches in terms of musical coherence, singability and listenability.

Keywords: Automatic music composition, Lyric to melody, Music theory, Linguistic naturalness for melody, Controllability

1 Introduction

Automatic music composition is one of the most interesting and challenging tasks in generative AI (such as ChatGPT [1]), as interest in generative AI has grown in recent years. How would users want to use automatic composition technologies? We believe automatic composition technologies should be an assistive tool that users can use for their creative activities so that beginners can compose music with only their intention without knowledge of composition theory which takes time to learn, and experts can gain new inspiration from the generation from AI composers. We are particularly interested in building a universal model for generating song for singing automatically based on user-given lyrics, that follows Western musical norms.

* This work was supported by Grant-in-Aid for Scientific Research (B) No. 21H03462 from Japan Society for the Promotion of Science (JSPS).

[1] https://openai.com/blog/chatgpt

State-of-the-art data-driven methods of deep learning from large amounts of data [1–3] have been proposed for automatically generating music from lyrics. Although such methods can cleverly learn patterns in the training data and generate coherent music that is close to the training data, difficulties in collecting large amounts of paired data, controllability to reflect user intent, diversity of generation (avoiding plagiarism) and musical accuracy (adherence to music theory) are often discussed. For example, Zheng et al. [4] reported owing to their subjective experiments that melodies generated by the deep learning models proposed by Sheng et al. [2] and Ju et al. [3] are difficult to sing and listen to the lyrics. Besides, DeepBach trained on Bach chorales using deep learning cannot generate pieces that adhere to music theory such as musical prohibition as Fang et al. [5] and Karatsu et al. [6] argued. Such reports may suggest that it is difficult for state-of-the-art deep learning approaches to learn singability, listenability and music theory.

Can we then rule and model the composition process in the automatic composition of songs, in addition to learning patterns from data like deep learning methods? For example, Oliveira et al. [7] investigated the relationship between stress syllables and melodic rhythm in 42 Portuguese songs and reported a correlation between stress and melodic rhythm (referred to as the stress-rhythm constraint). In an attempt to generate melodies from English lyrics, a method based on the stress-rhythm constraint and n-gram models was proposed by Monteith et al. [8]. Zhang et al. [4] report improving melodies generated by the latest deep learning methods (Sheng et al. [2], Ju et al. [3]) adjusting melody generation process based on linguistic naturalness constraints for tone languages and stress-accent languages similar to Monteith et al. [8] However, the approach of Zhang et al. [4] requires large amounts of data to train base deep models and leaves issues in terms of diversity, adherence to music theory and user controllability. We (Fukayama et al.) [9] previously proposed Orpheus, which generates the pitch of a melody based on the Japanese lyrics, music theory and user intentions (melodic rhythm, chord progression, register, etc.). As statistically demonstrated by Watanabe et al.[10], for lyrics in pitch-accent languages such as Japanese, the correlated nature of lyric prosody and the vertical movement of melodic pitch (called prosody-pitch constraints) is incorporated as a linguistic naturalness constraint in the melodic pitch generation model of Orpheus [9]. Orpheus [9] has been operating as a Web service (Orpheus v3) for more than 10 years, has over 15,000 subscribers, and has composed more than 500,000 songs. For simplicity, this approach is referred to as “Orpheus v3”.

In addition to the principle of melody generation from lyrics in pitch-accented languages in Orpheus v3, melody generation from lyrics in stress-accented languages was also expected. In this study, a model that can automatically generate a natural melody based on a given lyric written in pitch/stress-accent languages and user’s intention was realized by a combination of pitch generator from Orpheus v3 considering prosody-pitch constraints and music theory and a newly proposed rhythm generation model considering stress-rhythm constraints and music theory. Since each generator in the proposed model is formulated in probabilistic models, it can learn the probabilistic parameters from domain knowledge with explicit consideration of linguistic naturalness for both aspects of pitch and rhythm, music theory and user’s intention without a large amount of data as in deep learning. Moreover, as Orpheus v3 users had commented that

