5S-06

音楽言語モデルと採譜誤りモデルに基づく歌声採譜結果の訂正

平松 祐紀[1]　柴田 剛[2]　錦見 亮[2]　中村 栄太[2]　吉井 和佳[2]

[1]京都大学 工学部情報学科　[2]京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

歌声採譜は，音楽音響信号から歌声の音符列（単旋律を仮定）を推定する問題であり，楽曲検索や演奏補助に有用である．歌声のスペクトルや音高軌跡のダイナミクスは非常に複雑であるため，歌声採譜では，音高の誤り，発音時刻の誤り，音符数の誤りなどは避けられない．誤りを削減するため，音楽言語モデルの利用が提案されているが，採譜結果には依然として音楽的に不自然な部分が含まれる [1]．この理由として，言語モデルの表現力が十分でないことや，採譜システムにおける言語モデルの効果が限定的であることが原因と考えられる．

本研究では，何らかの手法で推定された歌声の音符列を，音楽的に妥当な音符列に訂正する統計的手法を提案する．具体的には，歌声の音符列の生成過程を表現する言語モデルと，そこに挿入・削除・置換誤りが付与されて誤りを含む音符列が生成される誤り付与モデルを統合した確率的生成モデルを定式化する．誤りを含む音符列が与えられた際には，この逆問題を解くことで，音楽的に正しい音符列を推定する．

関連研究として，メロディーのスタイル変換や楽曲のハミング検索が挙げられる．メロディーのスタイル変換は，あるスタイルの音符列を別のスタイルの音符列に変換する問題である．変換先のスタイルに対応する音符列の音楽言語モデルと，変換先からもとのスタイルに音符列を変換する編集モデルとを統合した確率モデルに基づく手法が提案されている [2]．我々の提案手法は，音楽言語モデルと編集モデルを使う点で共通であるが，変換前と変換後の音符数が異なることを許すように編集モデルを拡張する点で複雑になっている．ハミング検索では，ユーザーのハミングを採譜し，データベースから類似度が最も高い楽曲を選ぶ手法が多く，採譜結果とデータベース内の楽曲の比較に HMM を使った手法が提案されている [3]．一方，我々の研究では，採譜結果に対して正解となる楽譜の候補は与えられず，訂正後の音符列を生成するところが楽曲検索と異なる．

2. 提案法

本章では，音楽言語モデルと誤り付与モデルに基づく歌声採譜結果の訂正手法について述べる．

2.1 問題設定

いま，誤りを含む音符列 $\mathbf{X} = (x_1, \ldots, x_N)$ が与えられたときに，訂正結果として，音符列 $\mathbf{Z} = (z_1, \ldots, z_M)$ を推定したい．ここで，N は訂正前の音符数，M は訂正後の音符数である．各音符は発音時刻 $o \in \{0, \ldots, 47\}$ と

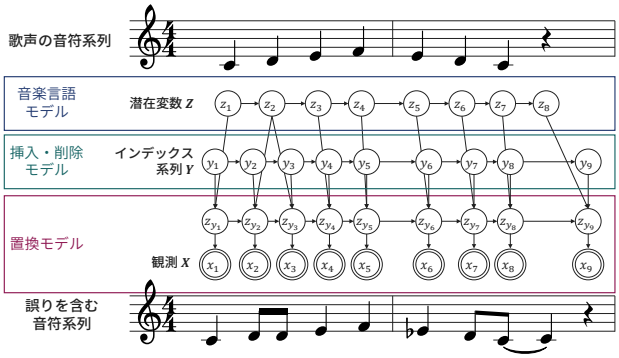

図 1: 誤りを含む音符列の生成モデル

音高 $p \in \{-1, 0, \ldots, 35\}$ のペアで表され，発音時刻は小節内の相対位置で表す．音高の -1 は休符に，0 から 35 は 12 で割った余りがピッチクラス $\{C, C\#, \ldots, B\}$ に対応する．1 オクターブ以内の音高の変化に対応するため，ピッチクラス数の 3 倍の状態を使う．本稿では，4/4 拍子の楽曲のみを扱う．

2.2 確率モデルの定式化

音符列の音楽的な妥当性を評価する音楽言語モデルと，音符列に対して挿入・削除誤りが付与される挿入・削除モデル，置換誤りが付与される置換モデルを統合した確率モデルを定式化する（図 1）．

