Vol.2019-MUS-124 No.2
2019/8/27

教師なしスタイル変換によるメロディーの自動生成

中村 栄太[1,a)]　柴田 健太郎[1]　錦見 亮[1]　吉井 和佳[1]

概要：本稿では，教師なし統計学習を用いて，与えられたメロディーを参照にして目標の音楽スタイル（例えばポピュラー音楽のスタイルや演歌のスタイル）のメロディーを生成する手法について論じる．統計機械翻訳と同様の定式化を行い，目標のスタイルのメロディー生成過程を記述する音楽言語モデルと，参照曲と生成曲との類似性を評価する編集モデルとの統合に基づくスタイル変換の統計的枠組みを提案する．従来の教師あり学習によるスタイル別の言語モデルの構成法では，音楽スタイルを適切に指定するデータを用いることが不可欠であった．人手によるデータ選択と情報付与への依存を低減するため，データから自発的に音高とリズムの構造に関するスタイルを見つけ出す新規の統計モデルを提案する．また，主音などの音符の統語機能を捉える編集モデルを教師なしで学習する方法を構築する．提案する方法が生成曲の品質の向上に有効であることを主観評価および生成楽曲の分析により確認する．

1. はじめに

音楽創作行為を情報学的に理解することは人工知能の問題として研究されており [1–4]，特に近年は自動音楽生成などの応用が盛んに研究されている [5–17]．ある曲を参照して，音楽スタイルを変換することで新たな楽曲を生成する過程（例えば，クラシック音楽のスタイルからポピュラー音楽のスタイルへなど）は多様な音楽を創る過程として重要である [18–22]．この過程に関して少し考察すると，「音楽スタイルをどのように計算論的に定義して記述するか？」や「スタイル変換では参照曲と生成曲との類似性を保つためにどのような音楽的特徴が不変になっているか？」など興味深い疑問が持ち上がってくる．ここでは，これらの問題を統計学習の観点で調べる．また，音楽の事前知識やラベル付きデータをなるべく用いないでメロディーのスタイル変換が可能な方法を構築する．

ここでは，スタイル変換手法が満たすべき条件として，目標の音楽スタイルに適合することおよび参照曲の特徴を保つことの二つを考える．これは，目標の言語に適合し，かつ原文の意味を保つように文を生成することを目指す，機械翻訳の問題と類似している．そこで，本研究では統計的機械翻訳 [23] の定式化に従い，目標のスタイルを表すスタイル別の言語モデルと，参照曲と生成曲との類似性を表す編集モデルとの統合によるスタイル変換の枠組みを提案する [9, 14]．

[1] 京都大学
Kyoto 606-8501, Japan
[a)] enakamura@sap.ist.i.kyoto-u.ac.jp

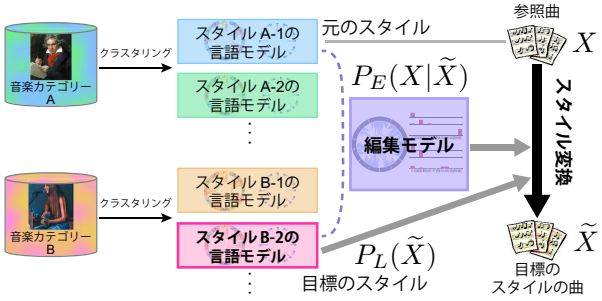

図 1　提案するスタイル変換の枠組み．スタイル別言語モデルと編集モデルは教師なし学習により得られる．

音楽生成に関する最近のほとんどの研究では，スタイル別のモデルは特定の音楽カテゴリー（ジャンルや作曲者など）に属する曲からなる学習データを用いて構築している [8, 11, 12, 17, 20, 22]．しかし，音楽カテゴリーは明確な音楽スタイルとは対応しないことが多い [24–27]．例えば，「モーツァルトの音楽」や「ビートルズの音楽」の中には，異なる音高組織のスタイル（例えば長調と短調など）や異なるリズムのスタイル（例えば 4 ビートや 8 ビートや付点リズムなど）が混在している．スタイルが混ざったデータは，学習されるスタイルを不明瞭にしたり，学習自体を困難にすることから，音楽生成の品質に問題を与えることがある [15–17]．逆に，入念に選択されて（調情報など）ラベル付けがされた学習データを用いた研究では高品質の結果が得られているが [11, 12, 28]，世の中に存在する全ての音楽スタイルに対してこうしたデータを得るには甚大なコストがかかる．人手によるデータ選択とラベル付けのコストを低減するには，データから自発的にスタイルを見出せる

方法が必要である [25,29].

機械翻訳では，編集モデルを構築するために二言語のパラレルコーパスが用いられることが多い．音楽のスタイル変換では，このようなパラレルデータを整備するコストは非常に高いため，教師なし学習による編集モデルの構築が必要である．先行研究では，音楽類似性には音符列の幾何的な距離に加えて，（主音や属音など）音符の調的機能が重要であることが明らかになっている [30,31]．参照曲のスタイルと目標のスタイルが異なり，一方あるいは両方が未知の音楽組織を持つ一般の場合を考えると，こうした音符の統語機能および二つのスタイルの間での機能の関係性を付加情報なしのデータから習得することが大きな課題である．

本稿では，メロディーを対象として，教師なし学習に基づく音楽スタイル変換の枠組みを提案する（図 1）．まず，言語モデルと編集モデルの統合に基づくスタイル変換の統計的定式化を行う．次に，確率的系列モデルの混合モデルの教師なし学習に基づいて，データから音高とリズム組織の特徴をクラスタリングすることでスタイルを発見する方法を構築する．ここでは，拍節構造，音階など音高の移調に対して不変な構造，および音高とリズムの相互依存性を取り入れた構造を持つ，新規のマルコフモデルを構築する．さらに，音符列の幾何的距離と統語機能の両方を捉える編集モデルを構築する．これは二つのスタイル別音楽言語モデルの背後にある共通の統語構造を自発的に学習する隠れマルコフモデル（HMM）を用いて実現する．提案する手法の効果を主観評価および生成楽曲の分析により調べる．