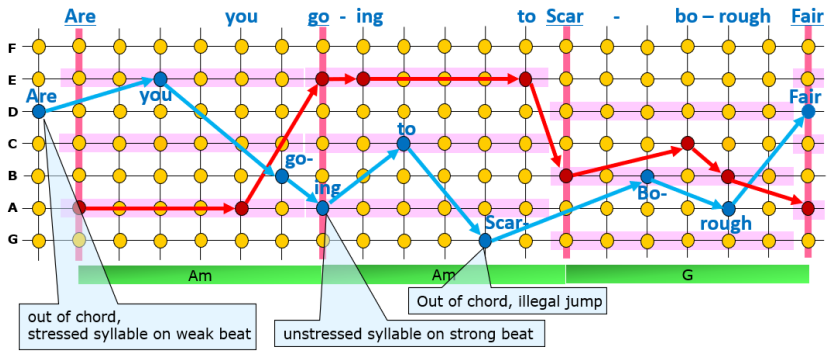

Fig. 1. Conceptual diagram of a singing song generation model based on path-finding. The red path represents acceptable melodies and the blue represents unacceptable melodies.

they found it difficult to sing due to the fact that the sentence breaks did not match the bar line, it was also hypothesised that placing sentence beginnings on stronger beats would improve singability and listenability in pitch-accented languages. The compositional principles of the proposed model were evaluated objectively in terms of linguistic naturalness, as well as subjectively by the audience.

2 Melody Generation Model

Since melodic composition from lyrics is the problem of assigning notes to lyric syllables, it can be understood as the problem of finding a path on a grid of points on a two-dimensional plane of time (e.g. 16th note resolution) and pitch (e.g. semitones) to each syllable, as shown in Figure 1. Although pathways, i.e. pitch and rhythm combinations, are vast, the pathway cost of finding a valid melody with respect to domain knowledge of music theory, linguistic naturalness and user intent can be defined mathematically. Such a model for simultaneously generating the rhythm (onset time and duration) sequence ($\hat{r}_{1:N}$) and the pitch sequence ($\hat{p}_{1:N}$) optimised to the compositional conditions including lyrics ($\mathbf{z}$) can be formulated as the maximisation of a probabilistic model:

$$\hat{r}_{1:N}\hat{p}_{1:N} = \underset{r_{1:N},p_{1:N}}{\arg\max}\, p(r_{1:N}, p_{1:N}|\mathbf{z}) \tag{1}$$

where $r_{1:N}$ and $p_{1:N}$ are random variables. However, since Equation 1 is too computationally complex, we assumed in Orpheus v3 that pitch and rhythm are independent and rhythm is given in advance, and the melodic pitch generator $\mathrm{argmax}_{\mathrm{p}_{1:N}} p(p_{1:N}|\mathbf{z})$ was realised by applying the Viterbi algorithm in a probabilistic model that follows Markov processes. In this paper, by introducing a rhythm generator ($\mathrm{argmax}_{r_{1:N}} p(r_{1:N}|\mathbf{z})$), which considers the musical domain knowledge, to the pitch generator of Orpheus v3, we propose a new melody generation model with high controllability that optimised to all aspects of user intention, music theory and linguistic naturalness in terms of both

