

# 入力文章の内容に沿った新たな歌詞を生成する 作詞支援システムと剽窃リスクを下げる歌詞生成手法

渡邊研斗 後藤真孝  
産業技術総合研究所

{kento.watanabe,m.goto}@aist.go.jp

## 概要

本研究の目的は、ユーザが歌詞に書きたい内容を文章として入力した時、その入力文章の内容に沿いつつ字面が似ていない歌詞を生成することで、ユーザに新たな作詞の発想を与える作詞支援システムを開発することである。この目的達成のために、(a) 既存の画像生成技術を用いることで、「入力文章の字面に関する情報を取り除きつつ入力の内容を表した画像」を生成し、(b) 本研究で提案する画像の内容を表した歌詞を生成するエンコーダ・デコーダを用いることで、「入力文章の内容に沿いつつ字面が似ていない歌詞」を生成する。更に、自動歌詞生成技術を用いた創作活動を円滑にするために、剽窃リスクを下げる歌詞生成手法を提案する。

## 1 はじめに

歌詞情報処理 [1] の研究として、自動歌詞生成技術を活用した作詞支援システムが提案されてきた [2, 3, 4, 5, 6, 7]。Watanabe ら [4, 5] のシステムは、ユーザが選択したトピックに沿った歌詞を生成できるが、選択可能なトピックは限られているためユーザの自由度は低く、生成される歌詞の作風が似てしまう問題点がある。Oliveira ら [6, 7] のシステムは、ユーザが入力した単語を含む歌詞を生成できるが、入力単語が生成歌詞に必ず含まれるため、入力単語以外の単語で表現した歌詞が生成されにくい。例えば、単語「海岸」を入力しても「ビーチ」や「海辺」等の同じ意味だが別表現の単語は生成されにくい。

本研究の目的は、「海岸で車を運転する」のような歌詞に書きたい内容を文章として入力した時、「海辺のハイウェイを走りたい [改行] ハンドルを握る君とならば」のような入力文章の内容に沿った歌詞を生成して提示することで、ユーザに発想を与える作詞支援システムを開発することである。本システム



図1 提案する作詞支援システムと歌詞生成手法の概要

の対象ユーザとして、歌詞に書きたい内容は決まっているがどんなフレーズで書けばよいか思いつかない人を想定している。よって、もしシステムが入力文章と字面が似た歌詞を提示してしまうと、ユーザに新たな発想を与えることは難しくなる。我々は「入力の内容に沿いつつ字面が似ていない歌詞を提示することが、ユーザに新たな発想を与える」という考えの元で作詞支援システムを開発する。

この目的達成のために、入力文章の内容に沿いつつ字面が似ていない歌詞を生成する手法を提案する。単純な手法として、文章と歌詞のペアデータを学習したエンコーダ・デコーダを用いて歌詞生成をすれば良いと考えられるが、そのようなペアデータは存在しない。そこで、文章要約技術や機械翻訳技術などを用いて歌詞から文章を自動生成することで、ペアデータを用意する方法が考えられる。しかし、このようにして用意された文章と歌詞のペアは字面が似ているため、このペアデータを学習したエンコーダ・デコーダは入力文章と字面が似た歌詞を生成する可能性がある。

そこで図1のように、(a) 入力文章の内容を保持したまま、入力の字面に関する情報を取り除いた中間表現へ変換した後に、(b) その中間表現の内容に沿った歌詞を生成することで、目的の「入力文章の内容に沿いつつ字面が似ていない歌詞」の生成を可能とする。本研究ではこの「中間表現」として画像を用いる。ここで画像を選択した理由には、Stable Diffusion[8]を始めとした拡散モデルによって入力文章の内容を表した画像生成が可能となったという背

景がある。この技術で生成された画像は、入力文章の内容を表しているだけでなく、入力文章の字面に関する情報が取り除かれた状態であるため、目的の中間表現の条件を満たしている。更に、入力文章の形式（単語・フレーズ・文・段落）に関わらず画像を生成できるため、ユーザの入力形式に依存しない柔軟性の高い作詞支援システムの開発が可能となる。

