Vol.2022-DBS-175 No.26
Vol.2022-IFAT-148 No.26
2022/9/10

楽曲歌詞の意味的な重畳性に基づいた印象的フレーズの抽出

佐々木 翔一[1,a)]　牛尼 剛聡[2,b)]

概要：音楽ストリーミングサービスでは，楽曲に対してサムネイル画像が表示されることが多い．楽曲に付与されているサムネイル画像は，視覚的に短時間で内容を把握することができるため，楽曲の選別に大きく貢献している．しかし，同一アルバム内の楽曲には，すべて同じジャケット画像やアーティスト画像が割り振られる場合には，それらが個々の楽曲内容を適切に表現していない場合がある．そこで，本研究では，歌詞情報に基づいて，楽曲に適したサムネイル画像を自動生成する手法を提案する．具体的には，まず歌詞をフレーズに分解し，それぞれのフレーズと同じ楽曲の他のフレーズとの意味的な重なりを調べることで，楽曲を特徴づけている印象的なフレーズを抽出する．その後，抽出したフレーズを画像生成モデルを利用して画像に変換することで，楽曲の特徴を表すサムネイル画像の自動生成を行う．

1. はじめに

近年，音楽を聞く手段として，音楽ストリーミングサービスを利用する人が増えている．日本レコード協会が行った調査によると，2012 年から 2021 年の 10 年間における音楽配信売上実績のうち，ストリーミングの占める割合が急増しているという [1].

多くの音楽ストリーミングサービスでは，個々のコンテンツに対してサムネイル画像が付与されている．サムネイル画像は，実際に映画を視聴したり音楽を聴取したりするのに比べて，視覚的に短時間で内容を把握することができる．そのため，ユーザのコンテンツに対する理解支援やユーザ体験の向上を目的とした，サムネイル画像およびストリーミング中の提示画像の最適化について盛んに研究が行われている [2][3].

しかし，これらの手法では楽曲の特徴と画像の特徴を事前に紐付け，そこから得られた関係性をもとにサムネイル画像を生成しているため，学習過程で用いられたデータと大きく異なる特徴が入力されたときに，適切な画像を生成することが難しい．そのためこれらの手法は，生成画像の多様性が高くない．

近年，短文を入力とし，その文の内容を表す画像を生成する Text to Image モデルが盛んに研究されており，「アボカドの形をした椅子」「写実的なスタイルの馬に乗った宇宙飛行士」など，教師データに含まれていない複雑な内容に対しても高水準な画像を生成できるようになった [4][5]. また，これらの手法では，画像と画像の内容を説明したテキストの関係を学習しているため，単一単語ではない複雑な文章を表す画像も生成することができ，生成画像の多様性という点において従来手法よりも秀でていると考えられる．そこで，本研究では楽曲におけるテキストデータである歌詞に着目し，これを Text to Image モデルに入力することで，楽曲に適したサムネイル画像を自動生成することを目指す.

Text to Image モデルに入力するテキストは「アボカドの形をした椅子」「写実的なスタイルの馬に乗った宇宙飛行士」など，比較的短く具体的な文章が想定されているため，歌詞のような長文をそのままの状態で入力したとしても，楽曲に適したサムネイル画像を出力することは難しい．そのため，歌詞全文を，生成する画像の内容を示す具体的で短い文章に変換する必要がある．そこで，本研究では楽曲歌詞における印象的フレーズに着目する．印象的フレーズはその楽曲を象徴する可能性が高いため，楽曲に適したサムネイル画像を生成する際に，画像の内容を示す短い文章として利用することができると考えられる.

楽曲の歌詞から，印象的フレーズを抽出することを目指した例として，山西らの研究では，楽曲におけるフレーズの繰り返しが印象に大きな影響を与えていることを，被験者実験の結果から報告している [6]．ここで言う繰り返しとは，同じ文字列の繰り返しという意味であるが，本研究ではこれに加えて，フレーズの意味的な繰り返し構造につい

[1] 九州大学大学院芸術工学府
Graduate School of Design, Kyushu University, Fukuoka 815–8540 Japan
[2] 九州大学大学院芸術工学研究院
Faculty of Design, Kyushu University, Fukuoka 815–8540 Japan
[a)] sasaki.shoichi.896@s.kyushu-u.ac.jp
[b)] ushiama@design.kyushu-u.ac.jp

