Vol.2021-MUS-131 No.41
Vol.2021-SLP-137 No.41
2021/6/19

LSTMを用いたメロディ素片間の接続コストの算出

平井 辰典[1,a)]　澤田 隼[2,b)]

概要：本稿では，LSTM を用いてメロディ素片間の接続コストを算出するモデルを提案する．これまで，メロディ生成のためにメロディ素片同士の接続コストを考慮する手法は提案されてきたが，本研究では，接続コストの算出を主目的としたモデルを提案する．既存のメロディを小節単位でシャッフルし，その接続箇所を判定できるように学習した LSTM モデルと，接続境界における音符遷移の尤もらしさを基に，与えられた 2 つのメロディ素片（小節）がどの程度自然に接続可能であるかを定量化する接続コストを提案する．本稿では，いくつかのデータ表現方法とネットワーク構成によるメロディ素片の接続箇所判定精度を比較し，接続コストとして適したモデルを探究する．その結果，BiLSTM を用いたモデルで最も高い精度が得られた．本稿では接続コスト算出モデルの提案に加え，接続コストを用いることでどのように音楽制作を支援することができるのかについても検討する．

1. はじめに

作曲を行う際，1 コーラス分のメロディを 0 から紡ぎあげていくという作業が必要となるが，実際にメロディを制作する際には 1 コーラス分には至らない短いフレーズ単位のメロディを思いつくことが多い．短いフレーズであれば鼻歌を歌うような感覚で比較的容易にメロディを思い浮かべることができるため，この行為自体は作曲に精通している人でなくてもできるものと考えられる．一方で 1 曲分のメロディを最初から最後まで作り上げることは誰もが簡単にできることとは言い難い．以上のことから，メロディ制作における困難な作業は，短いメロディ素片同士をうまく繋げる部分にあるのではないかと考えた．そこで，本研究ではメロディ素片同士の繋がりの自然さを測るための尺度として，メロディの接続コストを提案する．

自身が考えた短いメロディ素片を集め，それらを繋ぎ合わせることでより長いメロディを作り上げることができれば，何気なく鼻歌を歌うような行為をより本格的な音楽制作へと繋げることができると考えられる．他にも，自身が考えたメロディに限らずに既存のメロディを対象として，複数の曲からなるメドレーやマッシュアップ音楽を制作する際にメロディのどことどこが自然に繋がりやすいかを測ることができれば制作支援にも繋がると考えられる．このように，音楽制作を支援することを目的として本稿ではメロディ素片の接続コストについて検討する．

一般的に，音楽のメロディを扱う情報処理システムには，その評価が困難であるという大きな課題が存在する．主観的な評価を避けるため，一部の音楽生成に関する研究では，学習データに含まれるメロディを対象とした続く後続メロディの予測精度を用いたり，既存メロディの一部を欠損させて欠損箇所の推定精度を測ったりといったアプローチでモデル構築を行っている．そのような学習方法を採れば定量的な精度を扱ったモデル構築ができるため，音楽生成という結果の評価が難しい文脈においても，システムの良し悪しをある程度推測することができる．一方で，生成された音楽の良し悪しを主観評価実験等により直接評価するようなシステムの場合，そのシステムについて論文に書かれている以上の良し悪しを判断することができない．そのため，新たな手法が提案された際に該当のシステムと同じ条件で比較することができない．

そこで本稿では，メロディの接続コストを定義するにあたり，一部の音楽生成モデルで採用されているような定量的な評価尺度を学習タスクに取り入れ，システムの質をなるべく客観的に評価できるようなモデルを提案する．具体的には，学習データ内のメロディを小節毎にシャッフルし，シャッフルされたメロディとシャッフルされていないメロディをモデルに入力し，シャッフルされているか否かを判定するという学習タスクによりモデルを構築する．これにより，メロディが元々接続されている箇所と不自然に接続された箇所を判定可能なモデルが構築でき，さらにその判定精度を定量的に測ることもできる．また，メロディの小

[1] 駒澤大学
Komazawa University
[2] 東京理科大学
Tokyo University of Science
[a)] thirai@komazawa-u.ac.jp
[b)] sawada@rs.tus.ac.jp

節単位の接続の尤もらしさだけでは，接続箇所における音符間の繋がりが必ずしも自然になるとは限らないため，小節単位の繋がりやすさに加えて，小節を跨ぐ境界における音符の遷移確率を考慮する．これによって，小節単位での繋がりの自然さと，接続境界における音符単位での遷移の自然さを考慮した接続コストを実現する．

