芸術科学会論文誌 Vol. 20, No. 2, pp. 129 – 138

ボーカルメロディに応じた歌詞生成

宮野友弥 1)(非会員)　斎藤博昭 1)(正会員)

1) 慶應義塾大学大学院理工学研究科

Lyrics generation from a vocal melody

Tomoya Miyano[1)]　**Hiroaki Saito**[1)]

1) Graduate School of Science and Technology, Keio University

miyano_tomoya380 @ keio.jp

概要

ポピュラー音楽のボーカルメロディから，聴きやすさを考慮した自動歌詞生成についての研究を行った．歌詞については散文的意味に加え，音楽的な要素の双方を同時に考慮する必要がある．そのため生成段階においても評価段階においても困難な課題であると言える．本稿では人間の手助けを必要とするサポートシステムという形ではなく，ボーカルメロディを入力するだけで言語モデルが自動的に歌詞を出力するシステムを構築する．人間による解析を用いることなく，データを学習してメロディに対して柔軟性を持った歌詞生成することを目指した．具体的には音符列をシーケンスとみることにより，機械翻訳によく用いられる seq2seq と Transformer を適用した．評価には単語密度という簡易版の尺度を導入するとともに，各言語モデルが出力した歌詞をソフトウェアで音声合成し，7 人の被験者による主観評価で，聴きやすさ，意味，全体的なクオリティについての評価を行った．また，データ数による学習到達度の違いにも簡易的に評価して考察した．全ての手法でテスト曲の歌詞生成時にある程度の言語的な乱れが見られたが，その中では Transformer が最もメロディに応じた自然な歌詞を生成した．

Abstract

This paper reports research on automatic lyrics generation considering listenability from a vocal melody of popular music. In lyrics generation, it is necessary to consider both the meaning and the melody at the same time, which makes this task difficult both in the generation and evaluation steps. We built a system in which the language model automatically generates lyrics just from a vocal melody. We aimed to learn the data and generate lyrics with flexibility for the melody without using human analysis. By considering the notes as sequences, we applied seq2seq and Transformer, each of which is often used in machine translation. We introduce word density as a quantitative measure of evaluation. The lyrics output by each language model was voice-synthesized, and seven participants performed a subjective evaluation on listenability, meaning, and overall quality. We also briefly evaluated the difference in learning achievement depending on the number of data. All the methods showed some linguistic disturbance in generating the lyrics of the test songs. Among them, Transformer generated the most melodious and natural lyrics.

1 はじめに

ポピュラー音楽において，歌詞の 1 音 1 音がメロディに対して自然に乗ることは非常に重要である．しかしメロディと歌詞の対応付けデータは少ない上に，その評価としても歌詞がメロディに合っているかを機械的に計測することは困難である．ポエムやラップ歌詞といったジャンルでは，複雑な音韻規則などに沿って生成と評価が行われることが多い．しかしポピュラー音楽は定められたルールが少ないために，解析して生成を行ってみても提案手法によってどの程度結果が向上したのかを評価することが難しい．歌詞生成においては，従来，ストーリーやテーマ，音節数や韻など，一つの評価軸に沿って研究がなされてきた．しかし文字列や形態素列，単語列として歌詞を生成していく際に複数のタスクを同時に満たすことは難しく，それらを統合する研究もあまり行われていない．本稿では深層学習による自動的な解析をもとに歌詞を生成し，言語的な意味合い評価に加えて，日本語のアクセントをもとに音楽的な意味合いについても評価した．

2 関連研究

現在までに音楽理論を基盤とした情報処理研究は多数行われてきた．生成系の研究としては，メロディを言語条件なしに生成する場合，歌詞からメロディを出力する場合，メロディから歌詞を出力する場合に大別される．言語条件なしのメロディ生成研究として，石田ら [1][2] は旋律データベースと N-gram を用いて，人間による即興演奏の補正を行うシステムを考案した．この研究では被験者に即興演奏をしてもらい，補正すべき箇所を人手でラベル付けした上で，提案システムがどの程度正しく補正音を選べたかを評価している．深山ら [3] は日本語歌詞に由来する条件をもとに旋律を構成する研究を行った．ここでは Beckman ら [4] の，日本語のアクセントは強弱よりもピッチの上下を主とするという指摘を旋律の制約条件に用いている．本稿でも同様の制約条件を用いて，生成された日本語歌詞がどの程度旋律に沿っているのか判定することを試みた．