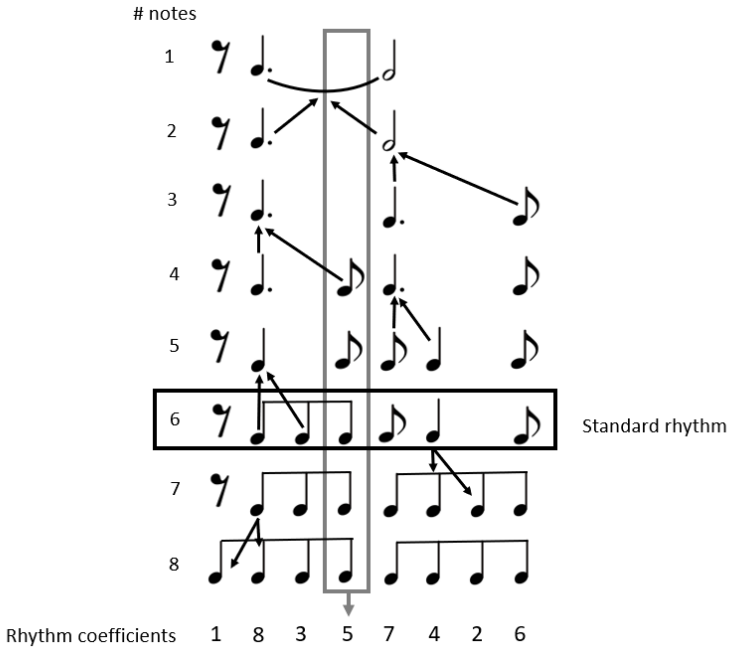

Fig. 2. Example of a rhythm tree for rhythm generation from each rhythm family. The rhythm coefficients at the bottom represent the importance of each onset event to the rhythm family.

rhythm and pitch. Note that this is an approximate solution of Equation 1 due to computational complexity, i.e. the vertical (pitch) and horizontal (rhythm) axes of Figure 1 are optimised separately. The next section describes the proposed rhythm generators.

2.1 Melodic rhythm generator

Overview What are the requirements for melodic rhythm generation models in creative automatic composition systems for users? It is not easy to create rhythm from scraps without studying the composition techniques. However, it is natural to want to compose the same section, e.g. the first and second choruses, with similar melodic-rhythmic patterns as suggested by Fukayama et al. [9]. This means that the rhythmic pattern should be controllable from section to section.

Melodic rhythm generation in Orpheus v3 In Orpheus v3, the rhythm generator is represented by "rhythm tree structure." As shown in Figure 2, a rhythm tree has a rhythm called a "standard rhythm," which is a good representation of its rhythm generator, and it is expanded and integrated so that they are perceived as similar to standard rhythms, even if the number of notes changes. By providing the rhythm tree that can generate similar rhythm patterns for each number of notes, users can control the melodic

rhythm by setting a rhythm family for each section. Rhythm trees in Orpheus v3 do not take into account the linguistic naturalness of the lyrics, as they are defined before the lyrics are input.

Melodic rhythm generation model considering rhythmic style, accent position and duration We aimed at an automatic generation model of melodic rhythms that preserves the rhythm family selected by the user as before, while also ensuring the linguistic naturalness of the lyrics by considering the relationship between syllable stress intensity and rhythmic intensity. Given a syllable feature vector sequence $mathbf{s}_{1:N}$, such a model that generates a rhythmic sequence $r_{1:N}$ containing N onset events can be modelled by dynamic Bayesian networks (DBNs) as follows:

$$\begin{aligned} p(r_{1:N}|\mathbf{s}_{1:N}) &\propto p(\mathbf{s}_{1:N}|r_{1:N})p(r_{1:N}) \\ &\approx p(r_1)p(\mathbf{s}_1|r_1)\prod_{i=2}^{N} p(\mathbf{s}_i|r_i)p(r_i|r_{i-1}) \end{aligned} \tag{2}$$

where rhythm sequence generation probabilities are approximated by Markov process and the i represents the consecutive numbers of syllables and not the rhythmic time.

Figure 3 shows the trellis of the proposed DBNs, in which horizontal nodes representing the onset events (16th-note resolution) and nodes are developed for the number of syllables in the vertical direction. Horizontal transition jumps are permitted to represent the duration of rhythmic events, while vertical jumps are not permitted as the syllables should be in sequence. However, the position of the rests has to be provided by the user same as the syllables.

The rhythm sequence with the highest likelihood generated from the proposed DBNs can be efficiently obtained using the Viterbi approach [11] and the rhythm sequence is assumed to respect rhythm family and linguistically naturalness. Finally, a melody is generated by combining the most likely rhythm generated by the proposed rhythm generator and the most likely pitch sequence generated by the Orpheus v3 pitch generator. The following sections describe how each probability parameter is learnt.

