Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Special Track on Human-Centred AI

XAI-Lyricist: Improving the Singability of AI-Generated Lyrics with Prosody Explanations

Qihao Liang[1] , **Xichu Ma**[1] , **Finale Doshi-Velez**[2] , **Brian Lim**[1] and **Ye Wang**[1] ,
[1]National University of Singapore
[2]Harvard University
qihao.liang@u.nus.edu, ma_xichu@nus.edu.sg, finale@seas.harvard.edu, {brianlim, wangye}@comp.nus.edu.sg

Abstract

Explaining the singability of lyrics is an important but missing ability of language models (LMs) in song lyrics generation. This ability allows composers to quickly assess if LM-generated lyrics can be sung harmoniously with melodies and helps singers align lyrics with melodies during practice. This paper presents XAI-Lyricist, leveraging *musical prosody* to guide LMs in generating singable lyrics and providing human-understandable singability explanations. We employ a Transformer model to generate lyrics under musical prosody constraints and provide demonstrations of the lyrics' prosody patterns as singability explanations. XAI-Lyricist is evaluated by computational metrics (perplexity, prosody-BLEU) and a human-grounded study (human ratings, average time and number of attempts for singing). Experimental results show that musical prosody can significantly improve the singability of LM-generated lyrics. A controlled study with 14 singers also confirms the usefulness of the provided explanations in helping them to interpret lyrical singability faster than reading plain text lyrics.

1 Introduction

Composing singable lyrics for melodies has long played a central role in music composition. It weaves stories and emotions into melodies that resonate with listeners and singers [Barradas and Sakka, 2022; Brattico *et al.*, 2011; Wang *et al.*, 2022], enhancing the expressiveness of music and adding layers of depth to the auditory experience. To make lyrics composition faster and easier, the notion of automatic lyrics generation is emerging, mainly via Language Models (LMs) [Ma *et al.*, 2021; Malmi *et al.*, 2016; Watanabe *et al.*, 2014; Potash *et al.*, 2015; Sheng *et al.*, 2021]. Despite their impressive performance in lyrics generation, most LMs only output plain text lyrics, making it difficult for people, particularly composers and singers, to understand whether the generated lyrics are suitable for singing melodies.

To address this limitation, we present XAI-Lyricist, leveraging *musical prosody* to guide language models in generating singable lyrics and providing prosody-based singability explanations. In the context of lyrics writing, musical prosody [Palmer and Kelly, 1992; Palmer and Hutchins, 2006; Heffner and Slevc, 2015; Everhardt *et al.*, 2022] generally refers to the alignment of strong and weak beat notes, long and short notes in melodies, with stressed and unstressed syllables, long and short vowels in lyrics, respectively. This alignment harmoniously integrates linguistic and musical elements, resulting in lyrics that are rhythmically adaptable to melodic developments. In linguistic studies, musical prosody is considered fundamentally decisive for the *singability* of lyrics, viz., whether a piece of lyrics is suitable for singing [Khoshsaligheh and Ameri, 2016; Güven, 2019; Franzon, 2008]. Psychological studies have further shown the importance of musical prosody in helping people better sing and comprehend lyrical content [Gordon *et al.*, 2011; Sahasrabuddhe, 2023], enhancing their linguistic and music skills [Patel and Iversen, 2007; Jansen *et al.*, 2023], and helping children develop speech abilities [Caccia and Lorusso, 2021]. These findings support the need and reasonableness to communicate musical prosody as singability explanations for LM-generated lyrics.

Inspired by these insights, we argue that a singable piece of song lyrics should share a similar prosodic structure with its corresponding melodies (or expert-composed lyrics). To implement this, we use an encoder-decoder Transformer [Vaswani *et al.*, 2017; Lewis *et al.*, 2020] as the foundation model of XAI-Lyricist. We extend the Transformer as a multi-target language model, which encodes a prosody template as conditional input, and generates multiple target sequences: (1) lyrics; (2) the number of syllables in each lyric word; (3) the strength and (4) length of each lyric word. Besides presenting the output lyrics to people, XAI-Lyricist also generates demonstrations of lyrics' prosody patterns as singability explanations. A demonstration visualises the correspondence between melody notes and lyric words, rendering notes in different colours, and words in lower and upper case based on musical prosody. Figure 1 displays an example demonstration based on "*Hey Jude*" by The Beatles.

We used objective metrics and a human-subjects study to evaluate XAI-Lyricist, focusing on the singability of generated lyrics, and the usefulness of explanations. For objective metrics, we employed perplexity [Jelinek *et al.*, 2005] and BLEU [Papineni *et al.*, 2002], two established benchmarks in natural language processing [Chen *et al.*, 2008]. We ad-