2. 関連研究

メロディの接続コストを用いて既存メロディを再利用することにより新たなメロディを生成する手法として，Bretan らは深層学習モデルによるメロディの Unit selection（素片選択）手法を提案している [1]．Unit selection は，text-to-speech の分野で活発に研究されてきた生成手法である．Bretan らの素片選択手法では，意味的な関連性と接続コストの二つを考慮することで，入力されたメロディに続くメロディをデータベース内から検索していくことでメロディ生成を行っている．まず，意味的な関連性を求めるために，bag-of-words のアプローチによって抽出された特徴量を対象として二層のオートエンコーダによって 500 次元の埋め込みベクトルを獲得する．その埋め込みベクトルの遷移に関する LSTM モデルを構築することで素片毎の意味的な関連性が求められる．次に，データベースの中から入力メロディに対して意味的な関連性が高い上位 5% のメロディ素片を絞り込み，それらに対して意味的な関連性と音符レベルの接続コストを求め，二つの尺度の組み合わせによって後続のメロディ素片を選択する．音符レベルの接続コストは，多層 LSTM によって次の音符を予測するようなモデルを学習することにより求められている．以上をまとめると，この手法では，まず入力メロディに対して意味的に近い上位 5% のメロディをデータセットから絞り込み，それらを対象に，意味的な近さと音符単位の接続コストを基に次のメロディの素片を選択していくことでより長いメロディを生成しているということである．本稿では，メロディの生成ではなく，メロディ素片同士の接続コストを算出することが主目的であるため，Bretan らの手法において素片同士の意味的な関連性を求める部分についても接続コストという観点に置き換えられないかを検討する．また，音符単位の接続だけではなく，より長い時間幅を考慮した接続コストを検討する．

既存のメロディを繋ぎ合わせることで新たな音楽を生成するというアプローチは Cope によっても提案されており，楽曲を細かくセグメンテーションし，それぞれの素片の特性に基づいてラベル付けをすることで，再利用，再結合をしながら新しい楽曲を制作する手法を提案している [2]．Cope による試みは，本稿で提案する手法や前述の Bretan らによる手法とは違い，機械学習等によるモデル化を行わずにルールに基づいてメロディの再構築を行うというアプローチとなっている．

既存のメロディを利用するアプローチは，これまでにも音楽生成において行われてきている．Kitahara らが提案した JamSketch では，ユーザが入力したメロディの概形に基づいて，既存のメロディを基に遺伝的アルゴリズムによって即興演奏のメロディをリアルタイムで生成している [3]．JamSketch では既存のメロディをそのまま使用しているわけではないが，既存のメロディをメロディ生成に活用している一つの事例と言える．他にも，Pachet が提案した The Continuator も既存のメロディを基に新たなメロディ生成を実現するシステムである [4]．The Continuator は，メロディを細かい素片に分割し，素片から素片への遷移を木構造のマルコフ連鎖によりモデル化することで，入力されたメロディに続くようなメロディとして適したものを学習データの中から探索するインタラクティブなシステムである．

既存のコンテンツを再利用することによって新たなコンテンツを生成するという試みは，メロディに限らず様々なドメインで試行されているアプローチである．例えば，画像合成手法である PatchMatch は，画像補間の処理として，補間領域に対応する小さなパッチ領域を同一画像内から探索して組み合わせることで補間された新たな画像コンテンツの合成を実現している [5]．他にも，平井らによる音楽動画の自動生成手法では，既存の動画データベースから，音楽に合うような動画素片を探索してそれらを繋ぎ合わせることで新たな音楽動画の自動生成を実現している [6]．

このように，コンテンツの再利用によるコンテンツの生成というアプローチは活発に行われている．本研究では，音楽メロディの再利用による新たなメロディの生成を実現するための要素技術ともなりうるメロディ素片同士の接続コストの算出方法について検討していく．

3. データの準備

メロディに関する Deep Learning モデルを構築する上で，どのようなデータベースを対象として，どのようにメロディデータを表現するかを決めることは非常に重要である．この選択によっては，精度や結果が大きく変わってしまうためである．また，これらは手法の再現性にも関わる内容であるため，本章で一つずつ詳述する．