Linguistic naturalness constraints The $p(\mathbf{s}_i|r_i)$ serves as a term to guarantee the linguistic naturalness of the lyrics. The s_i is the feature vector of the syllable, which includes the syllable stress intensity $s_{i,\mathrm{stress}}$ and the syllable length $s_{i,\mathrm{length}}$. Assuming the independence of each of them,

$$p(\mathbf{s}_i|r_i) = p(s_{i,\mathrm{stress}}|r_i)p(s_{i,\mathrm{length}}|r_i) \tag{3}$$

$p(s_{i,\mathrm{stress}}|r_i)$ can be formulated based on the findings of Oliveira et al. [7]. Assuming that $s_{i,\mathrm{stress}}$ follows a normal distribution with mean as the rhythmic intensity $r_{i,\mathrm{intensity}}$ and standard deviation as 1 empirically,

$$p(s_{i,\mathrm{stress}}|r_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(s_{i,\mathrm{stress}} - r_{i,\mathrm{intensity}})^2} \tag{4}$$

where $r_{i,\mathrm{intensity}}$ can be determined by music-theoretic knowledge, for example, the rhythmic intensity can be set heuristically as shown in Figure 3 based on music-theoretical

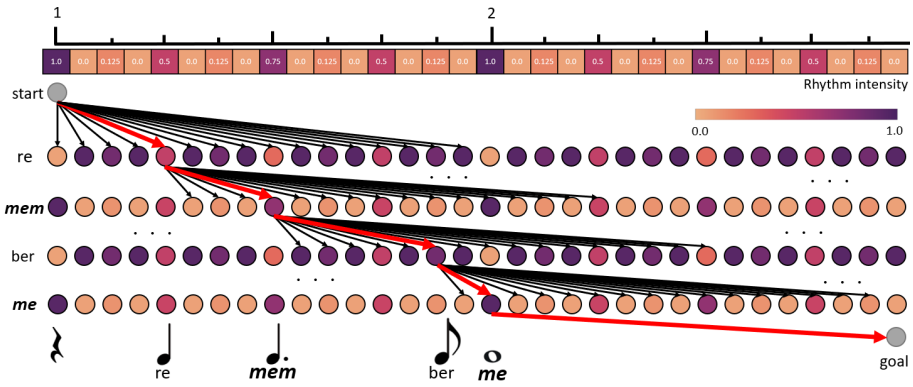

Fig. 3. An example of the path-finding trellis of the rhythm generation DBNs at 16th note resolution in 4/4 time when not syncopated. The rhythmic intensity is determined heuristically based on music theory knowledge and the state likelihood of each node is calculated based on the difference between the rhythmic intensity and syllable stress intensity as an example.

knowledge. In terms of $p(s_{i,\mathrm{length}})$, it would be possible to formulate syllable length ($s_{i,\mathrm{length}}$) following a normal distribution with mean as rhythmic duration ($r_{i,\mathrm{duration}}$) same as Equation 4. However, in this study, only rhythm events with stress-syllable and too short duration (sixteenth notes) were penalised from the point of view of singability, and all other values were given the same probability. In this way, the model can consider both music theory and linguistic naturalness.

Rhythm family constraints The $p(r_i|r_{i-1})$ is the transition probability of a rhythmic event and can be trained on the basis of the rhythm trees defined in Orpheus v3 to generate rhythm based on user-specified rhythm families. We hypothesise that in a rhythm tree, the smaller the number of pitch accents, the more characteristic rhythmic patterns remain, and the higher the number of pitch accents, the more redundant patterns are injected. Thus, by calculating how many times an onset event appeared vertically for the entire rhythm tree, as shown in Figure 2, rhythm coefficients, which indicate how important each onset event was for that rhythm family, can be calculated. This rhythm coefficient was normalised by dividing it by the sum of the rhythm coefficients within the rhythm family and was the likelihood for onset event. The parameters in the rhythm generation models trained with such likelihoods can generate essential patterns within the rhythm family with high priority. The inclusion of such likelihoods in the model is expected to generate melodic rhythms that respect the user's chosen rhythmic family.

3 Experimental Evaluation

In this experiment, the naturalness and coherence of the melodies generated by the proposed automatic composition system from Japanese and English lyrics were assessed objectively and subjectively.

3.1 Input data and proposed model setup

The stress accent intensity of lyrics was determined as follows.

- Japanese: Strong accent (1.0) placed at the beginning of a phrase. Accent values for other syllables were set to 0.0.
- English: 1.0 was placed on the primary accent and 0.75 on the secondary accent, and if the accented word was a weak form (e.g. preposition), the accent value was multiplied by 0.25. Accent values for all other syllables were set to 0.0.