ても，印象的フレーズに影響を与えているという仮説をたてた．楽曲には作者が一曲全体を通して伝えたいテーマがあり，この作者が表現したいテーマと意味的に最も近いフレーズが印象的フレーズになると考えられる．そして，作者が表現したいテーマは楽曲内のフレーズすべてに対して意味的な影響を与えていると考えられるため，楽曲のテーマと意味的に近い印象的フレーズは，その他のフレーズに対する類似度が大きいと考えられる．つまり，各フレーズのその他のフレーズに対する意味的な重畳性を確認すれば，印象的フレーズを推定することが可能であると考えられる．また，このとき抽出した印象的フレーズは，作者が伝えたいテーマと意味的に近いため，楽曲にふさわしい画像を生成するためのテキストとしても利用可能である可能性がある．

以上より，本論文では楽曲に適したサムネイル画像を生成するために，歌詞の意味的な重畳性に基づいて，印象的フレーズを抽出する手法を提案する．その後，抽出した印象的フレーズを Text to Image モデルに入力することで，サムネイル画像の自動生成を行う．

2. 関連研究

2.1 サムネイル画像の自動生成

映画配信サービスや音楽ストリーミングサービスでは，個々のコンテンツに対してサムネイル画像が付与されている．サムネイル画像は，実際に映画を視聴したり音楽を聴取したりするのに比べて，視覚的に短時間で内容を把握することができる．そのため，ユーザのコンテンツに対する理解支援やクリック確率の向上を目的として，サムネイル画像の最適化について活発に研究が行われている．

たとえば，映画配信サービスの Netflix は，ユーザの視聴履歴と映像内から抽出した映像的特性に基づきユーザの好みの傾向を予測することで，各ユーザに適したシーンのフレームを自動選択し，サムネイル画像を生成する手法を提案している [7]．一方，音楽は音コンテンツであるため，それ自体に内容を表す視覚的な要素が含まれていない．そのため，Netflix のようなシーン選択を利用した手法を用いることができず，楽曲から抽出した情報をもとに適切なサムネイル画像を自動生成・選択するというアプローチを取る必要がある．

楽曲情報をもとに，その楽曲に合う適切な画像の生成・選択を目指した研究として，梅村らは画像に付与した語ラベルと楽曲との共起性を Word2Vec により計算し，曲の流れにあったスライドショーを提示する手法を提案している [3]．また，Qiu らは短時間フーリエ変換（STFT）した楽曲音源を CNN および LSTM を用いたモデルに入力することで，楽曲内容を表現することに適した特徴量を獲得し，その後 DCGAN と組み合わせることで，楽曲の雰囲気に合うサムネイル画像を自動生成する手法を提案している [2]．

しかし，梅村らの手法では，あらかじめ存在する画像の中から適切なものを選択するため，対象楽曲の内容に近い画像が存在しない場合，適切なサムネイル画像を選択できない可能性がある．また，Qiu らの手法では，事前に sky，water，moutain，desert という 4 つのテーマを表す楽曲の特徴量と画像とを紐付けるような学習を行っている．そのため，これら 4 つのテーマのいずれかに関する楽曲の場合適切な画像を生成することができても，全く未知の楽曲であった場合，適切な画像を生成できない可能性がある．

このように，既存の手法では楽曲の特徴と画像の特徴を紐付け，そこから得られた関係性をもとにサムネイル画像を生成・選択しているため，事前情報として保有していない内容の入力があったときに，適切な画像を生成することが難しい．そのためこれらの手法は，一つ一つの楽曲に適したサムネイル画像を出力するという，生成画像の多様性が求められるタスクにおいて問題があると考えられる．

2.2 楽曲における印象的フレーズの抽出

楽曲に対する印象や評価は，個人の感性による部分も大きいが，その一方で多くの人間にとって印象深く刻まれるフレーズも存在すると考えられる．たとえば，楽曲のサビは印象的なフレーズであると考えられる場合が多い．