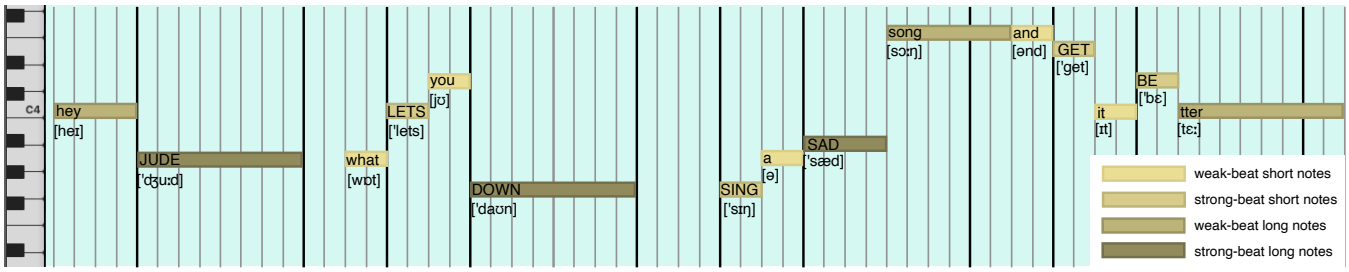

Figure 1: An example demonstration based on "*Hey Jude*" by The Beatles. The rectangles represent notes in melodies, with different colours indicating different types of notes (detailed in the legend). Vertical lines represent beats in music, with strong beats highlighted by bold black lines. Words with stressed syllables are capitalised and have a [ˈ] marker in their phonetic symbols (e.g., lets [ˈlets], sing [ˈsɪŋ]). Words with long vowels have a [ː] in their phonetic symbols, e.g., Jude [ˈʤuːd], song [ˈsɔːŋ]; or have diphthongs, e.g., down [ˈdaʊn], sad [ˈsæd].

justed BLEU to prosody-BLEU as an indicator of singability, which compares the similarity between melody rhythm and lyrics' prosody (detailed in subsection 4.4). For the human-subjects study, we invited 14 participants with musical expertise. They were required to try singing melody sentences with given lyrics and rate the singability of each lyric sentence. Besides their subjective ratings, we also recorded the average time and number of attempts participants needed to sing each lyric sample, as objective measurements of singability.

The results show that incorporating musical prosody in a language model can improve its performance in singable lyrics generation, with significant improvements in the metrics mentioned above. Through the human-subjects study, we also demonstrate that musical prosody can faithfully explain the singability of LM-generated lyrics and is considered useful by participants. Moreover, subjective responses from the post-study interview suggest that communicating prosody-based explanations can help singers understand and judge the singability of lyrics faster, even enabling some of them to recognise unsingable lyrics and edit them to be singable. In summary, our main contributions include:
(1) XAI-Lyricist framework for singable lyrics generation.
(2) The provision of musical prosody-based explanations to help people to interpret the singability of song lyrics.
(3) The evaluation of lyrics' singability and the usefulness of prosody explanations.

2 Related Work

This study uses musical prosody to improve and explain the singability of LM-generated song lyrics. We thus investigated previous studies and categorised them as follows:

2.1 Musical Prosody and Singability

In the context of lyrics writing, musical prosody generally refers to the alignment between lyrics' prosody and melodies' rhythm [Palmer and Hutchins, 2006]. To ensure a suitable lyric setting, human lyricists place stressed/unstressed syllables at strong/weak beats in melodies; long/short vowels at notes with long/short duration. The importance of musical prosody for singability has been discussed in psychological and linguistic studies. [Güven, 2019] proposes a hierarchical linguistic model of lyrical singability, where prosodic constraints have the most significant impact on singability among other layers. [Khoshsaligheh and Ameri, 2016] conclude that the prosodic match between lyrics and melodies is indispensable for producing a singable version of a song. [Franzon, 2008] suggests that the prosodic and poetic match suffices to make lyrics appear singable. Similarly, [Sogunro, 2022] states that the absence of musical prosody would make it technically impossible to sing lyrics. These findings suggest that the lack of musical prosody in automatic lyrics generation can be a research gap.

2.2 Automatic Lyrics Generation

Recent advances in AI have progressed automatic lyrics generation with generative models, such as recurrent neural networks (RNNs) for next-word prediction [Potash *et al.*, 2015; Wu *et al.*, 2019], Seq-GANs [Watanabe *et al.*, 2018; Chen and Lerch, 2020], and Transformer-based models [Zhang *et al.*, 2020; Tian *et al.*, 2023; Sheng *et al.*, 2021]. Moreover, some studies apply domain knowledge to LMs, providing finer-grained controls over the content of lyrics [Chang *et al.*, 2021], such as syllable counts, keywords [Ma *et al.*, 2021], and syllable lengths [Tian *et al.*, 2023]. To the best of our knowledge, most existing LMs only output plain text lyrics without singability explanations, nor have they explicitly used musical prosody to control or explain lyrical singability. These limitations make it difficult for singers to properly align words with notes and for composers to assess if generated lyrics are singable. Therefore, we argue for the need to communicate musical prosody as singability explanations in automatic lyrics generation tasks.

2.3 Explainable AI in Music Computing

Explainable artificial intelligence (XAI) aims to make AI understandable so that it can be useful to non-AI experts. In music computing, XAI can potentially advance music generative models by helping people understand music AI [Yan *et al.*, 2023]. [Bryan-Kinns *et al.*, 2023] explain the latent space of a variational autoencoder-based music generator. This work constrains four dimensions of the space to map four musical attributes and visualising their values. While the visualisation helps people comprehend the internal workings of a generative model, the four proposed features may be incomplete to fully explain music generation due to the intricate nature of music. Furthermore, [Zhao *et al.*, 2019] use an additional LM to generate the reason behind each music recommendation in

