Vol.2025-MUS-142 No.33
2025/3/7

歌声 MIDI 採譜における作曲スタイルを導入した言語モデルの効果の検証

杉本 悠[1]　中村 栄太[1]

概要：歌声採譜で課題となる大きな音高変動や曖昧な音符境界に対処するため，音楽言語モデルを制約として用いる方法が有効である．これにより音階から外れた音符誤りを低減できるが，音階などの作曲スタイルは多様であり，全ての楽曲に単一の言語モデルを用いる従来の方法には限界がある．本研究では，ポピュラー音楽の楽譜データをクラスタリングして，作曲スタイルごとにニューラル言語モデルを学習する．これを利用した RNN トランスデューサモデルで採譜を行い，言語モデルの効果を検証する．

1. はじめに

歌声 MIDI 採譜は，歌唱支援や音楽分析の基盤技術であり，近年は深層学習が用いられている [1–3]．歌声採譜では，音高変動や不明瞭な音符境界の影響で精度向上が課題であり，音楽言語モデルの統合による採譜精度の改善が研究されている [4]．従来で用いられたマルコフ言語モデルでは，音符間の関係を精緻に表せないという限界が存在する．また，多様な作曲スタイルを単一のモデルで表すことにも，作曲規則上の制約を精緻に表せないという限界がある．本研究では，ニューラル言語モデルと作曲スタイル変数を導入し，RNN トランスデューサ [5] を用いた統合を行う（図 1）．曲ごとの音高分布をクラスタリングしてスタイル変数を定め，採譜時に音響特徴から認識し，言語モデルへ入力する．本デモ発表では，ニューラル言語モデルと作曲スタイル変数を導入の効果を実際の採譜例を示しながら議論する．

2. 提案手法

2.1 問題設定

原曲 ($c = 1$) と歌声分離 [6] で得られた歌声のみ ($c = 2$) の音響信号の各々から抽出したメルスペクトログラムを $[x_{tcf}]_{t=1,f=1}^{T,F}$ とする（T はフレーム数，F は周波数ビンの数）．これらのスペクトログラム $\boldsymbol{X} = [\boldsymbol{x}_t]_{t=1}^{T} \in \mathbb{R}^{T\times 2\times F}$ ($\boldsymbol{x}_t = [x_{tcf}]_{c=1,f=1}^{2,F}$) が入力である．出力は，歌唱パートの音符列 $(t_l^{\mathrm{on}}, t_l^{\mathrm{off}}, p_l)_{l=1}^{L}$ である．t_l^{on} は l 番目の音符の発音時刻，t_l^{off} は消音時刻，$p_l \in \{0,\dots,127\}$ は音高（MIDI ノート番号），L は音符数を表す．本稿では，フレーム単位は 10 ms として，発音・消音時刻は 10 ms 単位で表す．