一方で，楽曲における印象的なフレーズは必ずしもサビであるとは限らないため，サビ以外からも印象的なフレーズを抽出することを目指した研究も報告されている．たとえば，山西らは楽曲内の共起語の特異性と繰り返しが，印象的なフレーズに影響を与える可能性を示唆している [6]．特に繰り返し表現は印象的なフレーズに影響する割合が多いことを，被験者実験により明らかにしている．しかし，文字列の一致という意味での繰り返し抽出では，楽曲内容の意味的な繰り返しまでは取ってこれない．楽曲内容のテーマに沿った適切なサムネイル画像を生成するためには，歌詞の意味に着目した重要フレーズの抽出が必要であると考えられる．

2.3 テキストからの画像生成

楽曲内容に適した画像を生成するには，印象的なフレーズを抽出した後，画像に変換する必要がある．しかし，2.1 で述べた手法では，事前に楽曲の情報と画像との関係を紐付けておく必要があり，多種多様な歌詞に対して適切なサムネイル画像を生成することが困難である．一方，近年では短文を入力とし，その文の内容を表すような画像を生成する Text to Image モデルが盛んに研究されている．DALL-E 2 や Imagen などでは，入力された文に対して，高水準で内容を表す画像を生成できることが報告されている [4][5]．また，これらの手法では，Qiu らの手法のように sky，water といった特定のクラスと画像とをセットとして学習が行われているのではなく，画像と画像の内容を説明

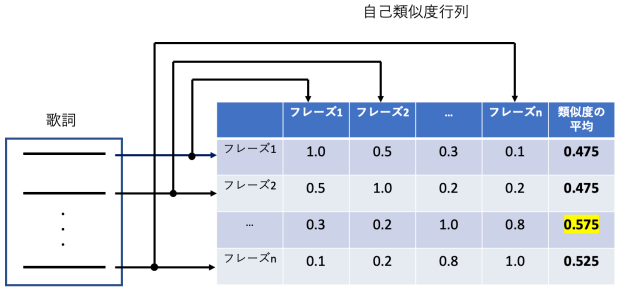

図 1: 自己類似度を用いた印象的フレーズの決定法

したテキストとの関係を学習している．そのため，単一単語ではない複雑な文章を表す画像も生成することができ，生成画像の多様性という点において従来手法よりも秀でていると考えられる．

3. 提案手法

本研究では楽曲に適したサムネイル画像を生成するために，歌詞の繰り返し構造に基づいて，印象的フレーズを抽出することを目的とする．そして，抽出した印象的フレーズを Text to Image モデルに入力することで，サムネイル画像の自動生成を行う．

3.1 印象的フレーズの抽出

3.1.1 ベースとなる印象的フレーズの決定方法

本研究では，楽曲の歌詞における繰り返し構造をもとに，印象的フレーズを決定する．具体的には，各フレーズ間での自己類似度を用いてフレーズの重要度を算出することによりこれを実現する．

フレーズ間での自己類似度を用いた印象的フレーズの決定法を図 1 に示す．楽曲の歌詞は，時間的な順序関係を有するフレーズの列と考えることができる．まず，対象とするフレーズ p に対して，同じ楽曲に含まれるすべてのフレーズとの類似度を求める．その後，求めたすべての類似度の平均を取ることで，フレーズ p の重要度を予測する．この計算を，対象とする楽曲に含まれるフレーズすべてを対象として行った後，求めた重要度が最も大きいものを，その楽曲の印象的フレーズとする．

我々は，印象的フレーズについて，文字列の繰り返し構造だけでなく，意味的な繰り返し構造が影響を与えているという仮説を立てた．そこで，提案手法における類似度の計算には，文字列の類似度だけでなく，意味的な類似度についても検証を行い，両者を比較する．

3.1.2 文字列の繰り返し構造を用いた類似度の算出

文字列の類似度を計算する代表的な手法の一つに編集距離がある．編集距離は，二つの文字列がどの程度異なっているかを示すことができる指標である．具体的には，1 文字の挿入・削除・置換によって，一方の文字列をもう一方の文字列に変形するのに必要な手順の最小回数として定義される．

文字列の繰り返し構造を用いた印象的フレーズの決定法では，3.1.1 で示した手法における類似度計算に，編集距離に基づく類似性を表す指標（0 から 1 の範囲をとり同一文章のとき 1 となる）を適用し，平均値が最も大きいフレーズを印象的フレーズと決定する．