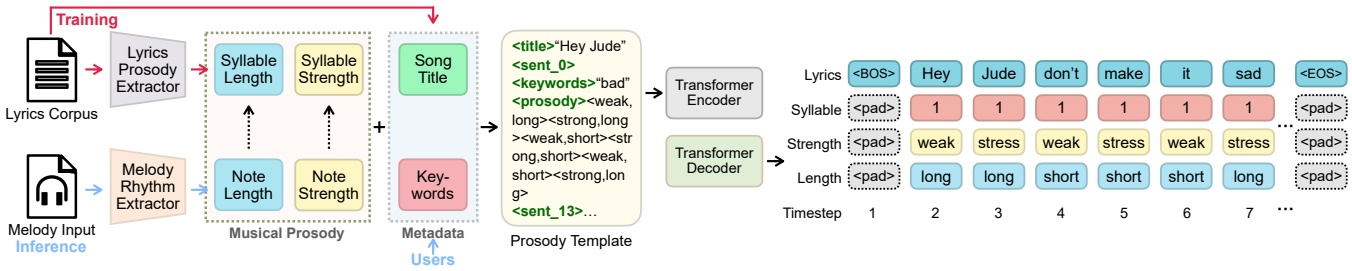

Figure 2: The lyrics generation pipeline of XAI-Lyricist. Overall, the pipeline uses an encoder-decoder Transformer model to generate lyrics from a given prosody template (in the middle). During training (top left corner), the input template comes from the prosody (including syllable length and syllable strength) and metadata extracted from the lyrics corpus. To make the model better aware of musical prosody, we optimise it on four targets (right corner): lyrics, syllable numbers, syllable strengths, and syllable lengths. During inference (bottom left corner), the template comes from the rhythm of a melody and customised metadata. Specifically, syllable length and syllable strength are replaced by note length and note strength, respectively. The model then generates lyrics aligned with the melody rhythm constraints.

a human-like tone. However, LM-generated reasons can be less grounded predictions than faithful explanations. Another study [Wang *et al.*, 2020] uses two music features, chord and texture, to control the generated music. This system generates a song from a given chord progression, enabling users to interpret the song by examining its chords. However, such music theory-based explanations are often complex and can overwhelm people. Furthermore, the intersection between explainable AI and automatic lyrics generation remains understudied. We are thus motivated to use musical prosody, a clearly defined concept in music psychology and linguistics, to develop a more explainable lyrics generation system.

3 Technical Approach

3.1 System Overview

Figure 2 shows the lyrics generation pipeline of XAI-Lyricist. Overall, this pipeline uses an encoder-decoder Transformer [Vaswani *et al.*, 2017], which generates lyrics from prosody templates. Given the lack of paired melody-lyric data available with accurate rhythm information, we resorted to a lyrics corpus for training, using the lengths (long/short) and strengths (strong/weak) of syllables in lyric words to create prosody templates (detailed in subsection 3.3).

To make the model better aware of musical prosody, we optimised the model on three prosody-related targets in addition to lyrics: (1) the number of syllables in each lyric word; (2) syllable strength (if the word has stressed syllables); (3) syllable length (if the word has long syllables). During inference, the model can take the rhythm of a melody to create the input prosody template, with syllable lengths and syllable strengths replaced by note lengths and note strengths, respectively. The model then generates lyrics whose prosodic structure aligns with the melody's rhythmic pattern, creating musical prosody. We also use the prosody of generated lyrics as singability explanations to people.

3.2 Musical Prosody

Melody Rhythm and Lyric Prosody

In music theory, beat refers to the basic unit of time. However, not all beats are equally important: some beats are considered stronger (strong beats), while others are weaker (weak beats). The rhythm of music is represented by a repeating sequence of strong/weak beats and long/short notes, analogous to lyrics whose prosody mainly comes from stressed/weak and long/short syllables. Musical prosody associates the rhythm of melodies with the prosody of lyrics. In this paper, we refer to the term *prosodic lyrics* as lyrics whose prosody is identical to the rhythm of their accompanying melodies.

Taxonomy of Notes and Syllables

We consider two attributes shared by melody notes and lyric syllables, *strength* and *length*, and categorise them as follows:

- **Strong-beat notes and weak-beat notes** refer to melody notes beginning at strong and weak beats, respectively. In 4/4 time music, the first and third beats are stronger than the second and fourth beats.
- **Long notes and short notes:** long notes are notes with relatively longer duration. Based on [Nichols *et al.*, 2009], we define long notes as those whose duration is greater than the average duration of all notes in a melody phrase, with all non-long notes being short notes.
- **Stressed syllables and weak syllables:** stressed syllables are accented when pronounced and are marked by special symbols (['], [ˌ]) in the international phonetic alphabet (IPA)[1]; Weak syllables (a.k.a. unstressed syllables) do not have special markers in their IPA notations.
- **Long syllables and short syllables:** Long syllables have long vowels (marked by [ː] in IPA) or diphthongs. A diphthong is a syllable with two vowel sounds, for example, sad ['sæd], hey [heɪ], etc. Syllables without long vowels or diphthongs are short syllables.

Taxonomy of Musical Prosody

Based on [Nichols *et al.*, 2009] and our proposed taxonomy of notes and syllables, we define the following two types of musical prosody that make lyrics singable to melodies.

- **Strength Alignment:** Strong/weak-beat notes should concur with stressed/weak syllables, respectively.
- **Length Alignment:** Long/short notes should concur with long/short syllables, respectively.

[1]https://www.internationalphoneticassociation.org/content/full-ipa-chart

3.3 Data Representation

Prosody Representation

We use strength and length symbols to represent the rhythm of melodies and the prosody of lyrics. Melody notes and lyric syllables with similar attributes are notated by the same symbols, as shown in Table 1. For example, a strong-beat note and a stressed syllable are both notated by a `<strong>`. Similarly, a short note and a short syllable are marked by a `<short>`, by analogy.

Attributes	Symbols	Melody Elements	Lyrics Elements
Strength	`<strong>`	strong-beat notes	stressed syllables
	`<weak>`	weak-beat notes	weak syllables
Length	`<long>`	long notes	long syllables
	`<short>`	short notes	short syllables

Table 1: Strength and length symbols in the prosody representation.