一方，メロディをもとにした日本語歌詞生成研究は近年の深層学習ブームが始まる前から行われてきた．とくに研究の初期段階から現在に至るまで，作詞支援システムとして人間と機械の共創を目指す研究がなされてきた．多くの場合では音節や，音節を部分的に分解した単位であるモーラ数などを主眼におき，メロディの休符間に適切なモーラ数の単語を配置するアルゴリズムが考えられた．これらの研究は人間と協力しながら音節数やモーラ数を調整するシステムとして有用であるが，モーラ数などの条件が合致すればメロディに適合し，かつ意味の通る歌詞が出力できるとは必ずしも言えず，人間の作為的なコントロールを不要とする自動歌詞生成の領域まで踏み込めてはいなかった．人為的なコントロールを必要としない研究としては，伊藤ら [5] の，日本人に馴染みのあるメロディに暗記したい単語を割り当てるシステムがある．深層学習を用いないシステムでは，歌詞を生成するというよりも，事前に与えられたメロディと単語を紐づけるというので限界であったと考えられる．Watanabe ら [6][7][8] は単語をラベル付けしてストーリー展開を考慮した研究，モーラ数に加えて歌詞の抑揚に着目する研究，メロディ条件付き言語モデルの研究などをそれぞれの観点から行っている．中でもメロディ条件付き言語モデルの研究は，これまで着目されてこなかった深層学習を用いた自然言語処理技術の応用であり，N-gram 言語モデルなどで候補を掲示するよりも自然な文生成を行えることが期待される．ここでは休符の位置と形態素のモーラ数の関係性を学習するために，再帰型ニューラルネットワーク (recurrent neural network: RNN) を用いて形態素と形態素のモーラ数を同時に予測するモデルを構築している．本稿ではメロディ条件からの言語生成だけに着目し，休符を除いた条件で平仮名と音符の関係性を学習した．RNN 言語モデルをさらに発展させ，音符列をシーケンスとみることでニューラル機械翻訳 (Neural Machine Translation: NMT) と同じ手法を適用し，音符の順番を考慮してどこに注意するかを学習する機構を持たせた．データから学習することで人間による歌詞の解析と複雑な条件付けを用いることなく自動歌詞生成することを目指す．

3 実験

図 1 に示すのが，音声合成までを含む自動歌詞生成システム全体の流れである．本稿ではまずメロディ条件付き言語モデルで歌詞を生成し，結果が音楽的制約条件をどの程度満たしているのか判定するとともに，既存の音

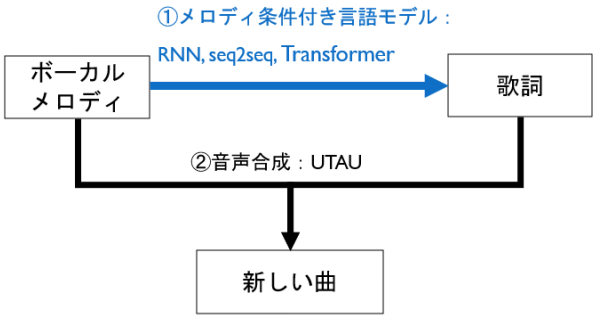

図 1　本稿システムの流れ

声合成システムを利用して主観評価した．

3.1　音声合成ツール

本稿で用いた UTAU は Windows 向けの歌声合成ソフトウェアで，音声ライブラリを元に歌唱を組み立てる機能を持つ．UTAU に向けて作成された音声ライブラリがネット上に非常に多く存在しており，それらをダウンロードして使用することで幅広い声質の歌唱を実現することが可能である．

3.2　データセット

UST ファイルは Web 上で有志が配布している UTAU 用の歌詞・メロディデータファイルである．本稿ではこれを 440 曲分集め，データセットとして利用した．図 2 に示すのが UST ファイルの例である．ファイルの冒頭で [#SETTING] としてテンポや用いる歌声データなどを指定し，[#0000] 以降にボーカルメロディの音の長さ，歌詞（文字），音符が 1 音ずつ書き込まれている．本実験では 430 曲を学習に用い，残りの 10 曲について評価した．また，その他にもビブラートなどの細かい設定が可能だが，本実験ではそこまでは触れないものとした．無音の区間については，それを学習に含める研究もあるが，歌詞データ数が少なく学習が難しいため本稿では除外した．その上で，メロディーが存在する部分についてのみ学習と実験を行った．また音声合成用のデータファイルなので，「何しよーかな雑誌めくって」が「なにしよおかなざあしめくうて」となるなど，必ずしも歌詞をそのまま平仮名にしたものと同じとは限らないことに注意する．