本研究では、入力文章から画像を生成するために既存の画像生成技術を利用するが、画像から歌詞を生成するエンコーダ・デコーダに関しては新たに学習をする必要がある。このエンコーダ・デコーダを学習するためには、入力文章から生成された画像と歌詞のペアデータを用意できれば理想的だが、上述したようにそのような入力文章データは存在しない。そこで本研究では、画像生成技術を用いて歌詞から画像を生成することで、画像と歌詞のペアデータを用意する。このように画像生成技術を活用すると、目的の歌詞生成が可能となるだけでなく、画像から歌詞を生成するエンコーダ・デコーダの訓練データを用意することも可能となる。

更に本研究では、訓練データの歌詞を剽窃するリスクを下げる歌詞生成手法を提案する。機械学習に基づく歌詞生成では、訓練データ中の歌詞フレーズを再利用して剽窃するリスクがある。ユーザが作詞支援システムを安心して使うには、そのリスクを下げられる歌詞生成手法が求められると我々は考えている。そこで本研究では、訓練データ中で少数の曲のみが使用している固有なフレーズを生成不可能にする制約を加えることで、剽窃リスクを下げる。

実験の結果、提案手法は比較手法よりも入力文章の内容に沿いつつ字面が似ていない歌詞を生成可能であることが確かめられた。また、剽窃対策を施さないと、生成歌詞中の一割以上のフレーズが剽窃になってしまう可能性があることが確かめられた。

## 2 作詞支援システム

本研究では、歌詞に書きたい内容を文章として入力すると、その文章の内容に沿った歌詞の候補を生成し提示する作詞支援システムを提案する。本システムはユーザが文章を入力する入力パネル（図2左側）と、生成された歌詞を表示する歌詞表示パネル（右側）で構成される。ユーザは基本的な操作として、以下の手順により作詞をすることができる。

1. 入力パネルに、作りたい歌詞の内容を文章として自由な長さ・スタイルで入力する。

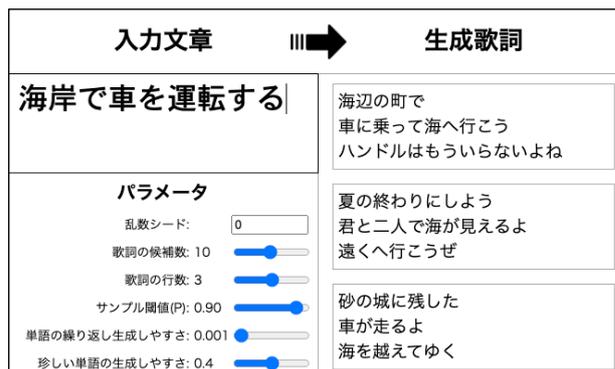


図2 作詞支援システムのスクリーンショット

2. 入力が完了すると、歌詞表示パネルに入力文章の内容に沿った歌詞の候補が複数生成される。

以上の操作で、本システムはユーザの意図に近い歌詞を複数生成するが、必要に応じて入力文章や歌詞を修正したり、生成する歌詞の条件（行数やシード値など）のパラメータを変更することで、ユーザが意図した歌詞に更に近づけることが可能となる。

### 2.1 自動生成が創作活動を妨げないために

自動生成技術が発展することで、作業が効率化されるという長所がある一方で、既存の創作物を剽窃してしまい創作者の権利を侵害してしまう可能性がある。本システムでは、多くの異なる曲で使用されているありがちなフレーズ（ありがち度 [9, 10] の高い単語列）を生成することは剽窃には当たらないと仮定し、少数の曲のみ（今回は仮に3曲以下とした）で使用された固有なフレーズを生成不可能にすることで剽窃の可能性を下げる機能を実装する。