The prosodies of lyrics were determined as follows.

- Japanese: Morphological analysis results from Mecab [2] are used.
- English: Only intonation within words was restricted, with reference to the F0 of the speech sound of the lyrics. Other pitch changes were allowed freely.

All rhythm families used in this experiment were non-syncopated 4/4 time rhythm patterns, and the intensity values for each onset event in one bar with 16th note resolution were set to the same values as in Figure 3, based on music-theoretical knowledge.

3.2 Objective evaluation

Experimental condition Objective evaluation experiments investigated the reflection of linguistic naturalness constraints in the melodies generated by the proposed model and Orpheus v3. A total of 12 songs, which were generated by each of Orpheus v3 and the proposed model based on a combination of 4 randomly selected rhythm families and their accompanying composition conditions (such as chord progression, accompaniment, drums, etc.), and 3 nursery rhyme lyrics (London bridge, Amazing grace, Scarborough Fair; opening eight bars of lyrics), were evaluated.

Results Rhythmic naturalness was assessed by the mean square error between the rhythmic intensity of each rhythmic event in the generated melody ($E_r(r_i)$ from Equation 4) and the stress intensity of the phoneme corresponding to each rhythmic event in the lyrics as defined in section 3.1. Rhythm naturalness was about 0.296 for Orpheus v3 and **0.195** for the proposed method. In this way, the melodies generated by the proposed method are more consistent regarding the relationship between stress and rhythm.

Pitch naturalness was similarly evaluated for all syllables in the lyrics by the percentage of match between the transition direction of the melodic pitch (up or down) and the prosody of the syllable (up or down). As a result, Orpheus v3 and the proposed method achieved almost the same values, 0.991 and **0.994** respectively. Therefore, even if the proposed rhythm generator is introduced, the existing pitch generators are still functioning adequately.

Figure 4 shows an example of the song actually generated from the proposed model and Orpheus v3 respectively, based on the same lyrics and compositional conditions. As can be seen from these figures, the knowledge obtained in the objective evaluation experiment can be found concretely. Further examples of generated scores and sound sources can be found on the URL [3].

[2] `https://taku910.github.io/mecab/`

[3] `https://coconuts-palm-lab.com/cmmr2023`

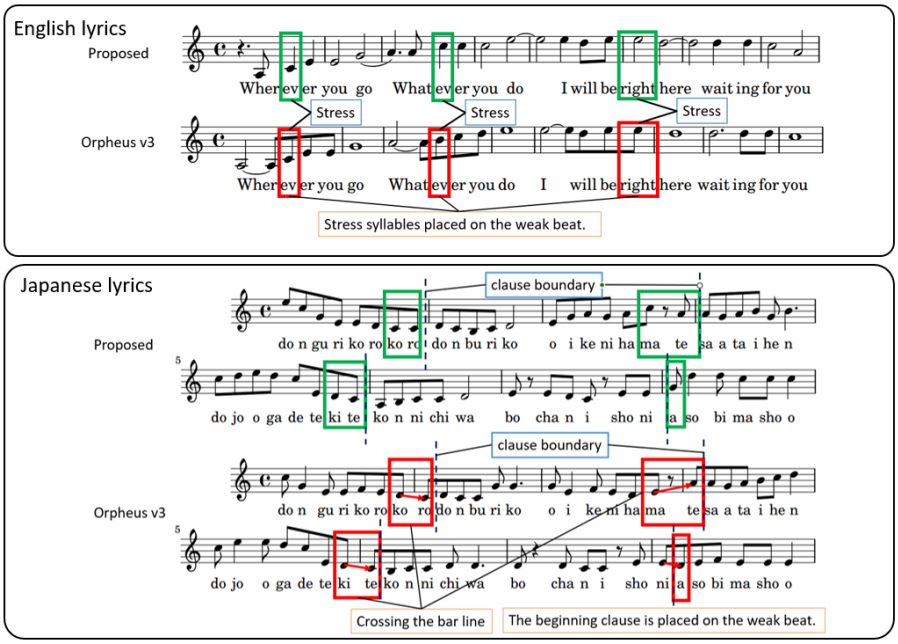

Fig. 4. Examples of generated melodies for the comparative and proposed methods when English and Japanese lyrics are used as input.