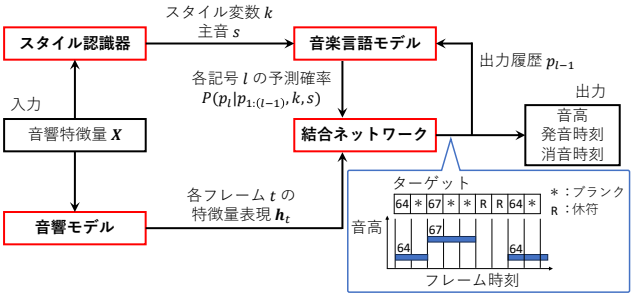

図 1: 提案手法の全体構成

2.2 RNN トランスデューサに基づく歌声採譜

提案手法は，音声認識などで用いられる RNN トランスデューサ [5] に基づいている．この手法は，音響モデルと言語モデル，結合ネットワークからなる（図 1）．音響モデルは $\boldsymbol{x}_{1:T}$ から各時刻 t における特徴量表現 $\boldsymbol{h}_t \in \mathbb{R}^H$ を出力する（H は出力次元）．具体的には，各時刻 t における（休符を含む）音高 $p_t \in \{0,\dots,128\}$ の予測確率 $\boldsymbol{a}_t^{\mathrm{pitch}} = [a_{tp}^{\mathrm{pitch}}]_{p=0}^{128}$ と発音ラベル $b_t \in \{0,1\}$ の予測確率 $\boldsymbol{a}_t^{\mathrm{onset}} = [a_{tb}^{\mathrm{onset}}]_{b=0}^{1}$ を出力するように学習する．消音時刻の認識に必要となる休符の認識を行うため，ここでは，音高の集合は通常の音高の集合 $\{0,\dots,127\}$ に $p = 128$ で表す休符を含めたものとしている．発音ラベルは通常の音符の発音時刻において $b_t = 1$ で，それ以外では $b_t = 0$ と定義する．最終的な出力 $\boldsymbol{h}_t = \boldsymbol{a}_t^{\mathrm{pitch}} \oplus \boldsymbol{a}_t^{\mathrm{onset}}$ の次元は $H = 129 + 2$ である．音響モデルには，畳み込みニューラルネットワーク (CNN) と LSTM ユニットを持つ双方向の再帰的ニューラルネットワーク (RNN) を組み合わせた CRNN [4] を用いて，クロス

[1] 九州大学
Kyushu University

エントロピー (CE) 損失で学習する．

言語モデルは時刻 t 以前に出力された音高列 $p_{1:l(t-1)}$（$l(t)$ は時刻 t における出力記号の数）から次の音高の予測確率 $\boldsymbol{g}_t \in \mathbb{R}^{128}$ を出力する．言語モデルにおける休符の役割は小さいと考えられるため，ここでの音高の集合は，休符を含めない通常の音高の集合 $\{0,\ldots,127\}$ とする．言語モデルについては，2.3 節で詳しく説明する．

結合ネットワークは，音響モデルの出力 $\boldsymbol{h}_t$ と言語モデルの出力 $\boldsymbol{g}_t$ を入力として，時刻 t における出力記号の予測確率 $\boldsymbol{a}_t^{\text{joint}} = [a_{tp'}^{\text{joint}}]_{p'=0}^{129}$ を出力する．ここでの出力記号の集合は，休符を含む音高の集合 $\{0,\ldots,128\}$ に音符の継続を表すブランク記号（129 で表す）を加えた集合である．結合ネットワークは，正解の音符列をフレームレベルの系列に変換したものをターゲットとして CE 損失により学習する（図 1）．推論時には，$\hat{p}'_t = \text{argmax}_{p'}\{a_{tp'}^{\text{joint}}\}$ の値から音符列を推定する．$\hat{p}'_t$ がブランクの場合は直前の出力記号の継続と見なして，出力系列や言語モデルは更新しない．それ以外の場合は，出力系列に $p_{l(t)} = \hat{p}'_t$ を追加する．さらに $\hat{p}'_t$ が通常の音高の時，言語モデルも更新する．このように，結合ネットワークはフレームレベルの音響特徴量と記号レベルの出力系列をつなぐデコーダの役割を持つ．結合ネットワークには，順方向 LSTM ネットワークを用いる．

実際には，フレーム単位を 10 ms とすると，ブランクや休符のフレームが増加し，学習が困難になる．そこで，結合ネットワークはフレーム単位を 100 ms に変更した上で学習と推論を行う．一方で，最終的に 10 ms 単位の分解能を持つ音符列を得るために，結合ネットワークの 100 ms 単位の出力を制約として，音響モデルの 10 ms 単位の出力を用いて発音時刻と消音時刻の微修正を行う．