For each melody note, we refer to its starting beat and duration to obtain its strength and length attributes. For each syllable in lyrics, we use a Python package named Prosodic[2] to query its IPA symbol and obtain its strength and length attributes (more details are included in section 1 of the supplementary document[3]). To reduce the length of prosody representations, we compress each pair of strength and length symbols into one compound word, following [Hsiao *et al.*, 2021]. For example, a strong-beat long note (or a stressed long syllable) is notated by a `<strong,long>`; a weak-beat short note (or a weak short syllable) is a `<weak,short>`.

Prosody Templates

A prosody template for a song includes the title and metadata of the song, followed by the prosody representation of each sentence in this song. Specifically, each template starts with a `<title>`, followed by the song name. Then, a `<sent_x>` is used to mark the beginning of a sentence, with `x` being the order of the sentence. After each `<sent_x>`, the keywords and prosody representation of the sentence are appended, prefixed by a `<keywords>` and a `<prosody>`, respectively. An example template is shown in the middle of Figure 2. Before being input into the Transformer model, each symbol in the template is tokenised and embedded as a vector. For compound words, the strength and length symbols are embedded as two vectors, which are then concatenated and linearly projected to the embedding size of the model, as in [Liang and Wang, 2024; Zhang *et al.*, 2023].

Lyrics and Other Target Sequences

The decoder end generates four target sequences, each using a different vocabulary. The lyrics target uses an English word vocabulary with punctuation marks and special beginning (`<BOS>`) and end of sequence (`<EOS>`) symbols. The syllable strength and length sequences adopt symbols in Table 1 as their vocabulary. For the syllable number target, we count the maximum number of syllables per sentence in our lyrics dataset and initiate a new vocabulary with 20 symbols formatted as `<syllable_s>`, where `s` ranges from `1` to `20`. The syllable number, strength and length vocabularies also include an additional `<pad>` symbol to accommodate special non-word symbols in the lyrics target, e.g. timestep 1 in Figure 2. At each input timestep, the embeddings of four symbols from four target sequences are concatenated and linearly projected to the model embedding size.

[2]https://pypi.org/project/prosodic/

[3]Codes and supplementary materials are available at: https://github.com/lqhac/XAI-Lyricist

3.4 Generating Lyrics and Explanations

Sampling Lyric Words and Prosody Symbols

The decoding uses top-k temperature sampling. At each decoding step, we first sample a symbol from the lyrics vocabulary. If the sampled symbol is a word, then the other three prosody-related symbols of this word are sampled. If the sampled symbol is not a word (e.g. a comma), the sampling of the other three symbols is skipped. The decoding ends once an `<EOS>` is sampled from the lyrics vocabulary.

Prosody Correction

Since the prosody symbols inferred by LMs may not be completely accurate, we also refer to the IPA of each sampled lyric word for prosody correction. Namely, inferred prosody should be replaced by standard IPA prosody if they conflict.

Demonstrations of Singability Explanations

We visualise the melody, generated lyrics, and strength/length symbols as a demonstration of singability explanations. An example is shown in Figure 1. First, we align all lyric words with melody notes using the two types of musical prosody in section 3.2. Then, all melody notes are rendered as rectangles on a piano roll, with their colours/widths indicating their strengths/lengths, respectively. For lyric words, all stressed syllables are capitalised and annotated on their aligned notes.

4 Experimental Setup

The following experiments aim to address two research questions (RQs) about musical prosody:

RQ1: Can musical prosody guide a language model to generate more singable lyrics?

RQ2: Is it useful to communicate musical prosody as singability explanations to people?

To this end, we implemented three LM-based lyrics generation models with/without prosody constraints and explanations. Using the lyrical content generated by these models, we conducted objective evaluations and a human-subjects study. This section describes the dataset, baselines, model configurations, metrics, and the setup of the human-subjects study. The results are reported in section 5.

4.1 Models

We implemented the following three models in the experiment. To allow for direct comparisons, we adopted the same lyrics vocabulary sfor all models.

- **Vanilla**, which directly generates lyrics from melodies without prosody constraints. We followed the methodology of [Sheng *et al.*, 2021] to implement this baseline.
- **Prosody**, which generates only prosodic lyrics but does not provide singability explanations.

- **Explainable Prosody** (XAI-Lyricist), which generates both prosodic lyrics and singability explanations.

Both Vanilla and Prosody generate only plain-text lyrics, while Explainable Prosody additionally presents demonstrations of singability explanations as in Figure 1.

4.2 Datasets

For Vanilla, we followed [Sheng *et al.*, 2021] and used their datasets for training and validation. For Prosody and Explainable Prosody, we used pop English song lyrics collected from a genre classification dataset[4] for training and validation. We excluded all non-English samples and used the Prosodic package to query the IPA of each lyric sample to obtain prosody templates. All samples with out-of-vocabulary words were filtered. The final dataset after cleansing includes 101,120 full song lyrics. For the test data in the human-subjects study, we randomly collected 210 sentences with low popularity on NetEase Music[5] as the test database. The selected sentences were manually transcribed into MIDI melodies. It is confirmed that none of the participants knew the selected songs before the human-subjects study.