3.3　各言語モデルへの入出力処理

Watanabe らの研究では前後の音符列の他に形態素列を入力して，形態素とモーラ数 s を出力するモデルを用

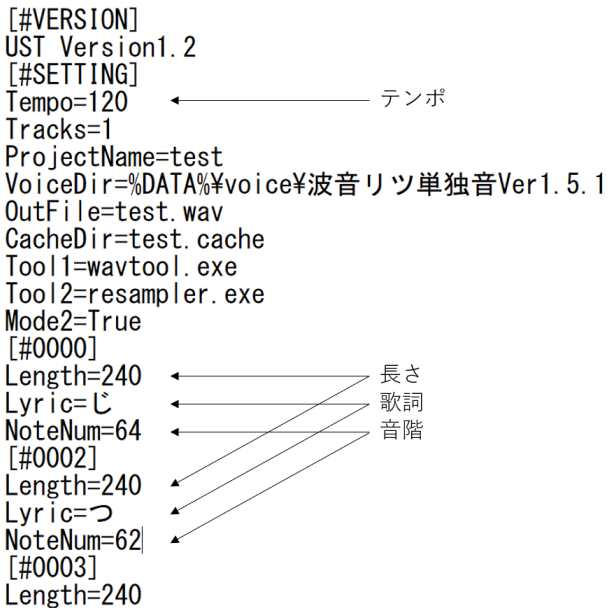

図 2　UST ファイルの例.

いていた．本稿では言語モデル自体の比較が目的であるため，文字で入出力を行うように簡素化して実験する．学習用のため，1 音ずつスライドして，10 音ごとの文字・音符データセットを作成した．10 音に対して 10 文字を出力することを 1 ステップとして学習，生成を行う．Teacher forcing を利用し，学習時の入力はデータセットをそのまま用いて，正解とのロスを計算する．一方で評価時には予測の最大値を出力し，次の入力に用いる．生成時にモデルに入力する最初の歌詞 1 文字は，テストデータ元曲の歌詞の 1 文字目を利用し，ステップごとに出力される最後の 1 文字を，次のステップ最初の入力文字として用いた．3.2 節で示したように，UST ファイルの音階は整数，歌詞は平仮名で表される．したがって歌詞（平仮名列）を各モデルに入力可能にするため，データセットの出現回数順の単語インデックスを用いて整数列に変換し，バッチサイズ 64 で各メロディーに対応する整数化した歌詞列と音階列のペアを作成した．

3.4　実験手法

本稿ではデータサイズによる学習結果の違いについて調べるため，まず各手法に対して 40 曲，100 曲，200 曲，300 曲，430 曲で学習したモデルそれぞれを，テストデータ 10 曲に対して用い，生成された歌詞の単語密度を調べた．次に 430 曲の場合で生成された歌詞についてテストデータ元曲ファイルの書き換えを行い，それら以外

の音符や長さについては同じ条件で音声合成ソフトウェア UTAU に歌わせるものとした．UTAU にはできるだけ自然な女性の声を発声できるものを使用し，音声合成の質が結果を左右しないように努めた．本稿に参加してくれた評価者は 7 人で，前節で述べた主観評価の尺度に従って歌詞を評価してもらった．評価者はスペースで 1 音ずつ区切られた平仮名歌詞を見ながら各 3 項目の評価を行った．

4 手法

Watanabe ら [8] の研究では，メロディ条件付き言語モデルのベースモデルとして LSTM が用いられていた．ここでは歌詞メロディ対応データセットが小さいという難点を克服するために，ベースモデルをもとに歌詞のみのデータと歌詞メロディ対応データを組み合わせる手法が提案されている．具体的にはファインチューニングや歌詞のみのデータに疑似メロディを付与する手法が提案され，ベースモデルよりもメロディと歌詞が合った生成をしやすいことが示されている．本稿ではメロディからの歌詞生成を，機械翻訳と同じシーケンス変換とみることにより，深層学習を利用したニューラル機械翻訳であるエンコーダ・デコーダモデルの活用を試みた．比較対象とする LSTM を用いた RNN 言語モデルに加え，Bahdanau ら [9] の注意機構付きの seq2seq と，Vaswani ら [10] の自己注意機構にもとづく Transformer を用いて実験した．seq2seq モデルでは，エンコーダ部でメロディをエンコードするが，デコーダ部は LSTM を用いた RNN 言語モデルとほぼ同一（メロディベクトルは当然入力しない）であるため，純粋にエンコーダを用いた結果として生成がどう変わったのかを比較することができる．メロディをベクトル化して入力する RNN モデルに比べ，seq2seq や Transformer ではエンコーダで音符の順番を把握し，シーケンスとしてどこに注意するかを学習する機構を持つため，より優れた結果が期待される．なお，これらのモデルでは入出力の系列長が必ずしも一致しなくてよいという機械翻訳上の利点があるが，今回は入力音符数と出力文字数は一致するので，言語モデルの生成は決められた数だけ行った．また，本稿では全ての RNN 層に LSTM を利用した．