## 3 実装

### 3.1 入力文章の内容に沿った歌詞を自動生成する手法

提案した作詞支援システムにおける自動歌詞生成は以下のような入出力を想定している。入力には歌詞に書きたい内容を表した文章であり、出力は入力文章の内容に沿いつつ字面が似ていない歌詞とする。この入出力の歌詞生成を可能にするエンコーダ・デコーダの訓練データとして、歌詞の内容を表しつつ字面は似ていない文章が多数必要となるが、そのようなデータは存在しない。また、文章要約技術等を用いて歌詞のペアとなる文章を自動生成したとしても、その文章は歌詞と字面が似ているため、このデータで学習されたエンコーダ・デコーダは入力文章と字面が似た歌詞を生成する可能性がある。

そこで本研究では図 1 のように、(a) 既存の画像生成技術を用いて入力文章の内容を表した画像を生成することで、入力文章の字面に関する情報を取り除く。そして (b) 画像の内容を表した歌詞を生成するエンコーダ・デコーダを用いることで、目的の「入力文章の内容に沿いつつ字面が似ていない歌詞」の生成を可能とする。このエンコーダ・デコーダを学習するためには、入力文章から生成された画像と歌詞のペアデータを用意できれば理想的だが、上述したようにそのような入力文章データは存在しない。そこで本研究では、画像生成技術を用いて歌詞から画像を生成することで、歌詞以外の文章データを一切用いずに訓練データを用意する。なお、一枚の画像は歌詞の一段落分の内容を表現すると仮定して、画像と歌詞段落のペアデータを用意する。

以上の着想を元に、提案エンコーダ・デコーダの詳細を説明する。まず、訓練データ用の日本語歌詞の各段落を日英機械翻訳<sup>1)</sup>により英語へ翻訳し、翻訳された英語歌詞から、アニメスタイルの画像を生成する技術 Anything v3.0<sup>2)</sup>(Stable Diffusion の改良型)を用いて解像度が 512×512 の画像を生成する。ここで、Anything v3.0 を用いる理由は、(1) 文章のまま入力してもその内容を表した画像が生成でき、(2) 一貫したスタイルの画像を用いた方がエンコーダ・デコーダを学習しやすいと考えたためである。

次に、Vision Transformer[11] を用いて画像から計算した特徴量 (10×768 次元) を、Transformer デコーダ [12] の Multi-Head Attention 層へ入力する。また、CaboCha[13] を用いて形態素解析された各単語を単語埋め込み層と位置埋め込み層を通してベクトル化し、Transformer デコーダの Masked Multi-Head Attention 層へ入力する。最後に、デコーダの出力を単語生成確率分布を表す全結合層へ渡す。

### 3.1.1 パラメータ

埋め込みの次元数は 768 とし、Multi-Head の数は 6、デコーダの層数は 2、Feedforward 層の次元数は 1024、活性化関数は GELU を用いる。Vision Transformer は学習済みモデル<sup>3)</sup>を用い、そのパラメータは再学習しない。最適化には AdamW[14] を用い、学習率は 0.001 とし、ウォームアップステップは 1 エポック分とする。学習は 40 エポック行い、

- 1) <https://huggingface.co/staka/fugumt-en-ja>
- 2) <https://huggingface.co/Linaqruf/anything-v3.0>
- 3) <https://huggingface.co/google/vit-base-patch32-224-in21k>

開発データにおける最良 Loss のモデルを採用する。

なお、BERT[15] や GPT-2[16] などの大規模言語モデルの学習済みパラメータはあえて利用せずに、歌詞データのみを用いてエンコーダ・デコーダを学習する。これは、大規模言語モデルの訓練データ中の文章を剽窃する可能性をゼロにするためである。

### 3.1.2 訓練データ

112,462 曲 (856,778 段落) の日本語のポピュラー音楽の歌詞を用い、8:2 で訓練/開発データに分割する。学習対象の語彙は文書頻度が大きい上位 50,831 の単語を使用し、その他の単語は未知語タグに置換する。また、語彙には改行を表す EOL タグと段落頭/末を表す BOP/EOP タグが含まれる。