4.3 Model Configuration and Training

All models employ a 6-layer encoder-decoder Transformer [Vaswani *et al.*, 2017; Lewis *et al.*, 2020], with 8 attention heads and a hidden size of 768. The feed-forward channel is 2048; the dropout rate is 0.3. The model was trained on three NVIDIA RTX A5000 GPUs. For the learning rate, we used Adam optimiser with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-5}$. For the loss function, Vanilla and Prosody were optimised only on the lyrics' cross-entropy loss, while Explainable Prosody (XAI-Lyricist) employed the summation of cross-entropy (CE) losses for four target sequences:

$$\mathcal{L}_{\text{train}} = \text{CE}_{\text{lyrics}} + \text{CE}_{\text{syllablles}} + \text{CE}_{\text{strengths}} + \text{CE}_{\text{lengths}} \quad (1)$$

4.4 Metrics

- **Perplexity**: Perplexity is a benchmark to assess language models [Jelinek *et al.*, 2005]. A lower perplexity can indicate higher fluency of LM-generated sentences.
- **Prosody-BLEU**: We adjust BLEU [Papineni *et al.*, 2002] to prosody-BLEU as an indicator of singability. It quantitatively measures the similarity between lyrics' prosody and melody rhythm under the prosody representation shown in Table 1.
- **Human Rating** reflects the degree to which a participant subjectively feels a lyric sentence is suitable to be sung with a melody, on a 5-point scale.
- **Average Time (T)** refers to the time taken by a participant to successfully align and sing a lyric sentence with its melody, normalised by the length of the lyric sentence. We suppose that, averagely, participants tend to take longer/shorter time to align less/more singable lyrics to melodies. T can be formulated as:

$$T(L) = \log(\frac{t}{|L|}) \quad (2)$$

where L denotes a lyric sentence; t denotes the actual time taken to align L to its accompanying melody; $|L|$ is the number of syllables in L. The logarithm scales the magnitude for better data analyses and visualisation.

- **Average Attempt Number (A)** is the number of attempts a participant needs to successfully align a lyric sentence to melodies, normalised by the length of the lyric sentence. An attempt refers to a one-time trial to sing out the lyrics that a singer aligns with the melody in a subjectively natural and fluent feeling. For example, if a participant finishes singing at one time without pausing or repeating the same parts, the attempt number is 1. If a participant pauses and repeats an identical part during singing once (e.g. "*Find my... Find my way ...*"), the attempt number is incremented by 1. A is formulated as:

$$A(L) = \frac{n_{att}}{|L|} \quad (3)$$

where L denotes a lyric sentence; n_{att} denotes the actual number of attempts to align L with its accompanying melody; $|L|$ is the number of syllables in L.

4.5 Human-subjects Study

Participants

We recruited 14 participants, each with at least 7 years of experience in formal music training (with degrees or certifications in music and lyrics composition, performance, singing, etc.). All participants have prior experience in singing English songs. Each participant received compensation equivalent to about US$14.89 after their successful completion of the study. Breaks were permitted during the study to counterbalance potential fatigue-related side effects.

Apparatus

Each participant was presented with 99 shuffled sentences randomly selected from the test database. We used sentence-level experimental design because: (1) it reduces the workload and does not overwhelm participants in each singing trial; (2) there are often repetitions between consecutive sentences in a song [Lloyd, 2020]; (3) with multiple sentences, participants tend to infer next sentences during singing [Hansen *et al.*, 2021], which can affect the study. The 99 samples were generated by the three models, each model taking up 1/3. The sentence length of each model ranges from 2 to 11 syllables (mean=6.17; median=6; SD=2.09). The lyrics and melody for each sentence were displayed on a digital audio workstation for playback and visualisation. Figure 1 is an example demonstration of Explainable Prosody. For the other two models without explanations, the visualisations only included notes on the piano roll, and participants read plain text lyrics annotated beside the piano roll (more examples are shown in section 2 of the supplementary). The duration of the experiment ranged from 45 to 90 minutes.

Procedure

1. **Participant Consent** Participants first read the introduction and sign the consent form with IRB approval. The study was audio recorded after participants' consent.

[4]https://www.kaggle.com/datasets/mateibejan/multilingual-lyrics-for-genre-classification

[5]https://music.163.com/

2. **Pre-study Tutorial** We offered a pre-study tutorial with 12 lyric samples to help participants understand the task. The 12 samples are equally generated by three baselines, but excluded from the 99 sentences for the formal study. Participants were instructed to use the samples to sing the melodies. Then, they needed to independently and correctly complete 6 new test samples meeting the pre-validated standard alignment, before the formal study.
3. **Melody Familiarisation** Since participants' unfamiliarity with melodies can affect their singing and judgement of singability, participants were required to first familiarise themselves with the melody by listening to its piano rendition before reading the lyrical content. They were allowed to play the rendition an unlimited number of times, until they could hum the melody at least three times without pauses in between or replaying it.
4. **Singing Trial** After the familiarisation, participants read the lyrics (and explanations). They needed to try to align the lyrics with the melody and sing the lyrics out. Participants were explicitly told not to care about how well they could sing. Namely, they needed not take more trials to improve singing techniques, singing voice quality, music intonation, etc. They could stop once they felt they had already made a satisfactory lyric-melody alignment, before which unlimited attempts were allowed.
5. **Singability Rating** After singing of each lyric sample, participants rated the singability of the sample on a 5-point scale. They were told to consider only how suitable they felt the lyrics were to be sung with the melody. Steps 3-5 were repeated for all 99 samples before step 6.
6. **Post-study Briefing** We finally conducted a post-study briefing for participants, clarifying the study's objectives and necessary details. Participants were encouraged to ask questions and share their thoughts about the study.

5 Results

5.1 RQ1: Can Musical Prosody Guide a Language Model To Generate More Singable Lyrics?

Musical Prosody Improves the Singability of LM-Generated Lyrics

From Table 2, we can observe that compared to the Vanilla baseline, Prosody and Explainable Singability have lower perplexity, higher prosody-BLEU and human singability ratings. These indicate that infusing musical prosody helps LMs generate more fluent and singable lyrics for melodies.