3.1 データセット

本稿では，The Lakh MIDI dataset[7] から抽出したメロディデータを使用する．The Lakh MIDI dataset は，Web から収集された 176,581 曲分の MIDI ファイルによるデータセットである．本研究で扱う対象はメロディであるため，このデータセットの中から，メロディのみを抽出する．メロディの抽出は Hirai と Sawada によるメロディの分散表現学習手法 [8] の前処理に則って行った．その結果，10,853 曲分のメロディが取得でき，その中でも，メロディ

の長さが 2 小節分（16 分音符 32 個分）以上の計 10,736 曲分のデータを使用することとした．さらに，上述の Hirai と Sawada の前処理手法により，取得したメロディの調を推定し，必要に応じて移調をすることでデータセット内のメロディの調の統一を図っている．

The Lakh MIDI dataset から得られるメロディの情報には，演奏表情が付加されているために楽譜通りの記号情報が得られない MIDI ファイルや，適切でないラベル付けによるノイズのようなデータも多く含まれている．よりクリーンなデータセットもあるが，そのようなデータセットでは楽曲数が絞られてしまうため，本稿ではなるべく多くのデータを対象に学習を行えることを優先した．また，The Lakh MIDI dataset は著者がこれまでに行ったメロディを対象とした生成 Deep Learning モデルの比較研究 [9] においても採用しているデータセットであるため，今後精度の比較実験等を行う際にも比較がしやすいということもデータセット選択の理由である．

3.2 メロディデータの表現方法

メロディは音高と音価との組み合わせからなる音符によって構成されるものであり，音符列によって表現される．MIDI データには演奏時の強弱を記録したベロシティなどの情報も付加されているが，本稿では音符列のみを処理の対象とする．音符列は，MIDI データからメロディを抽出した段階では，MIDI ノートナンバーとティック（4 分音符一つ分の長さ）によって表現されている．抽出して調を統一したメロディは，テキスト形式の音符列として処理を行う．

次にこの音符列を，接続コスト算出のためのモデルに入力可能な形式で表現する必要がある．データ表現方法の決定にあたって，本稿で提案するメロディ素片の接続コストを算出するための戦略を決める必要がある．詳細については後述するが，本稿で提案する接続コストでは，小節レベルでの接続の妥当性と音符（ノート）レベルでの接続の妥当性の二つの要素に注目して，それらを統合するようなモデルを構築する．

小節レベルの接続の妥当性を評価するモデルのデータ表現に関しては，1 小節の長さを 4 分音符 4 つ分（4 分の 4 拍子）と仮定し，メロディデータの先頭から 4 分音符 4 つ分の長さ毎にメロディをセグメンテーションする．さらに 1 小節分の音符列は，16 分音符毎に分割して，音名のみを記録した One-hot vector で表現する．この One-hot vector は，12 音の音名と休符を含んだ 13 次元のベクトルとなっており，これが 1 小節あたり 16 個あるため，13×16 の行列となる．ここで，小節を 16 分音符単位に分割してしまうことによって，同じ音名の音符が続けて発音された際の音符の境界に関する情報が消えてしまう．それを防ぐために前に発音された音が継続されている状態を示す hold という状態を導入することが多いが，本手法では，hold を導入した場合としなかった場合で実験した際に，導入しなかった場合の方が学習がうまくいった．そのため，小節レベルの接続の妥当性評価モデルについては上述の 13 次元の One-hot vector 表現を採用している．オクターブの情報を無視して音名のみを採用している理由としては，オクターブのばらつきに起因する入力データの状態の多様化によって学習がうまくいかないことを防ぐためである．

一方，ノートレベルの接続の妥当性を評価する際のメロディのデータ表現については，オクターブ情報や音符の長さについても重要となる．そこで，音名だけでなくオクターブ情報，さらに音の長さも考慮した音符（音高＋音価）をデータ表現の対象とする．メロディ内に登場する多種多様な音符を扱うために，メロディを文章と見立てて音符を単語としてボキャブラリー化する．まず，データセットのメロディ内に含まれるすべての音符を対象として，それぞれの音符に ID を付与する．ここでオクターブの情報を考慮している理由としては，実際にどの音符からどの音符に遷移する傾向にあるかという事例を基に妥当性を判定するためである．候補となる音符の遷移パターンがデータセット内に多く含まれていれば含まれているほど，妥当な遷移であると判定できるようなデータ表現を行う．