## 3.2 歌詞生成アルゴリズム

作詞支援システムでは、様々なバリエーションの歌詞を複数生成・提示することで、ユーザに新たな発想を与えることを想定している。そのため本研究では、ビーム探索のような最尤な単語列を探索する手法ではなく、サンプリングによる生成手法を用いる。ただし、単にサンプリングすると生成確率が低い単語が生成される可能性があるため、本研究では Top- $P$  サンプリング [17] を用いて、生成確率が低い単語のサンプリングを禁止する。Top- $P$  サンプリングでは、各単語列を独立に生成するため、複数の歌詞候補を同時に生成できる。さらに本研究では、ユーザが設定した行数などの詳細な条件を満たした歌詞を生成するために、これらの条件を満たさない単語列の生成を禁止する。

提案生成手法が既存歌詞を剽窃するリスクを下げるために、歌詞データ中の文書頻度が 1 から 3 である単語 4-gram の生成を禁止する。この処理は、多くの曲で使用されているありがちなフレーズは再利用できるが、少数の曲でしか使用されていない固有なフレーズを生成したら剽窃にあたる可能性が高い、という仮定の元で設計した。単語 4-gram の文書頻度を正確に計算するためには、歌詞データ中に同一楽曲の歌詞が重複して存在しない状態にする必要がある。ここで重要なことは、異なる歌詞が誤って同一楽曲と判定されることは問題ないが、同じ歌詞なのに同一楽曲でないと判定されることは避けなければならない点である。そこで、歌詞データ中のあらゆる歌詞のペアにおいて以下の 2 つの基準で判定することで、同一楽曲判定に取りこぼしがないように

する。(1) 単語 20-gram が一致している単語列が存在すれば同一楽曲とする。(2) もしくは、歌詞間の標準化編集距離 [18] が 0.5 以下のものを同一楽曲とする。この基準により、112,462 曲中で同一楽曲の歌詞を連結して 1 つの歌詞に置き換えた 109,829 曲の歌詞を用いて、単語 4-gram の文書頻度を計算した。

## 4 実験

提案エンコーダ・デコーダの性能を評価するためには、実験用の歌詞とその内容を表した入力文章のペアが必要となる。本研究では、小説がアニメ化された時のアニメ主題歌の歌詞は、小説の内容を元に作詞されたと仮定し、51 作品の Web 小説の概要文章と、その主題歌の歌詞 (2730 段落) を評価データとして用意した。実験では、小説の概要文章を入力した時のパープレキシティや生成歌詞を評価する。

実験の比較手法として、異なるペアデータを学習した以下のエンコーダ・デコーダを用意した。

**Image-to-Lyrics (I2L)** 提案エンコーダ・デコーダ。

**Summary-to-Lyrics (S2L)** 文章要約手法<sup>4)</sup>を用いて歌詞の一段落を一文へ要約することで要約文とオリジナル歌詞のペアを作成してから、そのペアデータを学習したエンコーダ・デコーダ。

**Back translation-to-Lyrics (B2L)** 日英翻訳手法と英日翻訳手法<sup>5)</sup>を用いて歌詞を日→英→日と逆翻訳することで逆翻訳歌詞とオリジナル歌詞のペアを作成してから、そのペアデータを学習したエンコーダ・デコーダ。

上記 S2L と B2L は Transformer を用いたエンコーダ・デコーダであり、そのパラメータは I2L と同じである。生成手法の Top- $P$  サンプリングのパラメータ  $P$  は 0.9 とし、1 つの入力に対して 5 つの歌詞候補を生成する。この比較実験では純粋な歌詞生成能力を評価するために、剽窃対策は施していない。

まず、各エンコーダ・デコーダが入力文章の内容に沿った歌詞を生成する能力を持つかを調べるために、パープレキシティを計算した。その結果、I2L は 149.30、S2L は 181.60、B2L は 1207.52 となり、提案エンコーダ・デコーダのパープレキシティが最も良いことが確かめられた。また、生成歌詞が入力文章の字面と似ているかを調べるために、生成歌詞と入力文章の標準化編集距離を計算した。その結果、I2L は 0.91、S2L は 0.87、B2L は 0.76 となり、提案