本研究の主な結果は次の通りである．

- スタイル別言語モデルと編集モデルの組み合わせによる，一般性の高く数学的に筋の通った音楽スタイル変換の定式化．同様の問題を扱う先行研究 [18–21] では，編集モデルは定式化されていなかった．
- 音階や典型的なリズムなどの意味あるスタイルを捉える音楽言語モデルの教師なし学習法
- パラレルデータや付加情報付きデータなしで音符の統語的関係を捉える編集モデルを教師なし学習で構築する方法
- 言語モデルの改良と編集モデルの改良の両方がメロディーのスタイル変換の品質を改善することの確認

2. 提案手法

2.1 スタイル変換の統計的定式化

まずメロディーのスタイル変換の問題の定式化を行う．メロディーは音符系列 $((p_{mn}, s_{mn})_{n=1}^{N_m})_{m=1}^{M}$ として表される．ここで，p_{mn} は m 番目の小節の n 番目の音高であり，s_{mn} はその発音楽譜時刻であり，これらは両方とも整数値で表される（M は小節数，N_m は m 番目の小節内の音符数）．ここでは簡単のため 4/4 拍子の曲を扱い，楽譜時刻 s_{mn} は 16 分音符の 1/3 の長さの単位で記述する．また，スタイル変換では各小節内の音符数と各音符のオクターブ範囲は保たれると仮定する．この仮定のもとでは，音高 p_{mn} はオクターブ内の相対値である「ピッチクラス」$q_{mn} \in \{0, \ldots, 11\}$（$q_{mn} \equiv p_{mn} \bmod 12$）で表し，楽譜時刻 s_{mn} は小節内の相対値である「拍節位置」$b_{mn} \in \{0, \cdots, 47\}$（$b_{mn} \equiv s_{mn} \bmod 48$）で表すことができる．よってメロディー X は $X = ((q_{mn}, b_{mn})_{n=1}^{N_m})_{m=1}^{M}$ と表される．メロディーのスタイル変換の手法は，ある音楽スタイルに属する参照メロディー X を目標の音楽スタイルに属する生成メロディー $\widetilde{X} = ((\tilde{q}_{mn}, \tilde{b}_{mn})_{n=1}^{N_m})_{m=1}^{M}$ へと変換するアルゴリズムとして定義される．望ましいスタイル変換の必要条件として，生成曲が目標のスタイルに合致すること，および人間が生成曲を聴いて参照曲を感じられることの二つを考える．

この統計的定式化においては，参照曲 X が与えられた時の生成曲 $\widetilde{X}$ の確率 $P(\widetilde{X}|X)$ をモデル化する．統計機械翻訳 [23] と同様に，この確率を $P(\widetilde{X}|X) \propto P_L(\widetilde{X})P_E(X|\widetilde{X})$ と分解する．ここで，P_L は「目標言語モデル」を表し，P_E は「編集モデル」を表す．目標言語モデルは目標の音楽スタイルの特徴を記述するためのものであり，編集モデルは参照曲 X と生成曲 $\widetilde{X}$ との内容の類似度を評価するためのものである．

2.2 音楽言語モデル

ピッチクラスの系列に対する極小モデルとして，ピッチクラスマルコフモデル（Pitch-class Markov Model; PcMM）を考える．これは，初期確率 $P(q_{11} = q)$ と遷移確率 $P(q_{mn} = q \,|\, q'_{mn} = q')$ により定義される．以下では，q_{mn} の一つ前の音符を q'_{mn}（同様に s'_{mn} など）と表すことにする．同様に，拍節位置に対する極小モデルとして，初期確率 $P(b_{11} = b)$ と遷移確率 $P(b_{mn} = b \,|\, b'_{mn} = b')$ により定義される拍節マルコフモデル（Metrical Markov Model; MetMM）[32,33] を考える．

音高とリズムの相互依存性を捉えるため，ピッチクラスと拍節位置の積空間を考えて PcMM と MetMM を組み合わせる．遷移確率は $P(q_{mn}, b_{mn} \,|\, q'_{mn}, b'_{mn}) = \Psi(q'_{mn}, b'_{mn}; q_{mn}, b_{mn})$ と表される．以降では，煩雑な表記を避けるため，初期確率は明示的に書かないが，同様に定義されるものと理解されたい．このモデルの状態空間は二つの次元が音高と拍節位置に対応するトーラス $S^1 \times S^1$ 上の格子と見なせるため，このモデルを「トーラスマルコフモデル」（Torus Markov Model; TMM）と呼ぶ．TMM では拍節位置が異なるピッチクラスは区別されるため，1 次のマルコフモデルでも小節の長さ程度の依存性が捉えられることに注意されたい．