2.3 音楽言語モデル

言語モデルとして，通常の音高の系列 $(p_l)_{l=1}^{L}$ の自己回帰型生成モデルを用いる．言語モデルは音符レベルで駆動し，各ステップ l で p_{l-1} を入力として，次の記号 p_l の予測確率 $\boldsymbol{a}_l^{\text{lang}} = [a_{lp}^{\text{lang}}]_{p=0}^{127}$ を出力する．

$$a_{lp}^{\text{lang}} = P(p_l = p \,|\, p_{1:(l-1)}) \tag{1}$$

言語モデルの出力 $\boldsymbol{g}_t = \boldsymbol{a}_{l(t-1)+1}^{\text{lang}}$ の次元は 128 である．

様々な音域や調を持つ歌唱曲の楽譜データから効率的に言語モデルを学習するため，音高の平行移動（移調）に関して対称性を持つモデルを考える．まず，音高列 p_l を主音 $s \in \{0,\ldots,11\}$（ハ長調およびイ短調で C に相当する音高クラス）と相対音高列 $\tilde{p}_l = p_l - s$ に分離する．これをオクターブ移調について不変な音高表現である拡張音高クラス $q_l \in \{0,\ldots,35\}$ を用いて表す [7]．拡張音高クラス系列の生成モデル $P(q_l|q_{1:(l-1)})$ と主音 s が得られれば，上記の要件を満たす音高列の言語モデルとして使用できる．言語モデルには，順方向 LSTM ネットワークを用いて，CE 損失により学習する．採譜時に主音を推定する方法は 2.4 節で述べる．

表 1: 歌声採譜の F 値 (%) の比較

手法	COn	COnP	COnPOff
CRNN	80.36	73.22	54.34
Transducer (クラスタ数 1)	83.81	75.93	52.43
Transducer (クラスタ数 5)	84.15	76.24	52.93
Transducer (クラスタ数 15)	83.11	75.61	51.49
VOCANO [1]	68.52	51.22	36.63
CTC&CE Loss [2]	88.44	81.18	64.37

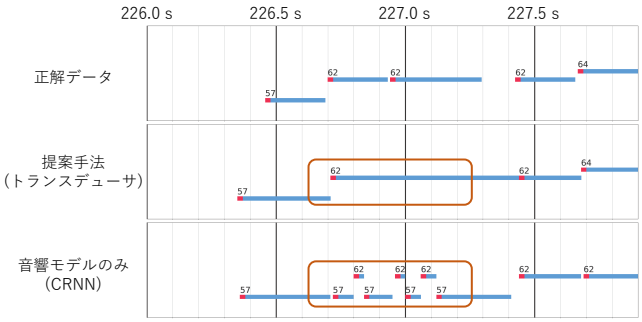

図 2: 採譜結果の例（ゆず「REASON」）

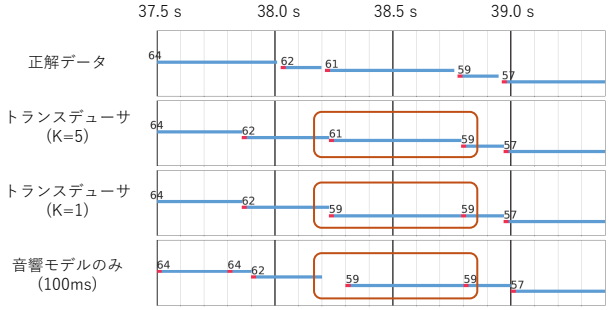

図 3: 採譜結果の例（コブクロ「灯ル祈リ」）

2.4 作曲スタイル変数の導入とスタイル認識

曲ごとの相対音高クラスの分布に対して，混合離散分布モデルと EM アルゴリズムに基づくクラスタリングによりスタイル変数を推定する．クラスタ数 K はスタイル分類の細かさを決めるハイパーパラメータとして扱う．スタイル変数 $k \in \{1,\ldots,K\}$ を入力に追加して 2.3 節の LSTM を学習することで，スタイル変数を導入した言語モデル $P(q_l|q_{1:(l-1)},k)$ を得る．

採譜時には，スタイル認識器により入力楽曲の各フレーム t のスタイル変数 k_t と主音 s_t を音響特徴量 $\boldsymbol{X}$ から推定して，言語モデルへと入力する．スタイル認識器には，音響モデルと同様の CRNN を用いる．