4) <https://huggingface.co/tsmatz/mt5-summarize-japanese>

5) <https://huggingface.co/staka/fugumt-en-ja>

表 1 入力文章に対して生成された歌詞の例

入力文章	生成歌詞
こんなあなたを好きだったなんて思った事はなかった。	唇で触れた指を絡めて あなたの瞳に映ってるから。
男性が絵を描いている。	色のない絵の具で塗り固めた 僕らは何色のキャンパスに描いて ただろう?
子供が海の波に走って入ったり出たりしている。	裸足で走る風を追い越して 君が残した足跡は波の上に立っている。

エンコーダ・デコーダが生成した歌詞は入力文章と最も文字列が異なっていることが確かめられた。これらの実験結果は、要約文-歌詞ペアや逆翻訳歌詞-歌詞ペアを学習した場合よりも、画像-歌詞ペアを学習した場合の方が、入力文章の内容に沿いつつ字面が似ていない歌詞を生成できることを意味する。

次に、剽窃対策を施していない歌詞生成手法がどれだけ剽窃リスクを負っているかを調べるために、生成された歌詞中のすべての単語 4-gram のうち、文書頻度が 1 以上かつ 3 以下の単語 4-gram の割合  $R$  を計算する。その結果、提案エンコーダ・デコーダを用いて歌詞を生成した場合、 $R = 17.5\%$  であることが確認された。これは、剽窃対策を施さないと生成歌詞の一割以上のフレーズは既存歌詞に固有なフレーズである可能性が高いことを意味する。ただし、提案した剽窃対策を施すことで  $R = 0\%$  となり、既存歌詞に固有なフレーズを再利用してしまう剽窃リスクを下げるのが可能となる。

表 1 に、入力文章とそれに基づいて提案エンコーダ・デコーダが生成した歌詞の例を示す。なお、ここでは剽窃対策を施した生成手法を用いている。これらの入力文章は、RWC 研究用音楽データベース [19] の歌詞と JSICK データセット [20] の文から選択した。この表より、入力文章の内容に沿いつつ字面が似ていない歌詞が生成されていることがわかる。また、その他の生成結果の例を付録に掲載する。

## 5 まとめ

本研究の主な貢献は以下の通りである。(1) 入力文章の内容に沿った歌詞候補を提示することでユーザーに新たな発想を与える作詞支援システムを提案した。(2) 入力文章の内容に沿いつつ字面が似ていない歌詞を生成するために、画像生成技術を活用した Image-to-Lyrics エンコーダ・デコーダを提案した。(3) 既存の歌詞を剽窃するリスクを下げる歌詞生成手法を提案した。今後は作詞支援システムのユーザー評価を行う予定である。