移調（曲全体の大局的な音高シフト）と転調（曲の一部の局所的な音高シフト）を記述するため，局所的な調変数

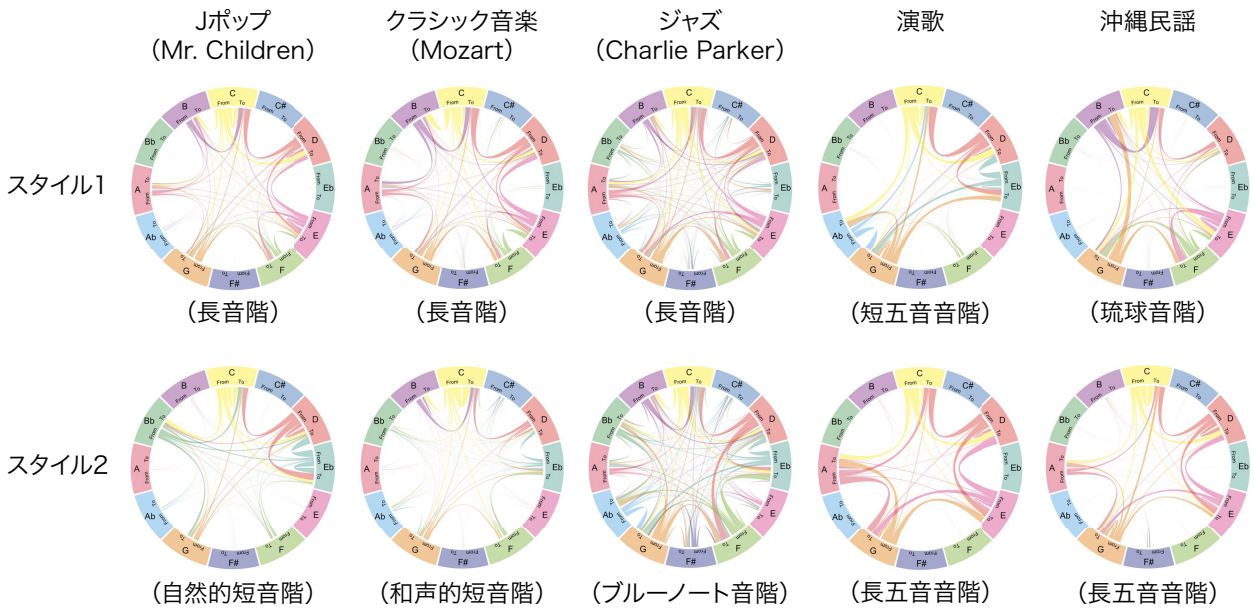

図 2 TSTMMixM の学習済みパラメータ．各データから学習された代表的な 2 つの要素モデルに対して，周辺化したピッチクラス遷移確率が帯で示されている．対応すると考えられる音階の名前が与えられている．これらの音階の主音がドになるように移調した結果が示されている．

$k_m \in \{0,\dots,11\}$ を各小節 m に対して定義する．例えば，$k_m = 0$ がハ長調を表す時，$k_m = 2$ はニ長調を表す．調変数はマルコフモデル $P(k_m|k_{m-1})$ により生成されると考えると，TMM は次のように拡張できる．

$$P(k_m = k \,|\, k_{m-1} = k') = \pi_{k'k} \tag{1}$$

$$P(q_{mn}, b_{mn} \,|\, q'_{mn}, b'_{mn}, k_m) = \Psi^{(k_m)}(q'_{mn}, b'_{mn}; q_{mn}, b_{mn}) \tag{2}$$

異なる主音を持つ調に対するパラメータを関係付けるため，任意の $\ell \in \{0,\dots,11\}$ に対して，次のようにモデルパラメータに対して移調対称性を課す．

$$\pi_{k'k} = \pi_{(k'+\ell)(k+\ell)} \tag{3}$$

$$\Psi^{(k)}(q', b'; q, b) = \Psi^{(k+\ell)}(q'+\ell, b'; q+\ell, b) \tag{4}$$

ただし，ピッチクラスおよび調変数は 12 を法として定義するものとする．このモデルを移調対称 TMM（Transposition-Symmetric TM; TSTMM）と呼ぶ．移調対称 PcMM も同様に定義できる．

1 章で議論した通り，あるカテゴリーの音楽には音高やリズムのモードがいくつかあることが多い．これらのモードが TSTMM の異なるパラメータ値で表現できると考えて，小節ごとにモード変数 $\rho_m \in \{1,\dots,N_{\mathrm{M}}\}$ を導入することで混合モデルを構成する．各モードに対して k_m で示される 12 の移調形を考える．生成過程は次のように記述される．

$$P(\rho_m = \rho, k_m = k \,|\, \rho_{m-1} = \rho', k_{m-1} = k') = \pi_{\rho'k',\rho k} \tag{5}$$

$$P(q_{mn}, b_{mn} \,|\, q'_{mn}, b'_{mn}, \rho_m = \rho, k_m = k) = \Psi^{(\rho,k)}(q'_{mn}, b'_{mn}; q_{mn}, b_{mn}) \tag{6}$$

ここでも式 (3) と (4) と同様に移調対称性を課す．このモデルは「移調対称トーラスマルコフ混合モデル」（Transposition-Symmetric Torus Markov Mixture Model; TSTMMixM）と呼ぶ．それぞれの TSTMM 成分が $P_L(\widetilde{X})$ に対応する．

TSTMMixMs は EM アルゴリズム [34] により教師なしで学習できる．原理的にはランダム初期化を使えるが，経験的にこれでは望ましくない局所解に落ちることが分かった．そこで，移調対称 PcMM の混合モデルと MetMM の混合モデルを別々にまず学習して，それらを用いて TSTMMixMs を初期化する方法を用いる．

2.3 音楽スタイルの学習結果

いくつかの音楽データから学習された TSTMMixM が図 2 に描かれている．データの内，J ポップ，クラシック音楽，演歌のデータに関しては，詳細を 3.1 節で述べる．この他，Charlie Parker の 50 曲 3,370 小節からなるジャズのデータと，様々な沖縄民謡 61 曲 2,154 小節からなるデータを用いた．全ての曲は 4/4 拍子あるいは 2/4 拍子であり，形式的に 4/4 拍子に変換して学習した．ここでは，TMM 成分を周辺化して得られた PcMM を可視化しており，ピッチクラスは対応すると考えられる音階の主音がドになるよ