4. メロディ接続コスト算出モデル

前章で述べたように，メロディの接続コストの算出にあたって，小節レベルの接続の妥当性とノートレベルの接続の妥当性を考慮する．実際には小節とノートの二種類に限らず，自由な解像度で接続の妥当性を考慮することも可能であると考えられるが，その拡張については今後の課題としている．本章では，小節レベルの接続の妥当性，ノートレベルの接続の妥当性，それらを統合する手法についてそれぞれ述べていく．

4.1 小節レベルの接続の妥当性

3.2 節で述べた小節単位の One-hot vector を入力とした小節レベルの接続の妥当性を判定するモデルを構築する．実装するのは，任意の 2 つの小節のメロディが入力されたときにその 2 つのメロディが元々繋がっているメロディ同士なのかを判定するような Deep Learning モデルである．

メロディに関するモデルを構築する際には，時系列情報を考慮可能な RNN をベースとしたモデルが望ましいため，本稿では複数種類の RNN モデルを実装し，判定精度を比較して最終的に採用するモデルを決定する．

本稿で構築した RNN ベースのモデルのネットワーク図を**図 1** に示す．本稿では複数種類の RNN モデルについて比較を行っており，モデルによって図 1 における RNN 層の箇所が LSTM や BiLSTM などに置き換わる．入力データは，16 分音符単位で表現された 12 音+休符の計 13 次

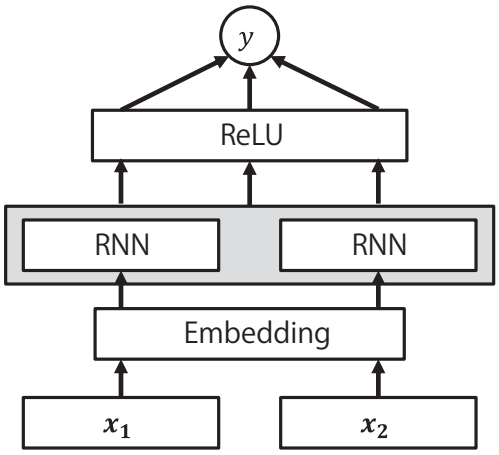

図 1 小節レベルの接続妥当性を判定するネットワークの枠組み

元の One-hot vector（13×16 の行列データ）を 2 小節分（13×32）である．本稿で構築する RNN モデルの枠組みにおけるデータの流れは以下のようなものである．

(1) 入力された 1 小節目のデータ x_1 を Embedding した後に RNN 層に入力し，最後の隠れ層の出力 h_1 を得る
(2) 2 小節目のデータ x_2 についても同じく Embedding した後に RNN 層に入力し，最後の隠れ層の出力 h_2 を得る
(3) h_1 と h_2 を結合し，全結合層（2〜3 層）に通して二値分類の出力 $\{0, 1\}$ を得る
(4) 入力された 2 つの小節が実際に繋がっていた小節の組み合わせである場合には 1，そうでない場合には 0 を教師データとしてネットワークを学習する

上記の枠組みで，3.1 のデータセットのメロディを対象として小節同士の接続判定を行うような RNN モデルの学習を行う．また，(1)，(2) の RNN 層を単純な RNN から，LSTM，2 層の LSTM，BiLSTM などに変えながら判定精度がどのように変化するかについても実験を行う．

学習にあたって，データセットの 10,736 曲分のメロディデータを 9:1 の割合で学習データとテストデータに分割する．学習データをそのまま利用するとすべてのデータが正例（前後の小節同士が接続関係にある例）となってしまう．そこで，データセットのメロディ 1 曲毎に，元のメロディのままの正例のデータと偶数番目の小節をすべてランダムなメロディに置き換えた負例（前後の小節同士が接続関係にない例）のデータを用意した．これにより，利用可能なデータの数が 2 倍に増える．ランダムなメロディへの置き換えにあたっては，完全にランダムなメロディを使用するのではなく，データセット内の別の曲の別の小節のメロディと置き換えている．

いずれの RNN モデルにおいても，Embedding の次元数は 25，RNN の隠れ層の出力の次元数は 50 とした．BiLSTM の場合のみ，隠れ層からの出力の次元数が 2 倍となる都合上，最後の全結合層の数を一層分増やしている．どのモデルでも損失関数を二値交差エントロピーとして，200 エポックの学習を行った．

以上のように構築したモデルに対して，任意の 2 小節のメロディデータを入力すると，$[0, 1]$ のスケールで表現される小節同士の接続の妥当性が求められる．本研究ではこれを小節レベルの接続スコア S_m と表現する．