3.1.3 フレーズの意味的な繰り返し構造を用いた類似度の算出

フレーズの意味的な類似度を計算するためには，まず各フレーズの文字列をそのフレーズが持つ意味を表現できる形式に変換し，その後類似度を計算する必要がある．フレーズの形式変換には，テキストの分散表現を利用する．フレーズの分散表現を求めるためには，フレーズを直接 Doc2Vec で変換する方法と，フレーズを構成する各トークンを Word2Vec で変換し，それらの平均をフレーズの分散表現とする方法の 2 つを採用する．

分散表現の類似度の計算には，cos 類似度を利用する．

これらの方法を，3.1.1 で示した提案手法に適用し，各フレーズで求めた cos 類似度の平均値が最も大きいフレーズを印象的フレーズと決定する．

なお，単純な平均では，各フレーズに対して全体的に意味的な類似性を示すフレーズが印象的フレーズとして選択される．そのため，強い意味的な繰り返しが行われているフレーズを考慮した抽出が行えない．そこで，各フレーズに対する cos 類似度が，一定のしきい値を超えた回数をカウントし，そのカウント数も印象的フレーズの選択に利用する．具体的に比較して用いる印象的フレーズの決定方法は，各フレーズに対する cos 類似度の算術平均，一定のしきい値を超えた回数，しきい値を超えた回数を重みとする重みつき平均の 3 つである．

3.2 抽出フレーズの画像化

3.1.1 で述べた手法により歌詞から抽出したフレーズを，Text to Image モデルに入力することで，サムネイル画像を生成する．具体的な手法としては，CLIP[8] と GAN[9] を組み合わせた FuseDream モデル [10] を使用する．なお，使用する FuseDream モデルは英文から画像を生成するようにトレーニングれている．本研究では，日本語の楽曲を対象とするため，抽出したフレーズを機械翻訳により英文化し，FuseDream モデルに入力する．フレーズの翻訳には，Google 翻訳[*1]を利用する．

4. 実験

4.1 データセットの作成

楽曲歌詞に対する聴取者の印象を評価するための基準となるデータセットを被験者実験により作成した．

[*1] https://translate.google.co.jp/?hl=ja

対象楽曲は，歌詞検索サービスの歌ネットにおける歴代人気曲ランキング，上位 25 曲（2022 年 7 月 31 日アクセス時点）とし，クラウドソーシングにより 30 名〜40 名程度の被験者に印象的フレーズを評価させた（楽曲により回答者数が異なる）．実験では，被験者に 1 曲ずつ楽曲の歌詞を提示し，その楽曲において印象的であると感じるフレーズを，1 行単位で任意数選択させた．その後，回答を集計し，被験者により選択された回数により各フレーズに対して印象的度合いのランク付けを行った．

4.2 印象的フレーズの抽出および評価実験

4.1 で述べた楽曲 25 曲に対して，提案手法を適用し，文字列の繰り返し構造を用いた手法（編集距離を用いた手法），意味的な繰り返し構造を用いた手法（Word2Vec を用いた手法，Doc2Vec を用いた手法）で，それぞれ印象的フレーズを求めた．なお，Word2Vec[*2]および Doc2Vec[*3]は日本語学習済みモデルを利用した．また，意味的な繰り返し構造を用いた手法では，印象的フレーズを求める際に，cos 類似度の算術平均，cos 類似度が一定のしきい値を超えた回数，cos 類似度が一定のしきい値を超えた回数を重みとする重みつき平均の 3 つを用いた．cos 類似度のしきい値は 0.9 とした．

その後，提案手法により予測された印象的フレーズと，被験者実験により定めた印象的フレーズとの一致度を求めた．

4.3 抽出フレーズの画像化および評価実験

4.2 で，提案手法により抽出したフレーズを，FuseDream モデルに入力し，サムネイル画像を生成した．

生成の手順は，まず，印象的フレーズを Google 翻訳に入力し，出力された英文を FuseDream モデルに入力することで，サムネイル画像を獲得した．

5. 結果および考察

5.1 印象的フレーズの抽出および評価実験

提案手法により予測した印象的フレーズと，被験者実験により定めた印象的フレーズとの一致度を，表 1, 表 2, 表 3 に示す．

表の各列は実験に用いた 25 曲に対して予測した印象的フレーズがそれぞれ，回答数 1 位のフレーズと一致した割合，回答数が 5 位以内のフレーズと一致した割合，予測したフレーズに付与された正解ランクの平均，予測したフレーズに付与された正解ランクの分散である．