Metrics	Vanilla	Prosody	Expl. Prosody
Perplexity↓	37.50	10.81	**4.76**
Prosody-BLEU↑	0.13± 0.020	0.92±0.022	**0.98±0.014**
Human Rating↑	1.68± 0.084	4.37± 0.066	**4.66± 0.056**

Table 2: The scores on computational metrics and human ratings. Expl. Prosody is short for the Explainable Prosody model.

Similarly, during the human-subjects study, all participants said that they encountered some lyric samples "incoherent" (P1, P3), "strange" (P2, P4, P6), "did not make sense at all" (P7). We observed that these problematic samples all came from the Vanilla model, with grammar issues (e.g. "*Stains the planets in the sea*"), odd coinages (e.g. "*burn the tears I arteaths*"), word repetitions (e.g. "*I fight fight forever. Chance I I could live*"), etc.

We suppose that these problems can be attributed to the "vagueness" of Vanilla. The Vanilla model uses a sequence of pitch-duration pairs (e.g. `<G3><1/16>`) to represent a melody and directly generates lyrics from this melody representation. However, the connection between pitch-duration pairs and lyric words is often vague, as each pitch and duration can irregularly correspond to many different words. Therefore, it can be confusing for LMs to converge well on such melody-lyric data, resulting in high perplexity, less singable, and even problematic lyrics as shown above. In contrast, Prosody and Explainable Prosody use prosody templates that explicitly relate lyrics to their rhythmic attributes (prosody). This strategy refines the original melody representation, guiding LMs to learn the rhythmic pattern of lyrics and effectively use the prosody information during inference.

Musical Prosody Faithfully Explains Lyrics Singability

To investigate the faithfulness of musical prosody to lyrics singability, we perform a correlation test between prosody-BLEU and human singability rating. The test results in a Pearson correlation coefficient $r = \mathbf{0.77}$ and $P < .001$, indicating a strong correlation. This suggests that musical prosody concurs with human perception of lyrical singability and constitutes a faithful singability explanation.

5.2 RQ2: Is It Useful To Communicate Musical Prosody as Singability Explanations to People?

Singability Explanations Help Singers Quickly Align Lyrics to Melodies

We analyse the average time (T) and average attempt number (A) that participants needed to successfully complete each singing trial. Figure 3 shows the aggregated T and A from all participants. Both T and A exhibit a decreasing trend from Vanilla to Prosody to Explainable Prosody, indicating that participants tend to take shorter/longer time and fewer/more trials to sing more/less prosodic lyrics.

For statistical evidence, a Friedman test was performed to compare T and A across three models. The results indicate a significant difference, with $Fr_T = 808.13$, $Fr_A = 355.10$, $P < .001$. Then, a Nemenyi post hoc test was conducted with holm correction applied to pair-wise comparisons between models. The results show significant differences between all model pairs as detailed in Table 3.

Model Pairs		T			A		
M_1	M_2	T_{M_1}	T_{M_2}	P	A_{M_1}	A_{M_2}	P
Prosody	Expl. Prosody	-0.17	**-1.13**	$< .001$	0.26	**0.19**	$< .001$
Vanilla	Expl. Prosody	0.93	**-1.13**	$< .001$	0.62	**0.19**	$< .001$
Prosody	Vanilla	**-0.17**	0.93	$< .001$	**0.26**	0.62	$< .001$

Table 3: The Nemenyi post hoc test analysis of Average Time (T) and Average Attempt Number (A) for model pairs. Expl. Prosody is short for the Explainable Prosody (viz., XAI-Lyricist) model.

We also summarise some findings from participants' responses in the post-study briefing.

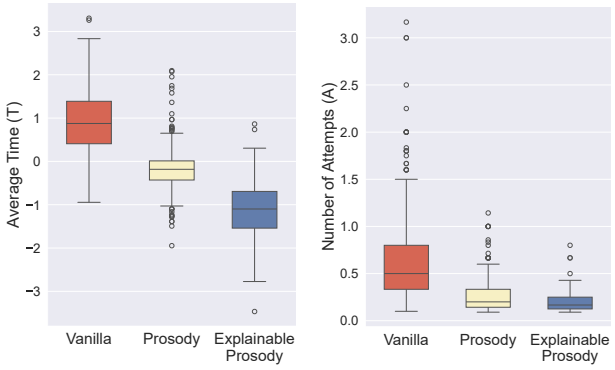

Figure 3: The visualisation of aggregataed Average Time (T) and Average Attempt Number (A) for each model.

Participants Can Feel the Connection Between Prosody and Singability but Cannot Clearly Define It

Before being told the definition of prosody, all participants stated that they could feel something related to singability.

> "I cannot tell exactly why, but I just feel that the arrangement of words in these sentences seems more consonant with melodies." – P2

> "I think there are some rhythmic features that make the lyrics sound natural." – P1

> "I find that capitalised words always appear at strong beats." – P8

However, when asked to summarise their descriptions into a clear definition, only 2 participants with lyrics writing experience precisely used the term *prosody*. The feedback shows the necessity of communicating prosody explanations to people, especially those without lyrics writing expertise.

Participants Confirm the Usefulness of Explanations

All participants agreed that the explanations were helpful, though they could use explanations for different purposes.

> "These explanations did help me quickly detect the rhythmic rises and falls. I needed not spend more time studying them before singing." – P1, singer

> "Besides singing, I love reading these explanations to quickly check whether the lyrics are rhythmically aligned to melodies." – P14, songwriter

P4 could also use the explanation to slightly adjust unsingable lyrics to be singable.