4.2 ノートレベルの接続の妥当性

4.1 節で述べた小節レベルの接続スコアのみでは，小節という粒度での接続の自然さを考慮できるものの，小節境界における音符の接続の滑らかさまでは考慮しきれない．小節レベルの接続スコアを算出するモデルは 1 小節内の音名の変化のみを扱っており，音符という情報が失われてしまっているため，音符の接続についての考慮が必要となる．そこで，小節境界における音符の接続の妥当性を評価する尺度を導入する．

具体的には，学習データに含まれる全てのメロディを小節単位に分割した際の，前の小節の最後の音符と後の小節の最初の音符が何であるのかを基に，小節をまたぐ際の音符遷移としてどのような遷移であれば尤もらしい接続であるかを記録する．3.2 で述べた音符の ID 表現を基に，小節境界において，どの ID からどの ID に遷移するのかをカウントし，音符の種類毎に，遷移の最頻値で除算し，遷移の妥当性に対応する $[0, 1]$ スケールの値を得る．これを，ノートレベルの接続スコア S_n とする．この尺度は，音符の遷移確率に対応するような尺度となっているが，スケールが最大値で正規化されているため，最もありがちな遷移の場合にスコアの値は 1 となるように設計されている．学習データに含まれない音符の遷移の場合にはスコアは 0 となるように設計されており，接続コストを算出する上で，学習データに事例がないような音符同士の接続が起こりづらくなるようにしている．

4.3 メロディの接続コスト

以上の手順で算出した小節レベルの接続スコア S_m とノートレベルの接続スコア S_n を組み合わせることでメロディの接続コストを算出する．S_m と S_n はともに $[0, 1]$ のスケールで，スコアが高いほど接続が妥当であることを示している．そこで，重み係数 α を用いてメロディの接続コスト C を，

$$C = 1 - \{\alpha S_m + (1 - \alpha) S_n\} \tag{1}$$

と定義した．ここで，α は，$[0, 1]$ の重み係数で，値が大きいほど小節レベルの接続を重視し，小さいほどノートレベルの接続の方を重視するパラメータとなっている．本稿では $\alpha = 0.5$ として実験を行う．

5. 結果

本章では提案した接続コストによるメロディの接続箇所判定精度や，モデルの学習に関する比較，実際のメロディ素片の組み合わせに対する接続コストの算出結果について記述する．

5.1 メロディ接続箇所の判定精度の比較

4 章で述べた手法で算出したメロディの接続コストを評価するために，2 小節分の入力メロディが接続されているかどうかを判定するメロディ接続箇所の判定精度を算出して比較を行う．本節では，まず小節レベルの接続スコア算出時の RNN モデルの比較を行った後に，小節レベルの接続スコア，ノートレベルの接続スコア，両者を考慮した接続コストによる判定精度の比較を行う．

5.1.1 RNN モデルの比較

まず，小節レベルの接続スコアの算出時に，どの RNN モデルを採用すべきかを検証するために，各 RNN モデルを用いた際のメロディ接続箇所の判定精度を比較する．データセットの 90% のデータ（小節数ではなく全曲数の 90% ）をモデルの学習に用い，残りの 10% のデータについて，2 小節分のメロディが元々繋がっていたものなのか別のメロディが接続されたものなのかを 4.1 節に記述したモデルにより判定する．RNN 層として，単純な RNN，LSTM，2 層の LSTM，BiLSTM を使った場合の判定精度（正解率）を**表 1** に示す．精度は，各モデルで出力された 0 から 1 の間のスコアを 0.5 で閾値処理し，閾値以上であれば元々繋がっていた小節同士であると判定しているものとした．学習は各モデルで 200 エポックずつ行った．

表 1 に示したように，RNN 層として BiLSTM を採用することで，最も高い 83.62% の精度で元々繋がっていた小節同士なのか，ランダムに接続された小節なのかを判定できた．ランダムな接続の中には，妥当な小節同士の接続がなされている箇所も存在することが考えられるため，この精度は必ずしも 100% に達するとは限らないものである．負例の中にも偶然自然な接続となった小節の組み合わせが含まれている可能性が十分に考えられるためである．ランダムな予測を行った場合の精度は 50% 前後となることから，本手法によって判定する小節レベルの接続スコアはある程度妥当な尺度として使用できるものであるといえる．