また，最もスコアが高かった，Doc2Vec を用いて他のフレーズに対する cos 類似度が 0.9 を超えた回数により印象的フレーズを決定した際の，抽出結果を表 4 に示す．表 4 には，例として歌ネットにおける歴代人気曲ランキング上位 10 曲を示す．

表 1: 予測した印象的フレーズと正解データとの一致度（編集距離）

1 位正解率	5 位以内正解率	平均順位	分散
0.32	0.68	4.16	13.31

表 2: 予測した印象的フレーズと正解データとの一致度（Word2Vec）

フレーズの決定方法	1 位正解率	5 位以内正解率	平均順位	分散
算術平均	0.08	0.32	8.68	24.48
0.9 を超えた回数	0.28	0.72	4.12	14.61
重みつき平均	0.24	0.76	3.96	13.71

表 3: 予測した印象的フレーズと正解データとの一致度（Doc2Vec）

フレーズの決定方法	1 位正解率	5 位以内正解率	平均順位	分散
算術平均	0.04	0.16	9.36	18.66
0.9 を超えた回数	0.36	0.84	3.24	10.27
重みつき平均	0.20	0.72	4.48	13.18

まず，ベースラインとなる文字列の繰り返し構造を用いた手法では，全体の 3 割程度，回答数 1 位のフレーズを予測できている．そのため，楽曲歌詞における文字列の繰り返し構造は，印象的フレーズに大きく影響を与えていることが示唆された．しかし，文字列の繰り返し部分以外には印象的フレーズを抽出することができないため，予測を大きく外している楽曲もあった．

次に，意味的な繰り返し構造を用いた手法では，Doc2Vec を用いて他のフレーズに対する cos 類似度が 0.9 を超えた回数により印象的フレーズを決定した際に，最も高いスコアを示した．特に，5 位以内正解率および予測したフレーズに付与された正解ランクの平均が，それぞれ 0.68 から 0.84，4.16 から 3.24 に向上している．つまり，印象度合い 1 位のフレーズを予測できない場合でも，予測を大きく外すことが少なくなっている．これは，cos 類似度のしきい値を 0.9 としたため，ある程度の文字列の類似性も考慮することができ，その上で，意味的に強く繰り返されている部分が抽出できているからではないかと考えられる．

一方で，予測を大きく外す楽曲もいくつか存在したため，該当する楽曲の正解データを確認したところ，文字列の繰り返しでもなく意味的な繰り返しでもない部分が印象的フレーズとして選択されていた．たとえば，楽曲においてアーティストが感情を込めて強く歌う部分や，タイトル名と関連のある部分，楽曲の音声的な構造で特異かつ目立つ部分などが選択されていた．このようなフレーズについては，意味的な繰り返し構造を用いた手法では抽出できな

[*2] http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/
[*3] https://yag-ays.github.io/project/pretrained_doc2vec_wikipedia/

表 4: 正解データにおける回答数 1 位のフレーズと提案手法により選択されたフレーズ

楽曲名／アーティスト名	回答数 1 位のフレーズ	提案手法によって選択されたフレーズ	選択されたフレーズの正解データ内での順位
Lemon ／米津玄師	今でもあなたは私の光	今でもあなたは私の光	1
クリスマスソング／ back number	君が好きだ	君が好きだ	1
キセキ／ GReeeeN	せめて言わせて 「幸せです」と	君に巡り会えた それって奇跡	2
花束／ back number	君と抱き合って手を繋いでキスをして	僕は何回だって何十回だって	2
前前前世（movie ver.）／ RADWIMPS	君の前前前世から僕は 君を探しはじめたよ	君の前前前世から僕は 君を探しはじめたよ	1
Pretender ／ Official 髭男 dism	君の運命のヒトは僕じゃない	グッバイ	2
LOSER ／米津玄師	アイムアルーザー どうせだったら遠吠えだっていいだろう	アイムアルーザー どうせだったら遠吠えだっていいだろう	1
紅蓮華／ LiSA	紅蓮の華よ咲き誇れ！ 運命を照らして	強くなれる理由を知った 僕を連れて進め	4
ハッピーエンド／ back number	あなたを好きなままで消えてゆく	なんてね 嘘だよ 元気でいてね	9
RPG ／ SEKAI NO OWARI	空は青く澄み渡り 海を目指して歩く	空は青く澄み渡り 海を目指して歩く	1