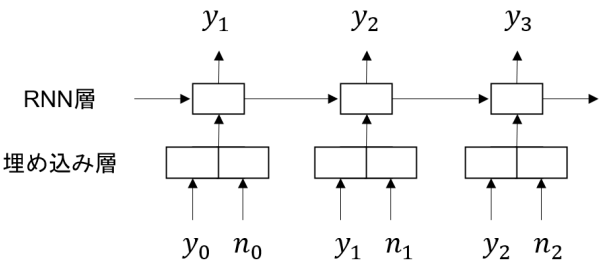

図 3　RNN モデルの概略図

4.1 RNN

まず本稿の比較対象として用いた RNN モデルについて解説する．図 3 に示すのが RNN モデルの概略図である．ここで，n_t は各タイムステップにおける歌詞 y_t の前後合わせて 10 音 $(m_{t-5}, ..., m_{t+4})$ である．歌詞 1 文字 y_t と n_t についてそれぞれ埋め込みベクトルに変換し，RNN 層に渡して次の 1 文字 y_{t+1} を出力する．形態素ではなく文字を入出力に用いることを除き，先行研究である Watanabe らのモデルにできる限りネットワークを合わせた．RNN モデルでは，文字 y_t の予測分布を前の文字 y_{t-1} と音符列 n_t，前の隠れ層の出力ベクトル s_{t-1} から推定する．つまり，単語系列 $Y = y_1, ..., y_T$ に対する生成確率 $P(Y)$ を次式のように計算する．

$$P(Y) = \prod_{t=1}^{T} P(y_t | y_{t-1}, n_t, s_{t-1}) \tag{1}$$

ここで s_{t-1} は，これまでのコンテキスト情報 y_1^{t-2} 及び n_1^{t-1} を埋め込んでいることに相当し，これにより長距離コンテキストを考慮したメロディ条件付き言語モデルを実現している．

4.2 seq2seq

次に本稿で用いた注意機構付きの seq2seq について述べる．図 4 に示すのが seq2seq モデルの概略図である．エンコーダとデコーダは埋め込み層と RNN 層を持つ．単純な seq2seq では各時点の隠れベクトルがすべての情報を保持しなければならないという制約があり，長い文ほど翻訳精度が下がるという問題点があった．そこで導入された注意機構は，デコーダのあるステップにおいて，エンコーダのどの時点での情報を利用するかを学習する仕組みである．これによって隠れベクトルの保持する情報を分散できるようになった．音符列 $m = (m_1, ..., m_N)$ をエンコーダへの入力とし，注

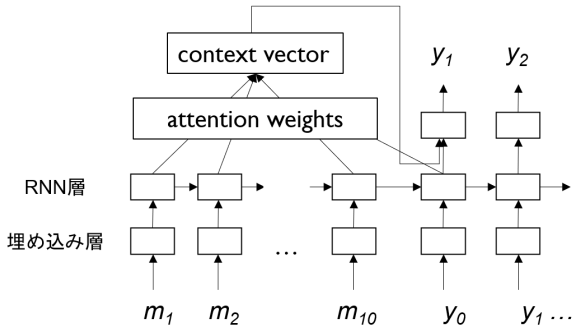

図 4　seq2seq モデルの概略図

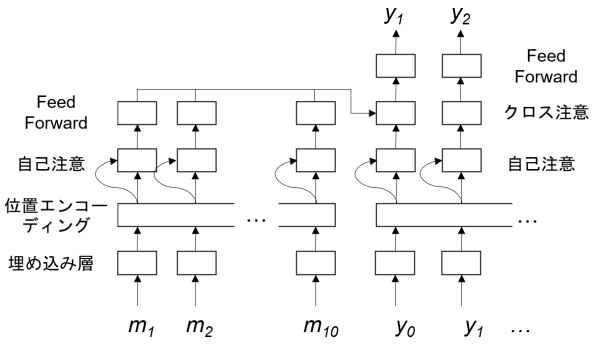

図 5　Transformer モデルの概略図

意機構での計算結果をもとに，目的である歌詞の文字列 $y = (y_1, ..., y_N)$ を出力する．なお，ここでは $N = 10$ である．系列情報であるメロディを各タイムステップごとに RNN 層に入力することで，相対的なメロディの流れを考慮した学習ができる上，注意機構を保持していることで入力された音符列のどこに注視すべきかを認知することが期待される．