うに移調してある．

Jポップ，クラシック音楽，ジャズのデータからは長音階と短音階に対応するモデルが学習された．Jポップ（具体的にはMr. Childrenの楽曲）のモデルでは短音階は，短7度を主に含む自然的短音階に近いものであり，クラシック音楽のモデルでは，長7度を主に含む和声的短音階に近いものが得られた．また，ジャズのデータではどちらのモデルも全音階の音階音以外の音を多く含むものが学習され，短音階はブルーノート音階と呼ばれるフラットを伴う5度の音を多く含んでいることが見てとれる．このように，それぞれのスタイルの特徴を捉えたモデルが人による音階ラベルや調ラベルなしに学習できることが分かった．

また，演歌と沖縄民謡のデータからは五音音階ないし六音音階に近いモデルが学習された．演歌データでは，メジャーとマイナーの五音音階が学習された．これらは通称，ヨナ抜き音階として知られるものであるが，曲によっては4度や7度を含む場合があることも分析結果から分かる．沖縄民謡では琉球音階として知られる音階と長五音音階に近いモデルが得られた．文献 [35] では琉球音階には2度の音（レ）を含む六音音階と含まない五音音階があることが分析されているが，図2の結果はこれと一致するものであり，人間の専門知識を用いない楽曲データだけからの分析による確認が得られたと言える．また同文献では，これらの音階の他に沖縄民謡では長五音音階が用いられることも分析されているが，今回得られた結果はこの点でも一致している．

以上の分析に見るように，今回のデータから学習された音階は既に音楽学でよく知られたものであり，データから獲得されることは予想できるが，調のラベルが全くないデータから教師なしでこれらが獲得できたことは強調するに値する．また，モデルは音階の構成音を学習するだけではなく，音高の遷移確率やリズムとの関係性も記述しており，音階分析よりも詳しいスタイルの分析が得られていると言える．

異なるカテゴリーの音楽データから異なるリズムのスタイルも学習できた（付随ページ [36] の可視化結果を参照）．例えば，クラシック音楽のスタイルでは音符のオンセット位置は強拍の上で最も頻繁に起こる一方で，Jポップのスタイルではそうとは限らないことが分かった．これは後者のスタイルでシンコペーションがしばしば現れることを反映した結果である．

2.4 編集モデル

参照曲メロディーの音符 $x{=}(q,b)$ が生成曲メロディーの音符 $\tilde{x}{=}(\tilde{q},\tilde{b})$ が対応する時，幾何的な距離に基づく「単純編集モデル」は次で定義される．

$$P(x|\tilde{x}) \propto \exp\left(-\frac{(q-\tilde{q})^2}{2\sigma_p^2}\right)\exp\left(-\frac{(b-\tilde{b})^2}{2\sigma_r^2}\right) \qquad (7)$$

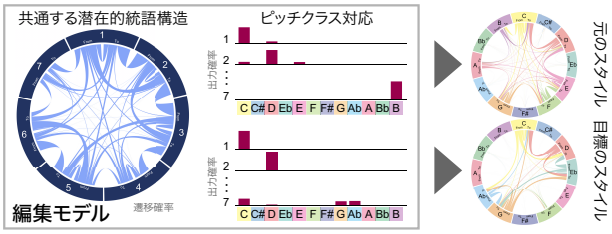

図 3　編集モデルで記述される音符の統語機能

ここで，二乗距離はピッチクラスと拍節位置の空間で定義されたものであり，σ_p と σ_r はスケールパラメータである．

この単純編集モデルは，特に教師なしの設定では本質的な問題を持つ．例えば，ハ長調のメロディーを短調のモードに変換する時，通常は生成メロディーの調としてハ短調が選ばれ，終止音として通常用いられる主音ドは保たれる．しかし，単純編集モデルでは，音符の幾何的距離を最小化するものとしてイ短調あるいは他の短調が選ばれ，主音のような音符の統語機能の構造が保たれない場合が起きてしまう．

この問題を解決するため，音符の統語構造を考慮した改良編集モデルを構築する．データから教師なしで音符の統語機能を獲得するために，文献 [37] で提案された方法を応用する．この方法では，記号の機能を表す潜在状態を保つHMMを系列的文脈，即ち前後にどういった記号が現れるかという情報を用いて学習する．潜在状態を $z_{mn} \in \{1,\ldots,N_{\mathrm{F}}\}$ と記すと（N_{F} はあらかじめ決める統語記号の数である），このHMMは遷移確率 $P(z_{mn}|z'_{mn})$ と出力確率 $P(q_{mn}|z_{mn})$ により定義される．これらの確率はTSTMMixMの確率 $P(q|q')$ を近似するように決められる．

二つのスタイルをつなぐ編集モデルを構成するため，このモデルを二つの出力確率を持つように拡張する．一つは参照曲のスタイルに対する $P(q_{mn}|z_{mn})$ であり，もう一方は目標スタイルに対する $P(\tilde{q}_{mn}|z_{mn})$ である（図3）．共通の統語構造を表すモデルを導き出すため，潜在状態とその遷移確率 $P(z_{mn}|z'_{mn})$ は二つのスタイルで共有する．これらの確率は $P(q|q')$ と $P(\tilde{q}|\tilde{q}')$ の両方を近似するように決められる．このモデルにより次の編集確率を構成できる．

$$P_F(X|\widetilde{X}) = \sum_{\boldsymbol{z}}\left[\prod_{m,n} P(q_{mn}|z_{mn})\right] P(\boldsymbol{z}|\tilde{\boldsymbol{q}}) \qquad (8)$$