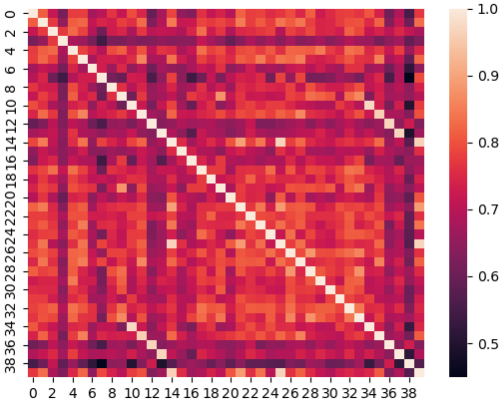

図 2: 「Lemon ／米津玄師」に Doc2Vec を適用した際の自己類似度行列

図 3: 生成した画像例（Lemon ／米津玄師）

いため，正解することが難しいと考えられる．

また，意味的な繰り返し構造を用いた手法において算術平均を用いて印象的フレーズを決定した際には，各スコアが大きく低下する結果となった．これは，各フレーズに対して全体的に意味的な類似性を示すフレーズが選択されているため，強い意味的な繰り返しが行われているフレーズを抽出できていないことが要因であると考えられる．例として，図 2 に，「Lemon ／米津玄師」の歌詞に対して Doc2Vec を用いた手法を適用した際の cos 類似度の自己類似度行列を示す．被験者により，選定された印象的フレーズは 14，25，39 番（「今でもあなたは私の光」）であり，cos 類似度のしきい値を用いた場合では類似度が 0.9 を超えた回数をカウントしているため，正解フレーズを抽出できている．しかし，算術平均を用いた場合では cos 類似度の平均をとるため，自己類似度が全体的に高いフレーズである 4 番（「戻らない幸せがあることを」）が選択されてしまっている．このように算術平均を用いた手法では，意味的に重要でないフレーズの影響も多分に受けてしまうため，結果的に目的とする印象的フレーズを抽出できない傾向にあると考えられる．

5.2 抽出フレーズの画像化および評価実験

5.1 において，最も精度が高かったのは，Doc2Vec を用いて他のフレーズに対する cos 類似度が 0.9 を超えた回数により印象的フレーズを決定する手法であった．この手法において，回答数 1 位のフレーズを抽出することができた楽曲を対象に，4.3 で述べた提案手法を適用し，サムネイル画像の生成を行った．

生成したサムネイル画像の例を図 3，図 4，図 5，図 6 に示す．概ね，フレーズの内容を考慮した画像を生成することができたが，中には楽曲の内容に適さない画像もあった．たとえば，図 3 の「Lemon ／米津玄師」や図 4 の「RPG ／ SEKAI NO OWARI」のサムネイル画像は比較的フレーズの内容を表すことができていると思われる．これは，生成に用いられた「今でもあなたは私の光」や「空は青く澄み渡り 海を目指して歩く」といったフレーズが，心象風景として画像化しやすいからであると考えられる．一方，図 5 の「白日／ King Gnu」や図 6 の「夜に駆ける／ YOASOBI」などは，印象的フレーズがそれぞれ「真っ新に生まれ変わって」「沈むように溶けてゆくように」であり，このフレーズだけでは具体的な心象風景として画像化につなげることが難しいため，適した画像を生成することが難しかったと考えられる．

6. まとめ

本研究では楽曲に適したサムネイル画像を生成するために，歌詞の意味的な重畳性に基づいて，印象的フレーズを抽出する手法を提案した．その後，抽出した印象的フレーズを Text to Image モデルに入力することで，サムネイル画像の生成を行った．