4.3　Transformer

Transformer は前述の seq2seq で用いていた RNN 層を用いず，自己注意機構を使って学習する．逐次的に計算を行う必要がある RNN 層を持つ seq2seq に比べ，自己注意機構でシーケンス変換を行う Transformer は並列化が容易で学習が高速である．機械翻訳のスコアでも高い結果を出し，様々なタスクへの応用が期待されている．図 5 に示すのが Transformer モデルの概略図である．ここでも seq2seq と同じように音符列 $m = (m_1, ..., m_N)$ を入力とし，目的である歌詞の文字列 $y = (y_1, ..., y_N)$ を出力する．単に各単語を独立にエンコーディングするだけでは文脈の情報を利用することができず，また固定数の周辺単語のみを補助的に利用する仕組みでは長距離の文脈を考慮できないため，自己注意という注意機構を導入することにより，入力文内，出力文内の文脈情報を考慮した翻訳を可能とする．seq2seq の注意機構がデコーダがエンコーダ側の情報を選択的に参照する方式であったのに対し，自己注意機構はエンコーダが入力文内の他の箇所の情報を，デコーダが出力文内の他の箇所の情報（デコーダ学習時には未予測の単語をマスクする）を，それぞれ選択的に参照する方式である [11].

5　評価

毎回生成された歌詞に対して主観評価を行うことはあまりに被験者の負担が重く現実的ではない．また，簡単な結果を迅速に得たいときに不便である．そのため本稿では主観評価に加え，簡易版としてのテキスト評価という尺度を導入した．しかしながらテキスト評価には後述の問題点が挙げられるため，あくまで研究途中で行う簡易結果であり，最終的な評価はすべて主観評価によるものとした．

5.1　テキスト評価

テキスト評価では生成した歌詞について簡易的な評価を行う．恣意的な評価を避けるため，単語密度を WD，単語として成り立つ文字数を c，全体の文字数を a として以下の基準を設けた．

$$WD = c/a \tag{2}$$

この評価の問題点は大きく三つ挙げられる．

- 歌詞の中で意味の分かる単語の割合と歌詞全体としての評価は必ずしも一致しないこと
- メロディーに応じた歌詞生成であるのか判別できないこと
- 単語かどうかの評価が曖昧であること

5.2　主観評価

主観評価実験では，7 人の被験者に対して 3 手法×10 曲をスペースで区切られた平仮名歌詞を見ながら聴いてもらい，表 1 参照の 3 項目について表 2 参照の 5 段階で評価した．評価した 3 つの項目について簡単に述べる．Listenability (L) とはメロディに対する歌声

表 1　評価項目.

評価項目	概要
Listenability (L)	メロディに対する歌声の自然さ
Document-level meaning (DM)	文字をみて意味のとれる割合
Overall quality (OQ)	全体的なクオリティ

表 2　評価の 5 段階.

レベル	概要
1	まったくダメ　ランダムに近い
2	10% くらい妥当と言える個所があった
3	30% くらい妥当と言える個所があった
4	50% くらい妥当と言える個所があった
5	70% 以上妥当と言える個所があった

の自然さで，歌詞の意味に関わらずどの程度歌詞とメロディが適合しているのかを調べた．Document-level meaning (DM) は歌詞の文字をみて意味のとれる割合であり，音声的な評価に関わらず文としての評価を求めた．Overall quality (OQ) は全体的なクオリティであり，歌詞や歌声として総合的にどの程度良い生成ができたのかを評価してもらった．なお，スコアが低評価に偏りすぎて各手法の評価差が出なくなることを避けるため，真ん中の評価であるスコア 3 を「30% くらい妥当と言える個所があった」として実験した．

5.3　音楽的評価

2 節で述べたように，日本語のアクセントは強弱よりもピッチ（音程）の上下を主とする．これを日本語歌詞のメロディに対する制約条件として，どの程度各単語がメロディに合致しているかを評価した．歌詞として破綻している箇所は判定には含めず，5.1 節の評価で単語として成り立つとしたものについてのみ評価した．具体的には，日本語歌詞の各単語についてアクセント辞書で音程の上下を調べ，それがボーカルメロディに沿っているかを一つずつ判定した．アクセント辞書には峯松らによる OJAD（日本語教育のためのコーパスを利用したオンライン日本語アクセント辞書）を用いた．部分的に破綻した歌詞に対する評価を定めるため，本稿では評価に AAMN(平均アクセント合致数) と AAMR(平均アクセント合致率) という評価尺度を用いることにした．1 曲ごとのアクセント合致率は式 (3) で計算する．