また，今回は RNN 層として 4 種のモデルを使用して実験を行ったが，RNN 層の検討と共に，Transformer を導入するなど，ネットワーク全体の構成要素や最適化手法を検討することによって，判定精度にはさらなる向上の余地があると考えられる．今後，より精度の高いモデルを探求する際にも，今回と同様の実験を行うことで，モデルの優劣を客観的に比較することが可能である．

表 1 小節レベルのスコアによるメロディ接続箇所の判定精度の比較

モデル	RNN	LSTM	二層 LSTM	BiLSTM
正解率 [%]	80.77	83.18	83.45	**83.62**

表 2 各スコア/コストによるメロディ接続箇所の判定精度の比較

モデル	小節レベルの接続スコア	ノートレベルの接続スコア	接続コスト
正解率 [%]	**83.62**	78.04	77.12

5.1.2 メロディ接続箇所の判定精度の比較

次に，ノートレベルの接続スコアによる判定精度，さらに両者を考慮した式 (1) の接続コストによる判定精度を算出し，比較する．ノートレベルの接続スコアについても，小節レベルのスコアと同じくデータセットの 90% のデータで音符遷移に関するデータを収集し，残りの 10% のデータで精度を求めている．ノートレベルの接続スコアに基づく接続の判定については，小節境界の 2 つの音符間の接続スコアが 0.1 以上だった場合に自然な接続であると判定するものとした．閾値を 0.1 とした理由は，単にその周辺の閾値における推定精度が高かったからである．

本稿で提案した接続コストに基づく接続箇所判定については，閾値を 0.5 として，接続コストが閾値以下だった場合に自然な接続であると判定し，データセットの 10% のテストデータを対象として判定精度を求めた．ここで，小節レベルの接続スコアについては，BiLSTM によるモデルを採用し，接続コストを算出する上での重み係数 α は 0.5 としている．**表 2** に，小節レベルの接続スコア，ノートレベルの接続スコア，接続コストに基づく接続箇所判定精度を比較した結果を示す．

表 2 に示したように，純粋に小節同士の接続判定を行うという観点では，小節レベルの接続スコアを用いた場合が最も高精度となる．これは，小節レベルの接続スコアが接続判定を行う際の損失を最小化するような学習によってモデル化されていることから，自然な結果であると言える．ノートレベルの接続スコアを考慮し，接続コストを導入した場合，メロディ同士の接続判定タスクの精度は下がってしまうが，それと引き換えにノートレベルの接続の妥当性も考慮できるため必ずしも効果がないとは言い切れない．前の小節の最後の音符と後の小節の最初の音符の遷移の自然さを考慮できることになるため，小節境界部の局所的な自然さが考慮できるという点で提案接続コストの導入にはメリットがあるといえる．また，接続コスト算出の際の重みを調整することで，ユーザがどこに重きを置いて接続コストを算出するかを決めることも可能であり，その調整については，今後インタフェースの実装により柔軟に行えるようにする予定である．

今回の比較は，各スコア，コストに対して閾値処理を行って評価したものである．そのため，閾値やパラメータの値を変えることで結果も容易に変わってしまう．実際

Vol.2021-MUS-131 No.41
Vol.2021-SLP-137 No.41
2021/6/19

表 3　各モデルの学習に要した時間と総パラメータ数

モデル	RNN	LSTM	2 層の LSTM	biLSTM
学習時間	1.0	1.19	1.62	1.73
総パラメータ数	13,126	36,226	77,026	87,126

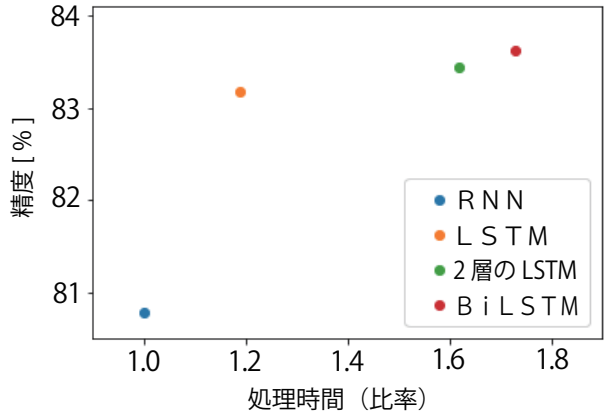

図 2　学習に要した時間と精度との関係

に，ノートレベルの接続判定の閾値を 0.5 にすると，ノートレベルの判定精度は 67.36% にまで下がり，接続コスト算出時の α の値を 0.75 にすると，接続コストによる判定精度は 82.28% に向上する．データについては同じものを使用してはいるが，判定精度の値については完全に同じ条件で比較できているわけではないことに注意されたい．