3.3 Subjective evaluation (Rhythm family)

Assessment conditions This experiment evaluated the similarity of melodic rhythms generated by the proposed model or Orpheus v3 for lyrics based on the same rhythm family but with different numbers of syllables and different accents. In this experiment, participants were asked to subjectively rate the rhythmic similarity of melodies generated by a combination of three nursery rhymes and three randomly selected rhythm families, similar to the objective assessment experiment. The evaluation is conducted with the XAB test, where X and A or, B are melodies generated from the same rhythm family. Users were instructed to listen to X first, then A and B, and to choose from A or B whichever they felt was closer to X in terms of rhythmic pattern.

Stimuli In addition to the melody, accompaniments and drum patterns generated from pre-prepared compositional conditions were assigned to each of the three rhythmic families to facilitate the capture of the beat and chord progression. The MIDI data generated from the models were synthesised using FluidSynth [12] at 44100 Hz, using the Fluid R3 sound font. Note, however, that as this is an experiment to assess the similarity of rhythmic patterns, the melody is played on a saxophone to make it easier to distinguish from the others, and the singing voice is not included in the sound source.

Participants and procedure The experiment was conducted online. The 25 participants in the experiment were all Japanese, with an average age of about 25. 80% of the participants had not trained musically for more than two years. The experiment began with an investigation of the participants' backgrounds, which included a survey of their age, country of residence and musical background based on Goldsmith-MSI [4]. In the main part of the experiment, participants were given just one question with answers in a similar format to the actual XAB test for a tutorial on the XAB test, and after gaining an understanding of the XAB test, they answered 10 XAB tests per person.

Results The results of the XAB test on rhythmic similarity showed that the proportion of selecting sound sources containing melodies generated from the same rhythmic family as sound source X was about 75% in both the proposed model and Orpheus v3. The results suggest that the proposed model can generate melodic rhythms with comparable rhythmic control performance to Orpheus v3 while it respects user-selected rhythmic families as well as linguistic naturalness constraints. There was concern that the constraints of linguistic naturalness might break the rhythmic patterns of the rhythm family, but this may suggest that probabilistic parameter learning with rhythm coefficients is a reasonable representation of the original rhythmic patterns.

3.4 Subjective evaluation (Generated melody assessment)

Assessment conditions This experiment aims to subjectively assess the consistency, singability and listenability of the melodies generated by the proposed model and the comparative method. For comparison, we used Orpheus v3 as a baseline for English lyrics, and also compared it with SongMASS[2], TeleMelody[2], SongMASS + Relyme [4], and TeleMelody + Relyme [4], respectively. In the case of melody generation with Japanese lyrics, since it was not possible to prepare Japanese lyrics/melody pair data for training the latest deep learning methods and Relyme [4] principles were targeted at Chinese and English lyrics, only the proposed method was compared with Orpheus v3.

Stimuli The English lyrics were picked from the three lyrics used by Zhang et al. [4] [5] as test data in order to conduct fair test. For the Japanese lyrics, three well-known Japanese children's songs (Donguri Korokoro, Okina Noppo No Furudoke and Urashima Taro) were selected. The singing voice based on the lyrics was synthesised by the Maki Tsurumaki sound source on Synthesizer V. The deep learning method under comparison does not have the support for generating accompaniment or drums, so for fair evaluation, the English lyrics experiment used only the singing voice and melody guide (played on a saxophone), excluding the accompaniment and drums. In the Japanese comparison experiment, the sound sources were synthesised using FluidSynth (same as section 3.3) according to the MIDI of the melody, accompaniment and drums generated from each model and combined with the singing voice.