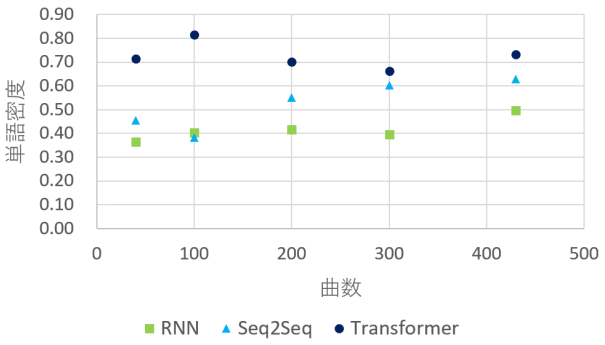

図 6　学習データの曲数と単語密度の関係

$$\text{アクセント合致率} = \frac{\text{アクセント合致数}}{\text{歌詞中の総アクセント数}} \tag{3}$$

6　結果

6.1　学習データサイズに関する予備実験

まず，学習データの曲数と単語密度の関係を図 6 に示す．RNN 系の 2 モデルは簡易的な評価であるが，概ね学習データ数が増えるほど生成結果は良くなった．一方で Transformer はほぼ曲数によらず，文字の出力から単語としてより意味の通るものを生成した．

6.2　主観評価の結果

図 7 に示すのが主観評価の結果である．ここでは RNN，seq2seq，Transformer モデルの各曲の L，DM，OQ の平均評価値に対して，Shapiro-Wilk（シャピロ・ウィルク）検定を用いて各データ群が正規性を示すか検定した．表 3 に示すのがその結果である．全ての p 値が

表 3　各データ群に対する Shapiro-Wilk（シャピロ・ウィルク）検定.

モデル	項目	p 値
RNN	L	0.46
	DM	0.42
	OQ	0.07
seq2seq	L	0.81
	DM	0.34
	OQ	0.95
Transformer	L	0.45
	DM	0.25
	OQ	0.18

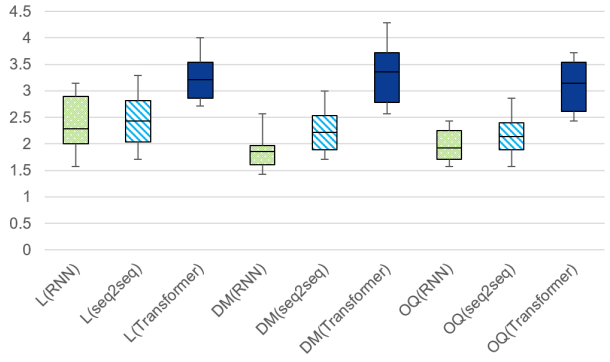

図 7　主観評価の結果

0.05 以上となったため，各データ群が正規分布に従うと判断した.

その結果を踏まえ，各モデルに対する結果に有意差があるかを調べるため，対応のあるデータ群に対する t 検定を行った．本稿ではメロディベクトルをデコーダに入力する RNN モデルより，メロディをエンコードする seq2seq モデルの方が DM について僅かに高くなり，t 検定の有意差がみられた．それ以外の L と OQ に関しては有意差は見られなかった．一方全く構造の異なる Transformer モデルは RNN 系の 2 モデルに比べて，3 項目すべてで明示的に高い結果となった.

6.3　音楽的評価

音楽的評価に関する結果を，表 4 に示す．平均アクセント合致率である AAMR が 0.5 未満である原因としては，歌詞にアクセントが入った場合にもメロディでは同じピッチが続いている場合は合致していないとカウントしたからだと考えられる．こちらについても Shapiro-

表 4　音楽的制約条件

モデル	WD	AAMR	AAMN
RNN	0.50	0.41	4.8
seq2seq	0.63	0.32	4.5
Transformer	0.73	0.32	5.1

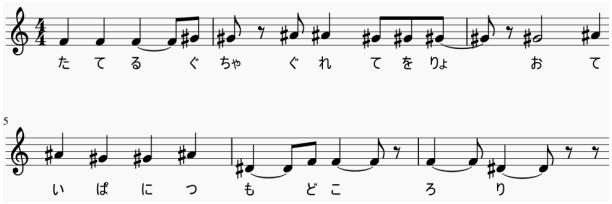

図 8　RNN の生成の一部

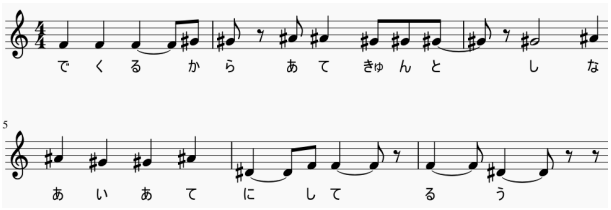

図 9　seq2seq の生成の一部

Wilk（シャピロ・ウィルク）検定を用いて各データ群が正規性を示すか検定し，全ての p 値が 0.05 以上となったため，各データ群が正規分布に従うと判断した．その上で対応のあるデータ群に対する t 検定を行い，AAMR と AAMN について全てのモデルで有意差があることを確認した.