ここで，$\boldsymbol{z}=(z_{mn})$ および $\tilde{\boldsymbol{q}}=(\tilde{q}_{mn})$ である．左辺の二番目の因子はフォワード・バックワードアルゴリズムにより計算できる．

幾何的距離に基づくモデルと統語機能に基づくモデルを結合して，次のように「改良編集モデル」を定義する．

$$P_E(X|\widetilde{X}) \propto P_F(X|\widetilde{X})^{\alpha_1} \prod_{m,n} P_D(q_{mn}|\tilde{q}_{mn})^{\alpha_2} P_D(b_{mn}|\tilde{b}_{mn})^{\alpha_3}$$

ここで $P_D(q|\tilde{q})$ と $P_D(b|\tilde{b})$ は式 (7) の二つの因子を表し，要素モデルの重み α_1，α_2，および α_3 を導入した．

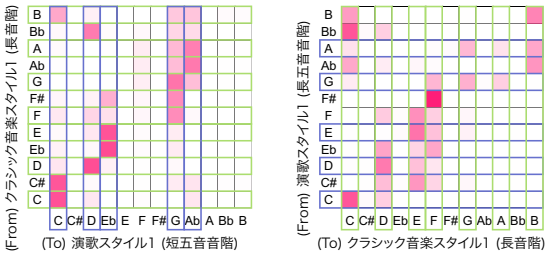

図 4 二対の音楽スタイルに対して学習された編集確率 $P(q|\tilde{q})$．太い長方形は主な音階音を示している．

編集モデルの学習されたパラメータの例を図 4 に示す．図中のヒートマップは確率 $P(q|\tilde{q}) \propto \sum_z P(q|z)P(\tilde{q}|z)\pi^*_z$ を表している（π^*_z は z の定常分布）．左側の図では，長音階からマイナー調の五音音階への音符の写像を示しており，中音と下中音（ミとラ）は主にフラット音（ミ♭とラ♭）にマップされていることが見て取れる．これは音楽的直感に合う結果と言える．右側の図では，メジャー調の五音音階から長音階への写像を示している．特に，五音音階の四番目と五番目の音（ソとラ）は全音階では複数の音に対応している．これらの音の機能は文脈により変化する．例えば，五音音階でドの前のラは全音階では導音のシに対応することがある．これらの例が示すように，教師なし学習により得られた改良編集モデルはしばしば音楽的直感に合致することが分かった．

2.5 メロディースタイル変換のアルゴリズム

メロディースタイル変換のアルゴリズムを 2.2 節の言語モデルと 2.4 節の編集モデルの組み合わせの統計的推論により導出できる．まず参照曲と生成曲に対応する二つの音楽カテゴリーのデータに対して一組の TSTMMixM を学習して，モード変数 ρ_{source} と ρ_{target} によりインデックスされる音楽スタイルを習得する．もし音楽スタイル ρ_{source} に対応する参照メロディー X が与えられると，元の言語モデルを用いればビタビアルゴリズムで調情報を推定できる．

目標の音楽スタイルは ρ_{target} の値の一つにより指定される．生成メロディー $\widetilde{X}$ とその目標言語モデルに対する調情報は確率 $P(\widetilde{X}|X) \propto P_E(X|\widetilde{X})P_L(\widetilde{X})$ の最大化により推定されるが，これもビタビアルゴリズムと同様の方法で可能である．調推定の精度は高いため（我々が用いた評価データでは 100% であった），参照メロディーの調情報が推定された後，この情報は編集モデルを用いて生成メロディーへと転移することで計算を効率化できる．

3. 結果と評価

3.1 実験設定

評価には，西洋クラシック音楽，J ポップ，演歌の三つのカテゴリーの音楽データを用いた．クラシック音楽データはモーツァルトが作曲したソプラノメロディー 7,133 小節からなり，J ポップデータは Mr. Children のメロディー 3,878 小節からなり，演歌データは [38,39] に掲載されている多様なアーティストによるメロディー 37,032 小節からなる．

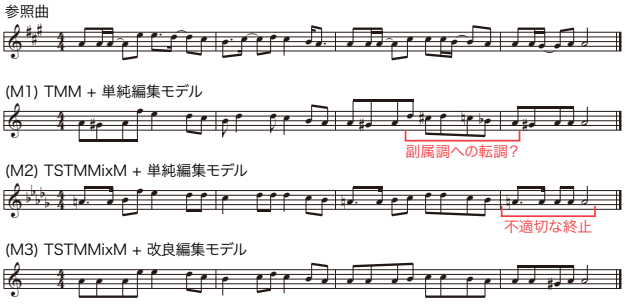

図 5 メロディーのスタイル変換の例．参照曲は J ポップのメロディーであり，手法 M2 と M3 ではクラシック音楽データから学習された二番目のスタイル（短調）を目標スタイルとして用いている．

各音楽カテゴリーの言語モデルは次のように学習した．まず，PcMM の N_{PM} 個の混合と MetMM の N_{RM} 個の混合を EM アルゴリズムで学習した．次に，これらのモデルの積の全ての組み合わせを用いて初期化した $N_{\text{M}} = N_{\text{PM}}N_{\text{RM}}$ 個の混合を持つ TSTMMixM を学習した．評価データに対する混合数は，クラシック音楽，J ポップ，演歌に対してそれぞれ $(N_{\text{PM}}, N_{\text{RM}}) = (3,1), (3,2), (3,3)$ とした．学習した TSTMM から大きな混合重みを持つ二つの成分を選び，各音楽スタイルの代表的なスタイルとして用いた．いくつかの試みの後，$\alpha_1 = 0.4$，$\alpha_2 = \alpha_3 = 0.8$，$\sigma_p = 0.7$，$\sigma_r = 3$，および $N_{\text{F}} = 7$ とした．比較のため，次の三つの手法を実装して評価した．