[4] https://www.gold.ac.uk/music-mind-brain/gold-msi/
[5] https://ai-muzic.github.io/relyme/

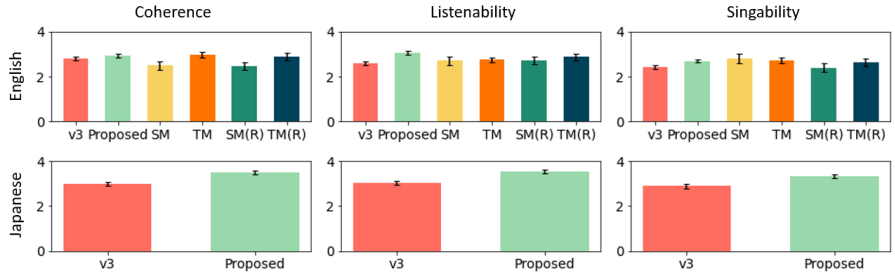

Fig. 5. Results of subjective experimental evaluation. v3 stands for the Oprheus v3 baseline method [9], SM for Songmass [2], TL for Telemelody [3] and R for Relyme [4] in combination with the methods of [2] and [3], respectively.

Participants and procedure The experiment was conducted online. The participants were 25 people who also took part in the experiment in section 3.3. Participants answered five-point Likert scale questions on three aspects of musical coherence, singability and listenability for songs generated from English (25) and Japanese (15) lyrics. However, to clarify which syllables were assigned to which notes and to minimize the effect of the singing synthesis, participants were presented with an image of the melodic score as well as the sound source simultaneously. In addition, we reminded participants that singing synthesis and playing instruments are not the scopes of our study.

Results The mean and standard error statistics of the experimental results are summarised in Figure 5. For English lyrics, the proposed method was rated significantly higher than the baseline method, Orpheus v3, on all items of coherence, listenability and singability. When comparing state-of-the-art deep learning methods with the proposed method, the proposed method obtained significantly higher ratings for coherence than some deep learning models (SM, SM(+R)) and for listenability than most deep learning models, respectively. Although there were no significant differences between the deep learning method and the proposed method, most of the participants in this experiment were not music experts and therefore had difficulty in assessing singability. Since there are no items where the proposed method is significantly inferior to the state-of-the-art deep learning methods, it suggests that the proposed method may be equal to or superior to the state-of-the-art deep learning methods in English lyrics. In addition, focusing on the standard errors, the fact that the latest deep learning methods have large standard errors while the proposed method has small standard errors seems to indicate the robustness of the proposed method.

For Japanese lyrics, all items were rated significantly higher for the proposed method than for the baseline method Orpheus v3. This may be because the placement of the initial syllable of a passage on a stronger beat might sharpen the semantic break in pitch-accent language. The results suggest that there is a clear relationship between semantic delimitation and melodic rhythm in the pitch-accent language, and that the proposed rhythm generator works effectively in the pitch-accent language.

However, as this experiment was conducted with 25 Japanese subjects, there is a bias, and therefore a similar experiment needs to be conducted with more participants and not only with native speakers of Japanese, but also with native speakers of other languages, in order to make a more generalised assessment.

4 Discussion

The results of the evaluation experiments show that the proposed method can generate melodies that respect the rhythmic pattern of the user-selected rhythm family, while taking into account linguistic constraints such as stress-rhythm constraints and prosody-pitch constraints. It has also been shown that the consideration of linguistic naturalness, as incorporated in the proposed method, improves singability and listenability compared to the baseline method. The proposed method, trained only on domain knowledge without training on large amounts of training data, was rated as good as or better than state-of-the-art deep learning methods.

Since pitch and rhythm are optimised separately in the proposed model, it is difficult to add constraints considering pitch and rhythm simultaneously. For example, when singing in the high register, a series of notes with short duration makes singing difficult and non-chord tones, such as neighbour and passing tones, are known to have weak beats and short duration. In order to apply such knowledge as a constraint for melody generation, it is necessary to consider path-finding on a 2D plane, as shown in Figure 1, and thus to study how to solve the problem of the computational cost of Figure 1.

In order to realise a universal compositional principle, it is expected to support lyrics in tonal languages in addition to stress and pitch-accented languages. For tonal languages, this can be resolved by DBNs with states that take into account the possibility of pitch transitions occurring within a single syllable.