3. 評価実験

日本のポピュラー音楽のデータを用いて比較評価を行った．収集した全 555 曲のうち，331 曲を学習用，112 曲を検証用，112 曲を評価用とした．言語モデルの学習には，これら 555 曲を含まない 5103 曲の歌唱パートの楽譜データを用いた．

提案法における言語モデルの効果を調べるため，音響モデルのみ（CRNN）との比較，およびクラスタ数を変化させた時の比較を行った．また，最新の既存手法である VOCANO [1] と CTC&CE Loss [2] との比較も行った．次の 3 つの F 値を評価尺度とした：(i) 発音時刻 (COn)，(ii) 発音時刻と音高 (COnP)，(iii) 発音時刻と音高，消音時刻 (COnPOff)．発音時刻および消音時刻のずれの許容範囲は，デフォルト値 (発音時刻は 50 ms 以内) とした．

表 1 の結果において，トランスデューサを利用した提案手法は，COn と COnP で音響モデルのみ（CRNN）の手法を大きく上回っていることから，言語モデルを導入した効果が確認できる．また，クラスタ数を変化させたトランスデューサの比較では，クラスタ数 1（作曲スタイル変数を導入しない場合）に比べクラスタ数 5 では，精度が向上しており，作曲スタイル変数の導入の効果が確認できる．一方，クラスタ数 15 では精度が低下しており，最適なクラスタ数の存在を示唆している．トランスデューサの結果では，音符の過剰な平滑化が確認されており，さらなる改善の余地が見出された．

図 2 の例では，コーラスパートの影響により，音響モデルのみでは誤推定となった箇所（茶色の枠の部分）が提案法では修正されており，言語モデルによる効果が表れたと考えられる．図 3 の例では，作曲スタイルのクラスタ数が 1 の結果では，誤推定となった箇所（茶色の枠の部分）がクラスタ数が 5 の結果では修正されており，作曲スタイルの変数の導入の効果が表れたと考えられる．

4. おわりに

評価実験ではニューラル言語モデルの導入と作曲スタイル変数の導入の有効性が示された．より効果的なクラスタリング手法の検討，要素モデルの改良，異なる時間スケール（10 ms と 100 ms）を統一的に処理する方法の研究などが今後の課題である．ニューラル言語モデルを利用する提案法は，音高に加えてリズムの認識も含む採譜や楽器演奏の採譜にも適用し得る．多要素や多パートを含む音楽の言語モデルの統合など，より広範囲の採譜問題の研究も今後の方向性として考えている．

謝辞 Li Su 氏と Jun-You Wang 氏との議論に感謝する．本研究は，JST FOREST No. JPMJPR226X 及び科研費 No. 23K24917 の支援を受けた．

参考文献

[1] J.-Y. Hsu et al.: “VOCANO: A note transcription framework for singing voice in polyphonic music,” *Proc. ISMIR*, pp. 293–300, 2021.
[2] J.-Y. Wang et al.: “Training a singing transcription model using connectionist temporal classification loss and cross-entropy loss,” *Proc. IEEE/ACM TASLP*, pp. 383–396, 2022.
[3] J.-C. Wang et al.: “Mel-RoFormer for vocal separation and vocal melody transcription,” *Proc. ISMIR*, pp. 3–12, 2024.
[4] T. Deng et al.: “End-to-end singing transcription based on CTC and HSMM decoding with a refined score representation,” *APSIPA TSIP*, Vol. 13, No. 5, e404, 2024.
[5] A. Graves: “Sequence transduction with recurrent neural networks,” arXiv:1211.3711, 2012.
[6] Demucs (v4): https://github.com/facebookresearch/demucs “Demucs: Deep extractor for music sources with extra unlabeled data remixed,”
[7] R. Singh et al.: “Dynamic cluster structure and predictive modelling of music creation style distributions,” *RSOS*, Vol. 9, 220516, 2022.