- (M1) TMM + 単純編集モデル
- (M2) TSTMMixM + 単純編集モデル
- (M3) TSTMMixM + 改良編集モデル

3.2 結果例

メロディースタイル変換の例を図 5 に示す．J ポップのメロディーをクラシック音楽のスタイルへ変換している．リズムの側面では，三つの手法の結果はどれも成功している．J ポップの曲に典型的であるタイされた音符が抑えられ，クラシック音楽スタイルで典型的であるリズムに置き換えられている．一方で，これらの結果は音高の側面では大きく違っている．M1 の結果では，第 3 小節の変化音が下属調への転調をほのめかしているが，これは参照曲にはない転調であり，終止を不安定にしている．これは TMM では調の構造が捕らえられていないことにより説明できる．M2 と M3 の結果では，調の構造は一貫している．しかし，M2 の結果は変ロ短調であり，終止付近の音符が適切に変換されていない．これは単純編集モデルでは主音などの音符の機能が捉えられないことにより説明できる．M3 の結果では音高組織にこのような問題は見つからない．

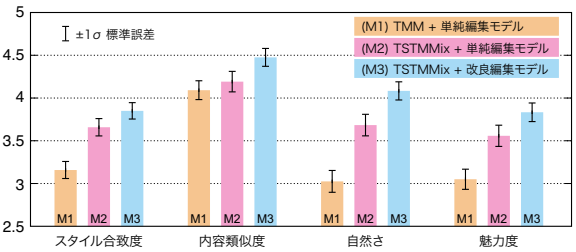

図 6　主観評価の結果．中心のバーは平均値を示し，小さいバーは 1σ の標準誤差を示している．

同様の傾向は他の結果でも見られる（付随ページ [36] 参照）．M1 の結果ではしばしば不自然な調の構造が見られ，時には調性を感じさせないこともあった．M2 と M3 の結果は，生成結果の調が同じ時にはしばしば類似していた．他の場合では，M2 の結果では，音符の機能や調構造が適切に生成メロディーに転移されない場合に，不自然な音高構造が見られた．

3.3　主観評価

スタイル変換の品質を調べるため，主観評価実験を行った．毎日平均一時間以上音楽を聴いている 10 人の評価者が実験に参加した．三つの音楽カテゴリー（クラシック，J ポップ，演歌）から 8 小節の長さを持つよく知られたメロディーを参照曲として二つずつ選び，参照曲とは異なる二つのカテゴリーのスタイルへと変換した．合計で各手法ごとに 12 のメロディーを生成した（生成結果は付随ページ [36] に記載されている）．評価者には，目標のスタイルは教えたが，手法は教えなかった．生成曲のメロディーを聴いたあと，次の各尺度について 6 段階で評価してもらった．

- **スタイル合致度** 生成メロディーが目標のスタイルと合致しているか
- **内容類似度** 生成曲が参照曲と似ているか
- **自然さ** 生成曲メロディーが自然か
- **魅力度** 生成曲メロディーが魅力的か

図 6 の結果が示す通り，手法の改良によって全ての尺度の平均値において改善が見られた．特に，M1 と M2 の比較では，言語モデルの改良によりスタイル合致度の点数が 0.5 向上し（t 検定での p 値 $< 10^{-5}$），M2 と M3 の比較では，編集モデルの改良により内容類似度の点数が 0.28 向上した（t 検定での p 値 3.3×10^{-3}）．自然さおよび魅力度の点数の改善も統計的に有意であった．これらの結果は提案法の有効性を明確に示すものである．

3.4　編集モデルの重みの効果

次に手法 M3（TSTMMixM + 改良編集モデル）について，生成結果のスタイル合致度と参照曲との類似度を客観的に定量して手法の挙動を調べる．同時に編集モデルの重みパラメータを変化させた場合の効果についても調べる．

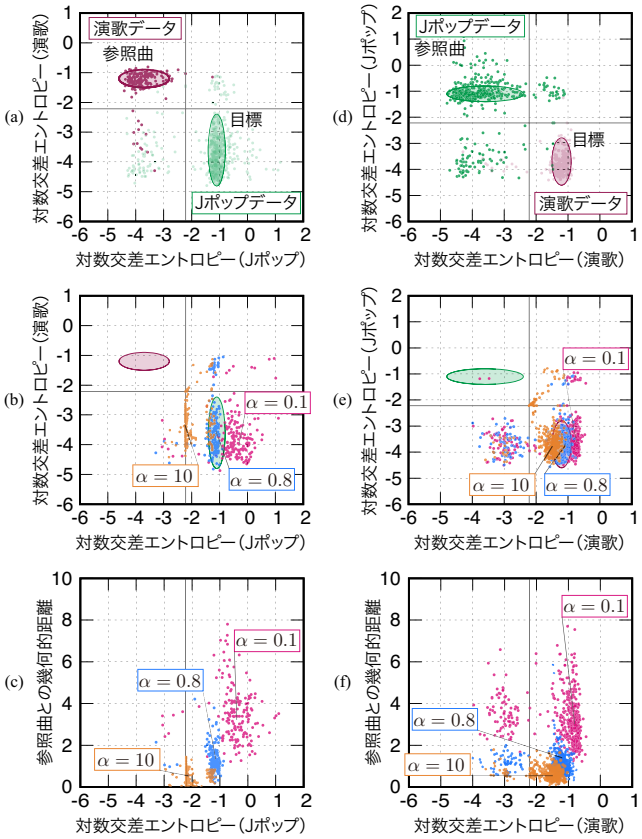

図 7　J ポップと演歌の間のスタイル変換の定量評価．(a) から (b) は J ポップの曲断片を参照曲として演歌スタイルでの生成を行った場合，(d) から (f) は演歌の曲断片を参照曲として J ポップスタイルでの生成を行った場合の結果を示している．図中の実線は，無情報モデル（一様分布）による交差エントロピーの値（およそ −2.2）を示している．