The proposed model is formulated by a hidden Markov model, which means that the computational complexity increases exponentially when trying to consider long contexts. From the results of the experimental evaluation, it seemed possible to generate melodies that could convince the audience by considering local music theory and linguistic naturalness. However, a longer context might allow the model to take into account song styles (e.g. genre, artist) in the training data and add user-selectable compositional styles to the compositional conditions.

5 Conclusion

This study proposed a probabilistic model that targets the generation of the most linguistically and musically natural song melody based on the user's input lyrics and compositional conditions. In addition to the melodic pitch generator considering relationship between prosody and melodic pitch of lyrics that have been considered in our previous research (Orpheus), the proposed system introduced a melodic rhythm generator in which the probability parameters are learned so that the stress of lyrics and the melodic rhythm intensity are consistent while respecting the rhythm style selected by the user. The results of the experimental evaluation showed that it is possible to generate melodies that are reasonable in terms of linguistic naturalness and music theory,

while maintaining the same level of controllability as the previous Orpheus v3. Subjective evaluation experiments showed that the melodies generated by the proposed model were equal to or better than state-of-the-art deep learning methods in terms of musical coherence, singability and listenability.

In the future, by solving the problem of computational complexity, the aim is to build a model that simultaneously considers pitch and rhythm, which can make use of vocal and other music-theoretical knowledge such as non-chord tones. We will also explore how knowledge from deep learning methods that can consider longer-term context can be used in the proposed model.

References

1. Hangbo Bao, Shaohan Huang, Furu Wei, Lei Cui, Yu Wu, Chuanqi Tan, Songhao Piao, and Ming Zhou. Neural melody composition from lyrics. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*, pages 499–511. Springer, 2019.
2. Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13798–13805, 2021.
3. Zeqian Ju, Peiling Lu, Xu Tan, Rui Wang, Chen Zhang, Songruoyao Wu, Kejun Zhang, Xiangyang Li, Tao Qin, and Tie-Yan Liu. Telemelody: Lyric-to-melody generation with a template-based two-stage method. *arXiv preprint arXiv:2109.09617*, 2021.
4. Chen Zhang, Luchin Chang, Songruoyao Wu, Xu Tan, Tao Qin, Tie-Yan Liu, and Kejun Zhang. Relyme: Improving lyric-to-melody generation by incorporating lyric-melody relationships. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1047–1056, 2022.
5. Alexander Fang, Alisa Liu, Prem Seetharaman, and Bryan Pardo. Bach or mock? a grading function for chorales in the style of js bach. *arXiv preprint arXiv:2006.13329*, 2020.
6. Yuki Karatsu and Satoru Fukayama. A comparative study of data augmentation methods for automatic harmony generation with a low number of forbidden violations. *Proc. Spring Meet. Acoust. Soc. Jpn.*, 2023.
7. Hugo R Gonçalo Oliveira, F Amilcar Cardoso, and Francisco C Pereira. Tra-la-lyrics: An approach to generate text based on rhythm. In *Proceedings of the 4th. International Joint Workshop on Computational Creativity*. A. Cardoso and G. Wiggins, 2007.
8. Kristine Monteith, Tony R Martinez, and Dan Ventura. Automatic generation of melodic accompaniments for lyrics. In *ICCC*, pages 87–94, 2012.
9. Satoru Fukayama, Kei Nakatsuma, Shinji Sako, Yuichiro Yonebayashi, Tae Hun Kim, Si Wei Qin, Takuho Nakano, Takuya Nishimoto, and Shigeki Sagayama. Orpheus: Automatic composition system considering prosody of japanese lyrics. In *Entertainment Computing–ICEC 2009: 8th International Conference, Paris, France, September 3-5, 2009. Proceedings 8*, pages 309–310. Springer, 2009.
10. Kento Watanabe, Yuichiro Matsubayashi, Fukayama Satoru, Nakano Michiyasu, Goto Masataka, and Inui Kentaro. Automatic lyric generation based on correlation between melody and lyrics. *IPSJ SIG Technical Report*, 2017(16):1–12, 2017.
11. Randal J Leistikow. *Bayesian modeling of musical expectations via maximum entropy stochastic grammars*. Stanford University, 2006.
12. Jan Newmarch and Jan Newmarch. Fluidsynth. *Linux Sound Programming*, pages 351–353, 2017.