6.4　歌詞の生成例と性質

図 8, 図 9, 図 10 に示すのが，各モデルのテスト生成の一部である．Transformer の生成に関しては楽器経験のある被験者から，音と文字のつながりが適切だったという感想をもらった．概ね，主観結果の通り Transformer, seq2seq，RNN の順に自然な歌詞を出力した．とくに RNN に関しては出力した文字が単語として意味を成すことはあっても，句や文節を形成することは殆どなかったのに対し，Transformer では例のようにかなりまとまった意味を出力する場合が多く見受けられた.

7　考察

7.1　学習データサイズに関する予備実験

まず，予備実験のテキスト評価について考察する．RNN 系の 2 モデルは概ね学習データ数が増えるほど生

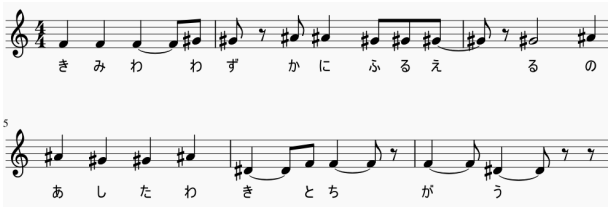

図 10　Transformer の生成の一部

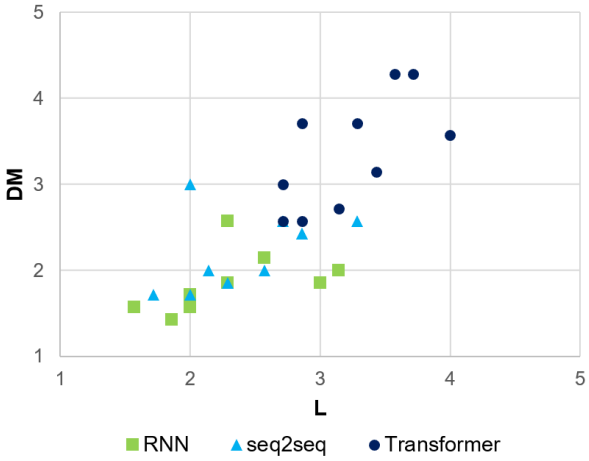

図 11　主観評価の L と DM の関係

成結果が良くなった．そのため，このグラフの範囲では Transformer の方が seq2seq よりもよい値を出したが，データ数を増やすことで Transformer よりもよい歌詞を生成することはありうる．一方で Transformer はほぼ曲数によらず，文字の出力から単語や文としても意味の通るものを生成した．よってデータ数を少しばかり増やしても意味的な改善はあまり期待できない．しかしながら関連研究の節で述べたように，自然言語処理研究では本稿とは比較にならないほど大規模なデータセットでかなり人間に近い文章が出力できるようになっている．歌詞生成においても同様の効果が期待される．

7.2　主観評価の結果

次に主観評価の結果について考察する．比較対象の RNN モデルより，メロディをエンコードした seq2seq の方が DM について僅かに高いスコアを出した．言語を出力する部分のモデル構造がほぼ同一のため，大きな違いが出なかったのではないかとみられる．僅かに DM に差が出た理由としては，RNN モデルへのメロディベクトルの入力が，とくに本稿のようなデータ数が少ない場合には，言語的な生成を妨げるように働いたのではないかと推察される．また，メロディをベクトル化するのではなく，エンコードしたことによる L の違いはほぼ見られなかった．一方，Transformer で比較的高いスコアが出たのは，逐次的な RNN 層の情報伝達を行う RNN 系のモデルに比べて，位置エンコーディング等で情報を扱う Transformer の方がより意味的に自然な歌詞を出力しやすく，メロディとの対応付けもよかったのではないかと考えられる．また，L と DM，OQ のスコアが持つ意味についても考察した．図 11 が示すのは主観評価の結果として，すべてのモデルの L と DM の関係をまとめたものである．L と DM の相関係数は 0.74 という強い正の相関がみられた．ここで DM は平仮名歌詞を見ながらスコア付けしてもらった値なので，L が DM に大きく影響を与えたとは考えにくい．よって，DM が高いほど L の評価が高くなりやすいという関係があったと考えられる．また，OQ は L と DM の値をまとめたものと考えられるので，3 項目すべてがかなり高く関連性をもっており，簡単に分離して評価できるものではないことが示唆される．L と DM を分離して評価するには，例えば平仮名歌詞を母音に変換して評価することが考えられるが，子音の影響を完全に無視してよいのかなど，さらなる詳細な実験評価が必要である．