> "If you asked me why I felt this was not singable, I could not explicitly tell the reason without these explanations. But with musical prosody, I know that a stressed syllable is missing here, and that I can insert one to make this sentence singable to me." – P4, singer

Moreover, all participants said that they dared try singing lyrics directly with explanations, instead of repeatedly reading and singing as they did without explanations. We analysed their attempt numbers (A) before being normalised by the sentence length $|L|$ and found that 85.71% (396 out of 462) test sentences were completed in one trial with Explainable Prosody (viz., XAI-Lyricist), in contrast to 27.06% (125 out of 462) with Vanilla and 58.44% (270 out of 462) with Prosody. These findings all support the usefulness of communicating musical prosody as singability explanations.

Individual Preferences for Explanation Demonstrations Can Differ

We also found that the individual preferences for presenting explanations often differed. Generally, the preference depends on the background and usage scenarios of participants. For example, participants with singing expertise said they preferred music sheet-styled visualisations:

> "I always read sheets when singing, so it'd be better if you wrote everything as sheets." – P1, singer

> "Maybe you could give us more options, such as sheets, other than the piano roll interface, cuz different people have different habits. " – P5, singer

Participants with songwriting expertise said they felt comfortable with both sheet music and piano-roll visualisations as in Figure 1. However, they added that they wanted to know more musicological details (e.g. the definition of prosody).

> "I felt it ok to see the piano roll, but knowing how these explanations are defined could further convince me." – P7, a songwriter with a degree in musicology

By contrast, participants with less expertise in songwriting mainly prefer the piano-roll interface:

> "I don't read western sheet music, so I prefer piano roll because it intuitively presents the keyboard, pitches, durations and lyrics." – P6

These findings suggest the importance of considering users' preferences and backgrounds in practical applications.

6 Conclusion

In this paper, we present XAI-Lyricist, which uses musical prosody to guide language models in generating singable song lyrics and providing singability explanations. We used both computational metrics and a human-subjects study to prove the effectiveness of musical prosody in (1) improving the singability of LM-generated lyrics; (2) faithfully explaining lyrics' singability, demonstrating the usefulness of communicating prosody-based explanations. XAI-Lyricist also offers valuable insights into the singability explanations of LM-generated lyrics, potentially fostering a more efficient and effective collaboration between humans and AI in the domain of lyrics creation. In the future, we aim to explore other finer-grained domain knowledge to improve machine lyric generators, meanwhile exploring the pragmatic applications that cater to both experts and amateurs.

Ethical Statement

This work involved human subjects in its research. All ethical and experimental procedures have been approved by the Departmental Ethics Review Committee (DERC), National University of Singapore.

Acknowledgments

We thank all anonymous reviewers for their valuable input. This work was funded by the Ministry of Education, Singapore, under research grant MOE-T2EP20120-0012. Finale Doshi-Velez was supported by the National Science Foundation under Grant No. IIS-1750358. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[Barradas and Sakka, 2022] Gonçalo T Barradas and Laura S Sakka. When words matter: A cross-cultural perspective on lyrics and their relationship to musical emotions. *Psychology of Music*, 50(2):650–669, 2022.

[Brattico *et al.*, 2011] Elvira Brattico, Vinoo Alluri, Brigitte Bogert, Thomas Jacobsen, Nuutti Vartiainen, Sirke Nieminen, and Mari Tervaniemi. A functional mri study of happy and sad emotions in music with and without lyrics. *Frontiers in psychology*, 2:308, 2011.

[Bryan-Kinns *et al.*, 2023] Nick Bryan-Kinns, Berker Banar, Corey Ford, Courtney N Reed, Yixiao Zhang, Simon Colton, and Jack Armitage. Exploring xai for the arts: Explaining latent space in generative music. *arXiv preprint arXiv:2308.05496*, 2023.

[Caccia and Lorusso, 2021] Martina Caccia and Maria Luisa Lorusso. The processing of rhythmic structures in music and prosody by children with developmental dyslexia and developmental language disorder. *Developmental science*, 24(1):e12981, 2021.

[Chang *et al.*, 2021] Jia-Wei Chang, Jason C. Hung, and Kuan-Cheng Lin. Singability-enhanced lyric generator with music style transfer. *Computer Communications*, 168:33–53, 2021.

[Chen and Lerch, 2020] Yihao Chen and Alexander Lerch. Melody-conditioned lyrics generation with seqgans. *2020 IEEE International Symposium on Multimedia (ISM)*, pages 189–196, 2020.

[Chen *et al.*, 2008] Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. Evaluation Metrics For Language Models. 1 2008.

[Everhardt *et al.*, 2022] Marita K Everhardt, Anastasios Sarampalis, Matt Coler, Deniz Başkent, and Wander Lowie. Speech prosody: the musical, magical quality of speech. *Front. Young Minds*, 10(698575):10–3389, 2022.

[Franzon, 2008] Johan Franzon. Choices in song translation: Singability in print, subtitles and sung performance. *The Translator*, 14(2):373–399, 2008.

[Gordon *et al.*, 2011] Reyna L Gordon, Cyrille L Magne, and Edward W Large. Eeg correlates of song prosody: A new look at the relationship between linguistic and musical rhythm. *Frontiers in psychology*, 2:352, 2011.

[Güven, 2019] Mine Güven. Why "sway" again? prosodic constraints and singability in song (re)translation. *Studies from a Retranslation Culture: The Turkish Context*, pages 177–194, 2019.

[Hansen *et al.*, 2021] Niels Chr Hansen, Haley E Kragness, Peter Vuust, Laurel Trainor, and Marcus T Pearce. Predictive uncertainty underlies auditory boundary perception. *Psychological science*, 32(9):1416–1425, 2021.

[Heffner and Slevc, 2015] Christopher C Heffner and L Robert Slevc. Prosodic structure as a parallel to musical structure. *Frontiers in psychology*, 6:1962, 2015.