3.1 節に述べた J ポップと演歌のデータを用いて，両方のデータからそれぞれ 20 曲をランダムに選択して参照曲として，もう一方のスタイルへの変換をする．ここでは，参照曲を 4 小節ごとに切り取り，これらの曲断片を変換する．ただし，音符数が 10 未満の断片は，メロディーの特徴が捉えにくいため，除外した．モデルのパラメータ値などの設定は 3.1 節に述べたものと同じであるが，編集モデルの重みパラメータに関しては次のように設定した．変数 α に対して，$\alpha_1 = \alpha/2$ および $\alpha_2 = \alpha_3 = \alpha$ とする．α の値としては，0.1，0.8，10 の 3 通りを試した．

図 7 に結果を示す．参照曲および生成曲のスタイル合致度を測るために，J ポップモデルと演歌モデルに対する交差エントロピーを計算した．各曲断片に対して，最尤となるモード変数と調変数を求め，それを用いて交差エントロピーを計算した．また，参照曲との類似度を測るために，幾何的距離を計算した．これは 3.1 節に値を記したスケールパラメータに基づく二乗距離である．図 7(a)(d) を見ると，2 つのスタイルモデルの交差エントロピーの空間において，2 つのスタイルの参照の曲断片がクラスターを形成していることが確認できる．ただし，一部の曲断片は交差

エントロピーが両方低かったり，両方高かったりすることも確認できる．次に，図 7(b)(e) では，自動生成された曲断片が目標のスタイルに近づいている様子が見てとれる．編集モデルの重みが小さいほど，言語モデルの効果が大きくなるため，スタイル合致度がより大きくなっていることも分かる．図 7(c)(f) では，編集モデルの重みに応じて，生成結果と参照曲の幾何的距離が変化していることが分かる．つまり，この重みが小さい時はスタイル合致度は大きいが，参照曲との幾何的距離も大きいこと，逆に，重みが大きい時は参照曲との幾何的距離は小さいが，大部分の生成曲断片が目標のスタイルから外れている様子が見える．この結果から，編集重みの効果が確認できるとともに，$\alpha = 0.8$ 辺りの領域がスタイル合致度と参照曲との類似度がバランスしやすい値であることが確認できた．

一方で図 7 の結果から，特に J ポップから演歌へのスタイル変換において，編集モデルの重みが小さい時でも，一部の曲断片が演歌スタイルから大きく外れることがあることが分かった．このことは，曲によってスタイル変換しやすいものとそうでないものが存在することを示唆している．より詳細な解析および解決法については今後の課題とする．

3.5 スタイル変換による楽曲生成のデモページ

スタイル変換により生成された楽曲は，デモページ[*1]から視聴できるようになっており，今後も更新していく予定である．また，本稿では参照曲として電子形式で楽譜が用意された楽曲を考えてきたが，スタイル変換の定式化自体は，音楽音声データからの自動採譜で得られた不完全な楽譜やランダム生成された楽譜などにも適用できる．また，自動採譜では出力結果に音楽的に不適切な誤りが含まれることがしばしばあるが，これを音楽的に自然な楽譜へと修正することも考えられる．こうした可能性についても，得られた結果をデモページにて発表していく予定である．

4. 結論

最初に上げた問いに戻ると，本研究の結果は音階や典型的なリズムなどの音楽スタイルの側面が統計的生成モデルにより定義されるクラスターにより記述できること，そしてそれが音楽の専門知識に大きくは依存せず学習できることを示している．音楽スタイルの教師なし学習はジャンル分類の文脈で研究されているが [25, 29]，本研究ではそれが音楽生成にも有用であることが示された．また，スタイル変換に基づく楽曲生成において音楽類似性を記述するためには音符の統語機能を考慮する重要性を明らかにし，調情報を自動的に推定しなければいけない時にそれが特に重要であることを明らかにした．提案する枠組みの一般性の高さと得られた結果の品質の高さから，このスタイル変換の定式化はコード進行や多声音楽など他の形式の音楽に対しても有用である可能性が見込まれる．

本稿で扱った枠組みは他の音楽スタイルに容易に応用できるので，幅広い種類の音楽スタイルに対して手法の有効性を検証する予定である．同じデータであっても異なる初期値からは異なる音楽スタイルが得られ，その結果はいつでも解釈可能な訳ではないことも明らかになっている．よって，情報量尺度（例えば尤度）により音楽的に意味のあるスタイルのクラスタリングを特徴付けことは重要な課題である．また適切な混合数を決める問題も今後の重要な課題である．

謝辞 本研究は，科研費 16H01744，16H02917，16K00501，16J05486，19K20340，JST ACCEL No. JPMJAC1602，柏森情報科学研究財団，および京都大学教育研究振興財団からの支援を受けた．