5.2　学習についての比較

本稿で使用した RNN モデルについての学習に関する情報について比較する．精度の比較については 5.1 節の表 1 に示した通りであるが，本節では各モデルの学習にどれだけの時間を要したか，その結果，どれだけの精度を得られたのかについて記述する．

まず，最も精度が低かった単純な RNN を用いた 200 エポックの学習に要した時間（4 時間 18 分 25 秒）を基準として，LSTM，2 層の LSTM，BiLSTM の 200 エポックの学習にそれぞれどの程度の時間を要したか，また，それぞれのモデルで学習するパラメータ数を**表 3** に示す．表 1 及び表 3 より，モデルのパラメータが増えて複雑になるほど精度は向上するが，その分だけ学習に要する時間は増加する傾向にあると言える．精度と学習に要する時間との関係に注目すると，LSTM によるモデルが処理時間に対する精度向上の幅が大きいと言えるが，この点についてはネットワークの構造を工夫することで結果が変わる部分であり，今後も引き続き検討が必要である．

5.3　既存のメロディ同士の接続コストの算出

提案したモデルを用いて，実際にいくつかのメロディ素片同士の接続コストを算出する．**図 3** に示す m_1 から m_5 の 5 つの小節のメロディの組み合わせに関する接続コストを算出し，その結果を**表 4** に示す．

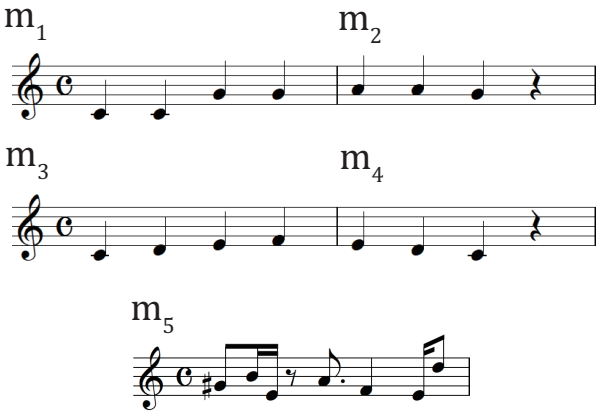

図 3　接続コスト算出実験に使用した 5 つの小節

表 4　既存メロディ同士の接続コストの算出結果

		$\boldsymbol{x_2}$				
		m_1	m_2	m_3	m_4	m_5
$\boldsymbol{x_1}$	m_1	0.560	0.65	0.72	0.78	0.97
	m_2	0.79	0.63	0.69	0.76	0.95
	m_3	0.75	0.74	0.57	0.43	0.89
	m_4	0.73	0.72	0.57	0.62	0.97
	m_5	0.84	0.84	0.79	0.65	0.75

図 3 に示したメロディは，m_1 と m_2 及び，m_3 と m_4 が元々繋がっている小節同士であり，m_5 は筆者が用意したランダムな 1 小節分のメロディである．表 4 において，下線が引かれた値が元々繋がっている小節同士（m_1 と m_2 及び，m_3 と m_4）の接続コストを表している．この結果から，元々繋がっている小節同士の接続コストの値は比較的低い値となっており，そうでない組み合わせの接続コストが比較的に高いことがわかる．同様に，m_4 から m_3 への遷移という，元の接続とは逆の接続のコストも低いが，これらについては，逆順でも自然に繋がるメロディであると考えられる．また，まったく同じ小節同士を接続する場合の接続コストの値も低くなっており，音楽のメロディにおいて繰り返しが多用されることに対応しているものと考えられる．

元々繋がっていない小節同士の繋がりでは，m_5 から m_4 への遷移が比較的低めの接続コストの値となっている．この組み合わせを実際に繋げて聴いてみると，筆者の主観の範囲内であるが，違和感が感じられる遷移ではなかった．最も高い接続コストの値となっている m_4 から m_5 への遷移については，m_4 の最後に休符を挟んでいるため，接続箇所における違和感こそなかったが，まったく別の二つのメロディのように感じられる接続であった．