[Hsiao *et al.*, 2021] Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):178–186, May 2021.

[Jansen *et al.*, 2023] Nelleke Jansen, Hanneke Loerts, Eleanor Harding, Deniz Baskent, and Wander Lowie. The influence of musical abilities on the processing of second language prosody: An eye-tracking study. *The Journal of the Acoustical Society of America*, 153(3_supplement):A157–A157, 2023.

[Jelinek *et al.*, 2005] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63, 08 2005.

[Khoshsaligheh and Ameri, 2016] Masood Khoshsaligheh and Saeed Ameri. Exploring the singability of songs in a monster in paris dubbed into persian. *Asia Pacific Translation and Intercultural Studies*, 3(1):76–90, 2016.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[Liang and Wang, 2024] Qihao Liang and Ye Wang. Drawlody: Sketch-based melody creation with enhanced usability and interpretability. *IEEE Transactions on Multimedia*, 26:7074–7088, 2024.

[Lloyd, 2020] Dan Lloyd. The musical structure of time in the brain: Repetition, rhythm, and harmony in fmri during rest and passive movie viewing. *Frontiers in computational neuroscience*, 13:98, 2020.

[Ma *et al.*, 2021] Xichu Ma, Ye Wang, Min-Yen Kan, and Wee Sun Lee. Ai-lyricist: Generating music and vocabulary constrained lyrics. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1002–1011, 2021.

[Malmi *et al.*, 2016] Eric Malmi, Pyry Takala, Hannu Toivonen, Tapani Raiko, and Aristides Gionis. Dopelearning: A computational approach to rap lyrics generation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 195–204, 2016.

[Nichols *et al.*, 2009] Eric Nichols, Dan Morris, Sumit Basu, and Christopher Raphael. Relationships between lyrics and melody in popular music. In *ISMIR 2009-Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 471–476, 2009.

[Palmer and Hutchins, 2006] Caroline Palmer and Sean Hutchins. What is musical prosody? *Psychology of learning and motivation*, 46:245–278, 2006.

[Palmer and Kelly, 1992] Caroline Palmer and Michael H Kelly. Linguistic prosody and musical meter in song. *Journal of memory and language*, 31(4):525–542, 1992.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[Patel and Iversen, 2007] Aniruddh D Patel and John R Iversen. The linguistic benefits of musical abilities. *Trends in cognitive sciences*, 11(9):369–372, 2007.

[Potash *et al.*, 2015] Peter Potash, Alexey Romanov, and Anna Rumshisky. Ghostwriter: Using an lstm for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1919–1924, 2015.

[Sahasrabuddhe, 2023] Hari Sahasrabuddhe. Role of prosody in music meaning. In *Computer Assisted Music and Dramatics: Possibilities and Challenges*, pages 71–76. Springer, 2023.

[Sheng *et al.*, 2021] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13798–13805, May 2021.

[Sogunro, 2022] Bolanle O Sogunro. Phonological and sociolinguistic challenges of translating yorùbá play, and game songs to singable english for children. *Yoruba Studies Review*, 7(1):1–14, 2022.

[Tian *et al.*, 2023] Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Tagyoung Chung, Jing Huang, and Nanyun Peng. Unsupervised melody-to-lyrics generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9235–9254, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[Wang *et al.*, 2020] Ziyu Wang, Dingsu Wang, Yixiao Zhang, and Gus Xia. Learning interpretable representation for controllable polyphonic music generation. In *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, pages 662–669, 2020.

[Wang *et al.*, 2022] Zihao Wang, Kejun Zhang, Yuxing Wang, Chen Zhang, Qihao Liang, Pengfei Yu, Yongsheng Feng, Wenbo Liu, Yikai Wang, Yuntao Bao, et al. Songdriver: Real-time music accompaniment generation without logical latency nor exposure bias. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1057–1067, 2022.

[Watanabe *et al.*, 2014] Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, and Masataka Goto. Modeling structural topic transitions for automatic lyrics generation. In *Proceedings of the 28th Pacific Asia conference on language, information and computing*, pages 422–431, 2014.

[Watanabe *et al.*, 2018] Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. A melody-conditioned lyrics language model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 163–172, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[Wu *et al.*, 2019] Xing Wu, Zhikang Du, Yike Guo, and Hamido Fujita. Hierarchical attention based long short-term memory for chinese lyric generation. *Applied Intelligence*, 49:44–52, 2019.

[Yan *et al.*, 2023] Zihan Yan, Chunxu Yang, Qihao Liang, and Xiang 'Anthony' Chen. Xcreation: A graph-based crossmodal generative creativity support tool. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.

[Zhang *et al.*, 2020] Rongsheng Zhang, Xiaoxi Mao, Le Li, Lin Jiang, Lin Chen, Zhiwei Hu, Yadong Xi, Changjie Fan, and Minlie Huang. Youling: an AI-assisted lyrics creation system. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 85–91, Online, October 2020. Association for Computational Linguistics.

[Zhang *et al.*, 2023] Kejun Zhang, Xinda Wu, Tieyao Zhang, Zhijie Huang, Xu Tan, Qihao Liang, Songruoyao Wu, and Lingyun Sun. Wuyun: exploring hierarchical skeleton-guided melody generation using knowledge-enhanced deep learning. *arXiv preprint arXiv:2301.04488*, 2023.

[Zhao *et al.*, 2019] Guoshuai Zhao, Hao Fu, Ruihua Song, Tetsuya Sakai, Zhongxia Chen, Xing Xie, and Xueming Qian. Personalized reason generation for explainable song recommendation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(4):1–21, 2019.