[*1] https://melodyarrangement.github.io/demo-ja.html

参考文献

[1] G. Papadopoulos and G. Wiggins, “AI methods for algorithmic composition: A survey, a critical view and future prospects,” in *Proc. AISB Symposium on Musical Creativity*, 1999, vol. 124, pp. 110–117.
[2] G. Nierhaus, *Algorithmic Composition*, Springer, 2009.
[3] J. D. Fernández and F. Vico, “AI methods in algorithmic composition: A comprehensive survey,” *J. Artificial Intelligence Res.*, vol. 48, pp. 513–582, 2013.
[4] J.-P. Briot, G. Hadjeres, and F. Pachet, “Deep learning techniques for music generation—A survey,” *arXiv preprint arXiv:1709.01620*, 2017.
[5] F. Pachet, “The continuator: Musical interaction with style,” *J. New Music Res.*, vol. 32, no. 3, pp. 333–341, 2003.
[6] H. Maekawa et al., “On machine arrangement for smaller wind-orchestras based on scores for standard wind-orchestras,” in *Proc. ICMPC*, 2006, pp. 268–273.
[7] S. Fukayama et al., “Automatic song composition from the lyrics exploiting prosody of the japanese language,” in *Proc. SMC*, 2010, pp. 299–302.
[8] F. Pachet and P. Roy, “Markov constraints: Steerable generation of Markov sequences,” *Constraints*, vol. 16, no. 2, pp. 148–172, 2011.
[9] G. Hori, H. Kameoka, and S. Sagayama, “Input-output HMM applied to automatic arrangement for guitars,” *J. Info. Processing Soc. Japan*, vol. 21, no. 3, pp. 264–271, 2013.
[10] M. McVicar, S. Fukayama, and M. Goto, “AutoLeadGuitar: Automatic generation of guitar solo phrases in the tablature space,” in *Proc. ICSP*, 2014, pp. 599–604.
[11] B. L. Sturm et al., “Music transcription modelling and composition using deep learning,” in *Proc. CSMC*, 2016, pp. 1–16.
[12] G. Hadjeres and F. Pachet, “DeepBach: A steerable model for Bach chorales generation,” *arXiv preprint arXiv:1612.01010*, 2016.
[13] L. Crestel and P. Esling, “Live orchestral piano, a system for real-time orchestral music generation,” in *Proc. SMC*, 2017, pp. 434–442.
[14] E. Nakamura and K. Yoshii, “Statistical piano reduction

controlling performance difficulty," *APSIPA Trans. on Signal and Information Processing*, 2018, to appear.
[15] H.-W. Dong et al., "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment," in *Proc. AAAI*, 2018.
[16] L.-C. Yang, S.-Y. Chou, and Y.-H. Yang, "MidiNet: A convolutional generative adversarial network for symbolic-domain music generation," *arXiv preprint arXiv:1703.10847*, 2017.
[17] H. H. Mao, T. Shin, and G. Cottrell, "DeepJ: Style-specific music generation," in *Proc. IEEE ICSC*, 2018, pp. 377–382.
[18] D. Tzimeas and E. Mangina, "Jazz Sebastian Bach: A GA system for music style modification," in *Proc. IEEE ICSNC*, 2006, pp. 36–42.
[19] F. Zalkow, S. Brand, and B. Graf, "Musical style modification as an optimization problem," in *Proc. ICMC*, 2016, pp. 206–2011.
[20] W-T. Lu and L. Su, "Transferring the style of homophonic music using recurrent neural networks and autoregressive models," in *Proc. ISMIR*, 2018, pp. 740–746.
[21] G. Brunner et al., "Symbolic music genre transfer with CycleGAN," in *Proc. IEEE ICTAI*, 2018, pp. 786–793.
[22] N. Mor et al., "A universal music translation network," *arXiv preprint arXiv:1805.07848*, 2018.
[23] P. F. Brown et al., "The mathematics of statistical machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
[24] R. L. Crocker, *A History of Musical Style*, McGraw-Hill, 1966.
[25] J.-J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *J. New Music Res.*, vol. 32, no. 1, pp. 83–93, 2003.
[26] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.
[27] B. L. Sturm, "Classification accuracy is not enough," *J. Intelligent Information Systems*, vol. 41, no. 3, pp. 371–406, 2013.
[28] J. Sakellariou et al., "Maximum entropy models capture melodic styles," *Sci. Rep.*, vol. 7, no. 9172, pp. 1–9, 2017.
[29] X. Shao, C. Xu, and M. S. Kankanhalli, "Unsupervised classification of music genre using hidden Markov model," in *Proc. ICME*, 2004, vol. 4, pp. 2023–2026.
[30] C. L. Krumhansl, "The psychological representation of musical pitch in a tonal context," *Cognitive Psychology*, vol. 11, no. 3, pp. 346–374, 1979.
[31] P. Hanna, P. Ferraro, and M. Robine, "On optimizing the editing algorithms for evaluating similarity between monophonic musical sequences," *J. New Music Res.*, vol. 36, no. 4, pp. 267–279, 2007.
[32] C. Raphael, "A hybrid graphical model for rhythmic parsing," *Artificial Intelligence*, vol. 137, pp. 217–238, 2002.
[33] M. Hamanaka et al., "A learning-based quantization: Unsupervised estimation of the model parameters," in *Proc. ICMC*, 2003, pp. 369–372.
[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
[35] 金井喜久子, 琉球の民謡, 音楽之友社, 1954.
[36] Supplemental material (available online), `http://melodyarrangement.github.io/demo.html`.
[37] H. Tsushima et al., "Generative statistical models with self-emergent grammar of chord sequences," *J. New Music Res.*, vol. 47, no. 3, pp. 226–248, 2018.
[38] Y. Goto (ed.), *Grand Collection of Enka Songs by Male Singers 5th Ed. (in Japanese)*, Zen-on Music Co., 2016.
[39] Y. Goto (ed.), *Grand Collection of Enka Songs by Female Singers 5th Ed. (in Japanese)*, Zen-on Music Co., 2016.