実験の結果，Doc2Vec を用いて他のフレーズに対する cos 類似度が 0.9 を超えた回数により印象的フレーズを決

図 4: 生成した画像例（RPG ／ SEKAI NO OWARI）

図 5: 生成した画像例（白日／ King Gnu）

図 6: 生成した画像例（夜に駆ける／ YOASOBI）

定した際に，文字列の繰り返し構造に着目した手法（ベースライン）よりも，高精度で抽出が行えることがわかった．一方，アーティストが強く歌う部分や，音声的な構造で目立つ部分など，文字列の繰り返しでもなく意味的な繰り返しでもない部分が印象的フレーズとして選択されている楽曲は，抽出を行うことが難しかった．

サムネイル画像については，概ねフレーズの内容を考慮した画像を生成することができた．一方で，人間が頭の中で映像化することが難しいようなフレーズについては，楽曲の雰囲気に適さない画像が出力されることがあった．

本研究で提案した手法では，歌詞の文字列の繰り返し構造や意味的な繰り返し構造に関する印象的フレーズは概ね抽出することができた．しかし，楽曲の印象的フレーズは，これら繰り返し構造以外の部分にも現われることが示唆された．たとえば，楽曲においてアーティストが感情を込めて強く歌う部分や，タイトル名と関連のある部分，楽曲の音声的な構造で特異かつ目立つ部分などがあげられる．今後は，このような要因に影響される印象的フレーズも正しく抽出が行えるように，分散表現取得モデルの再選定，楽曲の音声を考慮したフレーズ抽出方法の提案を行っていく予定である．

また，サムネイル画像の生成については，楽曲を聞いて人間が思い浮かべる心象風景に，より近い内容の画像を生成することが課題である．現在，抽出した印象的フレーズを直接 Text to Image モデルに入力しているが，楽曲を聞いて人間が思い浮かべる心象風景は，印象的フレーズやその他の歌詞との関係，音声的な特徴など様々な要因が相まって形作られるものである．今後は，楽曲を聞いて思い浮かべる心象風景をより具体的に説明できるテキストを出力できるような仕組みについても研究を行っていく予定である．当初の実施内容としては，生成した画像の妥当性に関する評価（被験者実験），Text to Image モデルを日本語で利用できるように再トレーニングすること，楽曲全体の内容をまとめるような説明文を生成するために文書要約の機構を取り入れること，楽曲を聞いて思い浮かべる心象風景に関する特徴的な名詞や形容詞を抽出することを行う予定である．

参考文献

[1] 日本レコード協会：音楽配信売上実績　過去 10 年間　全体, 入手先 ⟨https://www.riaj.or.jp/f/data/annual/dg_all.html⟩ (2022.08.08).
[2] Yue Qiu, Hirokatsu Kataoka: *Image generation associated with music data*, CVPR workshop (2018).
[3] 梅村允康, 保利武志, 土屋駿貴, 嵯峨山茂樹: *Word2Vec* を用いて歌詞と写真を対応づけたスライドショー生成システム, 情報処理学会第 81 回全国大会 (2019).
[4] Aditya Ramesh et al.: *Hierarchical Text-Conditional Image Generation with CLIP Latents*, arXiv:2204.06125 (2022).
[5] Chitwan Saharia et al.: *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*, arXiv:2205.11487 (2022).
[6] 山西良典, 鍵田里沙子, 西原陽子, 福本淳一: 共起語の特異性と繰り返しに着目した歌詞からの印象的フレーズ抽出, 日本感性工学会論文誌, Vol.14, No.1(特集号), pp.29-35(2015).
[7] AVA: The Art and Science of Image Discovery at Netflix, 入手先 ⟨https://netflixtechblog.com/ava-the-art-and-science-of-image-discovery-at-netflix-a442f163af6⟩ (2022.08.08).
[8] Alec Radford et al.: *Learning Transferable Visual Models From Natural Language Supervision*, arXiv:2103.00020 (2021).
[9] Ian J. Goodfellow et al.: *Generative Adversarial Nets*, arXiv:1406.2661 (2014).
[10] Xingchao Liu et al.: *FuseDream: Training-Free Text-to-Image Generation with Improved CLIP+GAN Space Optimization*, arXiv:2112.01573 (2021).