本手法で得られた接続コストと聴感上の接続の自然さとの関係に関する調査については，今後評価を行っていきたい．

6. 今後の方針

メロディ素片同士の接続コストは，Bretan らの素片選択型のメロディ生成手法 [1] に挙げられているように，メロディの生成に応用することが可能である．任意の入力メロディ素片を基に，それに続く接続コストの低いメロディを順次選択していくことで，より長いメロディの系列を得ることができる．一方で，既存のメロディを再利用することは，新しい音楽を創作することとは異なる行為であると考えられる．そのため，今後の研究において既存のメロディの再利用による新しい音楽の創作の可能性について探求していきたいと考えている．例えば，鼻歌で作ったような短いメロディ素片をいくつもストックしておき，それらを組み合わせることにより長いメロディの制作を支援する手法や，短いメロディフレーズを大量に自動生成しておき，ユーザが制作したメロディの続きを対話的に選択していくことができるインタフェースなどが考えられる．

今後，本稿で提案したメロディ素片の接続コストを応用したメロディ創作支援のためのインタフェースの実現を目指したい．その先に，短いメロディフレーズ単位の作曲の可能性についても探求していきたい．例えば，数十人のユーザがそれぞれ制作した短いメロディ素片を組み合わせることによる，大量の作者による音楽制作行為の実現可能性についても検討していくつもりである．実際に，Splice などのサウンドライブラリでは，大量の短い音素材が公開されており，世界中のクリエータが作る作品の一部として活用されている．メロディのように，楽曲を構成する重要な要素についても，音素材と同様に自身の創作物の一部として取り込むような形で扱える可能性があると考えており，今後その可能性を探究していきたい．

7. まとめ

本稿では，LSTM を用いたメロディ素片間の接続コストの算出モデルを提案した．接続コストは，小節レベルでの接続の妥当性とノートレベルの接続の妥当性を任意の割合で組み合わせられる形で定義した．いくつかの種類の RNN モデルを対象として，メロディの接続箇所であるかどうかを判定するタスクを行い，その精度を比較した．その結果，BiLSTM を用いたモデルでメロディの接続箇所の判定精度が高いという結果が得られた．また，小節レベルの接続スコア，ノートレベルの接続スコア，接続コストについて，それぞれの接続箇所判定精度を算出して比較を行った．本稿で評価尺度として使用した，メロディの接続箇所の判定というタスクは，モデルを評価するための客観的な指標として活用できるもので，今後，より精度が高いモデルの構築を目指して実験を行う場合にも，本稿で採用した評価タスクをそのまま使うことができると考えている．

今後の研究の方向性として，本稿で提案したメロディ素片の接続コストを活用したメロディの創作支援システムの実現を目指している．

謝辞 本研究は JSPS 科研費 JP 19K20301 の助成を受けたものである．

参考文献

[1] Bretan, M., Weinberg, G. and Heck, L.: A unit selection methodology for music generation using deep neural networks, *Proceedings of the International Conference on Computational Creativity 2017* (2017).
[2] Cope, D.: One approach to musical intelligence, *IEEE Intelligent Systems and their Applications*, Vol. 14, No. 3, pp. 21–25 (1999).
[3] Kitahara, T., Giraldo, S. and Ramírez, R.: JamSketch: Improvisation Support System with GA-Based Melody Creation from User's Drawing, *International Symposium on Computer Music Multidisciplinary Research*, Springer, pp. 509–521 (2017).
[4] Pachet, F.: The continuator: Musical interaction with style, *Journal of New Music Research*, Vol. 32, No. 3, pp. 333–341 (2003).
[5] Barnes, C., Shechtman, E., Finkelstein, A. and Goldman, D. B.: PatchMatch: A randomized correspondence algorithm for structural image editing, *ACM Transactions on Graphics*, Vol. 28, No. 3 (2009).
[6] 平井辰典, 大矢隼士 and 森島繁生: 既存音楽動画の再利用による音楽に合った動画の自動生成システム, **情報処理学会論文誌**, Vol. 54, No. 4, pp. 1254–1262 (2013).
[7] Raffel, C.: *Learning-based Methods for Comparing Sequences, with Applications to Audio-to-midi Alignment and Matching*, PhD Thesis, Columbia University (2016).
[8] Hirai, T. and Sawada, S.: Melody2Vec Distributed Representations of Melodic Phrases based on Melody Segmentation, *Journal of Information Processing*, Vol. 27, pp. 278–286 (2019).
[9] 平井辰典: メロディを対象とした生成 Deep Learning モデルの比較, **情報処理学会研究報告音楽情報科学（*MUS*）**, Vol. 2021-MUS-130, No. 15, pp. 1–12 (2021).