## 謝辞

本研究の一部は JST CREST JPMJCR20D4 と JSPS 科研費 JP20K19878 の支援を受けた。

## 参考文献

- [1] Kento Watanabe and Masaraka Goto. Lyrics information processing: Analysis, generation, and applications. In **Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA 2020)**, pp. 6–12, 2020.
- [2] Chihiro Abe and Akinori Ito. A Japanese lyrics writing support system for amateur songwriters. In **Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2012)**, pp. 1–4, 2012.
- [3] 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**, pp. 195–204, 2016.
- [4] 渡邊研斗, 松林優一郎, 乾健太郎, 中野倫靖, 深山覚, 後藤真孝. LyriSys: 歌詞の大局的構造に基づいた作詞支援インタフェース. 日本ソフトウェア科学会第23回インタラクティブシステムとソフトウェアに関するワークショップ (WISS 2015) 論文集, 2015.
- [5] Kento Watanabe, Yuichiro Matsubayashi, Kentaro Inui, Tomoyasu Nakano, Satoru Fukayama, and Masataka Goto. LyriSys: An interactive support system for writing lyrics based on topic transition. In **Proceedings of the 22nd Annual Meeting of the Intelligent User Interfaces Community (ACM IUI 2017)**, pp. 559–563, 2017.
- [6] Hugo Gonalo Oliveira, Tiago Mendes, and Ana Boavida. Co-PoeTryMe: a co-creative interface for the composition of poetry. In **Proceedings of the 10th International Conference on Natural Language Generation (INLG 2017)**, pp. 70–71, 2017.
- [7] Hugo Gonalo Oliveira, Tiago Mendes, Ana Boavida, Ai Nakamura, and Margareta Ackerman. Co-PoeTryMe: Interactive poetry generation. **Cognitive Systems Research**, Vol. 54, pp. 199–216, 2019.
- [8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022)**, pp. 10674–10685, 2022.
- [9] 中野倫靖, 吉井和佳, 後藤真孝. 確率的生成モデルに基づく音楽の類似度とありがち度の推定に関する検討. 情報処理学会研究報告 音楽情報科学 (MUS), Vol. 2014, No. 2, pp. 1–7, 2014.
- [10] Tomoyasu Nakano, Kazuyoshi Yoshii, and Masataka Goto. Musical similarity and commonness estimation based on probabilistic generative models of musical elements. **International Journal of Semantic Computing (IJSC 2016)**, No. 1, pp. 27–52, 2016.
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)**, 2021.
- [12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, pp. 1–11, 2017.
- [13] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In **Proceedings of the 6th Conference on Natural Language Learning (CoNLL 2002)**, 2002.
- [14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In **Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)**, 2019.
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)**, pp. 4171–4186, 2019.
- [16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. **OpenAI blog**, Vol. 1, No. 8, p. 9, 2019.
- [17] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In **Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)**, 2020.
- [18] Yujian Li and Bi Liu. A normalized Levenshtein distance metric. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Vol. 29, No. 6, pp. 1091–1095, 2007.
- [19] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC Music Database: Popular, classical and jazz music databases. In **Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)**, Vol. 2, pp. 287–288, 2002.
- [20] Hitomi Yanaka and Koji Mineshima. Compositional evaluation on Japanese textual entailment and similarity. **Transactions of the Association for Computational Linguistics (TACL)**, Vol. 10, pp. 1266–1284, 2022.

## A 入力文章に対して生成された画像および歌詞の例

入力文章	中間表現（生成画像）	生成歌詞
二人の人が戦っていて、観客が観ている		このままじゃ 遊びは終わらせない 勝ち上がるだけさ！ 全てを賭けて
探検者のグループが草原を歩いている		草原の芝生に立って 二人で歩いた道を行く まだ見ぬ場所へ 風が吹くよと言っていたけど
人々が海のそばの砂でいっぱいの浜に いて、晴れた一日を楽しんでいる		今日も空は青く 晴れた日には君がいるから 海へと続く道を歩いてくよ ここから先のストーリー一緒なら
会いたくてせつなくて泣きだしそうで		思い出の中で あなたを見つめているだけで 涙があふれそうなの いつまでも傍にいたいから
ご機嫌なナンバーだぜ ラジオから流れてる 俺を急がせるよ Crazy D.J		このビートが刻むリズムの鼓動 これは音に乗って 俺らの鼓膜を揺らすんだろ？ dj mc と mic で作れるかな？
未来はまだ誰も見えなくて それでもきっと素敵が待っている		今はまだ見えない 未来の僕等が立っているんだ 明日を信じてるよ 君にしか会えなかったから

上の例のうち、上3つの入力文章は JSICK データセットから選択した文であり、下3つの入力文章は RWC 研究用音楽データベースの歌詞から選択した文章である。上3つの例では、入力文章に記された具体的な情景が画像として生成され、この画像の内容を表した歌詞が生成されていることがわかる。一方、下3つの例は感情的・抽象的な内容の入力文章である。この時、入力文章の内容と同じ感情の表情をした人物の画像が生成され、その感情を表した歌詞が生成されていることがわかる。