7.3　音楽的評価

数値上は有意差が出たが実際に耳で聴いて確かめてみたところ，メロディと単語のアクセントが合致していない場合にも，それほどの違和感は覚えなかった．例えば Transformer モデルのテスト生成 1 曲目に出てくる「世界」という単語は，通常は「せ」が高く発音されるが，歌声の場合には「か」でピッチが上がったとしても不自然ではない．歌詞の韻を踏むために単語の発音を変形して歌うのはポピュラー音楽，とくにラップ歌詞によく見られることである．よって先行研究のように，日本語歌詞のアクセントをメロディの生成時に用いるのは妥当であるにしても，本稿のように歌詞の評価として用いるのは適当ではないことが分かった．ただし，どんな日本語文でも必ずメロディに適合するとは考えにくいので，単語ごとに可能なアクセント変化をまとめた辞書を作るなど，より発展した評価手法が必要であると考えられる．

8 まとめ

事前の想定と異なり，メロディをエンコードしたことによる結果は，単純にメロディベクトルを入力する場合と大きくは変わらなかった．また，現状では元の学習データ曲レベルの生成というのには大きな隔たりがあるということが分かった．その上で Transformer を用いた場合には，位置エンコーディングや，自己注意機構の仕組みによって，RNN を用いるモデルよりも意味が通り，かつメロディに応じた歌詞生成ができたのではないかと考えられる．また，日本語のアクセントをもとに歌詞の音楽的評価をしてみたが，上手くいかなかった．本実験ではデータ数がかなり限られた場合について実験を行ったが，曲数が多い場合にどうなるのか調べてみる必要がある．Transformer の発展や，データ数を増やすことで，より自然な歌詞が生成可能になることを期待したい．

参考文献

[1] 石田克久, 北原鉄朗, 武田正之. N-gram による旋律の音楽的適否判定に基づいた即興演奏支援システム. 情報処理学会論文誌, Vol. 46(7), pp. 1548 – 1559, 2005.

[2] 石田克久, 北原鉄朗, 武田正之. N-gram による即興演奏の旋律補正. 情報処理学会論文誌, Vol. 45(3), pp. 743 – 746, 2004.

[3] 深山覚, 中妻啓, 酒向慎司, 西本卓也, 小野順貴, 嵯峨山茂樹. 音楽要素の分解再構成に基づく日本語歌詞からの旋律自動作曲. 情報処理学会論文誌, Vol. 54(5), pp. 1709 – 1720, 2013.

[4] M. Beckman and J. Pierrehumbert. Intonational structure in Japanese and English. *Phonology yearbook*, Vol. 3, pp. 255 – 309, 1986.

[5] 伊藤悠真, 寺田努, 塚本昌彦. Mnemonic DJ: 暗記学習のための替え歌自動生成システム. 情報処理学会論文誌, Vol. 56(11), pp. 2165 – 2176, 2015.

[6] 渡邉研斗, 松林優一郎, 乾健太郎, 深山覚, 中野倫靖, 後藤真孝. ストーリー展開と一貫性を同時に考慮した歌詞生成モデル. 人工知能学会全国大会論文集 第30 回全国大会, 2016.

[7] 渡邉研斗, 松林優一郎, 深山覚, 中野倫靖, 後藤真孝, 乾健太郎. メロディと歌詞の相関に基づく自動歌詞生成. 情報処理学会 研究報告自然言語処理 (NL), pp. 1 – 12, 2017.

[8] K. Watanabe, Y. Matsubayashi, S. Fukayama, M. Goto, K. Inui, and T. 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 (NAACL-HLT 2018)*, Vol. 1, pp. 163 – 172, 2018.

[9] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations(ICLR 2015)*, 2015.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, Vol. 30, pp. 5998 – 6008, 2017.

[11] 須藤克仁. ニューラル機械翻訳の進展—系列変換モデルの進化とその応用—. 人工知能, Vol. 34(4), pp. 437 – 445, 2019.

宮野 友弥

2019 年慶應義塾大学理工学部情報工学科卒業．同年同研究科修士課程に入学し，2021 年修了．

斎藤 博昭

慶應義塾大学工学部数理工学科卒業．現在，同大学理工学部情報工学科准教授．工学博士．音声言語理解などに興味を持つ．言語処理学会，人工知能学会，他会員．