音楽言語モデルは，調で条件付けしたマルコフモデルである．調は主音を表す $k \in \{0, \ldots, 11\}$ と，長調か短調かを表す $\rho \in \{$ 長調, 短調 $\}$ のペアで表される．音符列の調 k, ρ の生成確率を $P(k, \rho) = \pi_{k,\rho}$ と記す．調 k, ρ が与えられたとき，正しい音符列 $\mathbf{Z} = (z_1, \ldots, z_M)$ は以下の確率に従って生成される．

$$P(z_1|k, \rho) = P(o, p|k, \rho) = \xi^{k,\rho}_{o,p} \tag{1}$$

$$P(z_{m+1}|z_m, k, \rho) = P(o', p'|o, p, k, \rho) = \phi^{k,\rho}_{o,p,o',p'} \tag{2}$$

移調対称性により，次の関係を仮定する．

$$\xi^{k,\rho}_{o,p} = \xi^{0,\rho}_{o,p-k} \tag{3}$$

$$\phi^{k,\rho}_{o,p,o',p'} = \phi^{0,\rho}_{o,p-k,o',p'-k} \tag{4}$$

次に挿入・削除モデルについて述べる．正しい音符列 $\mathbf{Z} = (z_1, \ldots, z_M)$ に対して，挿入・削除を行って得られる音符列 $(z_{y_1}, \ldots, z_{y_N})$ は，インデックス列 $\mathbf{Y} = (y_1, \ldots, y_N) \in \{0, \ldots, M-1\}^N$ を定めることで決まる．$\mathbf{Y}$ は次の left-to-right 型のマルコフモデルで記述する．

$$P(y_1 = i) = \eta_i \tag{5}$$

$$P(y_{n+1} = j|y_n = i) = \psi_{i,j} \tag{6}$$

$$i > j \Rightarrow \psi_{i,j} = 0 \tag{7}$$

置換モデルは，挿入・削除された音符列の各音符を独立に置換する．挿入・削除された音符列の n 番目の音符 $z_{y_n} = (o, p)$ が与えられたとき，置換後の音符 $x_n =$

Error Correction of Singing Transcription Results Based on a Music Language Model and a Transcription Error Model: *Yuki Hiramatsu, Go Shibata, Ryo Nishikimi, Eita Nakamura, and Kazuyoshi Yoshii (Kyoto Univ.)*

表 1: 正解楽譜に対する誤り率 (%)

楽曲番号	訂正前	訂正後
2	34.0	34.2
8	26.5	25.5
17	28.9	29.8
51	19.4	19.4
100	32.9	32.7
平均	28.3	28.3

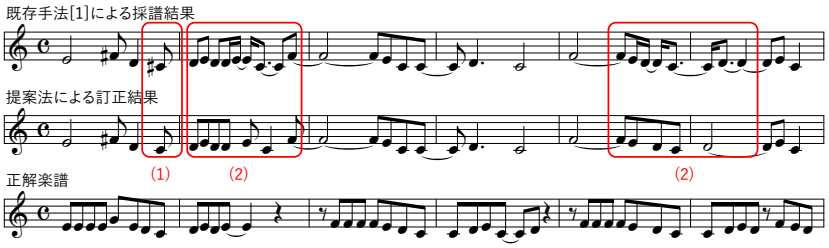

図 2: 提案法による (1) 音高誤りの修正と (2) 不自然なリズムの修正.

(o', p') は以下の確率に従って生成される.

$$P\left(o', p' | o, p\right) \propto \exp\left(-\frac{\|o' - o\|^2}{2\sigma_o^2}\right) \exp\left(-\frac{\|p' - p\|^2}{2\sigma_p^2}\right) \tag{8}$$

ここで発音時刻と音高は独立に置換され，σ_o^2 と σ_p^2 はそれぞれの分散パラメータである.

2.3 パラメータの学習・設定

音楽言語モデルのパラメータは，既存の楽譜から歌声の音符列 $\mathbf{Z}$ を大量に集め，EM アルゴリズムに基づく教師なし学習を行うことで推定できる．このとき，パラメータの初期値は，音階を構成する音高の初期確率とそれらの音高への遷移確率を高く設定しておく．一方，挿入・削除モデルおよび置換モデルのパラメータも，理論的には，正しい音符列 $\mathbf{Z}$ と誤りを含む音符列 $\mathbf{X}$ のペアが大量にあれば学習可能である．しかし，含まれる誤りの傾向は歌声推定手法に大きく依存することから，本稿では手動で与えるものとした．このとき，$\mathbf{Z}$ の原型をとどめないほど編集を受けた $\mathbf{X}$ が生成されない，($\mathbf{X}$ と近い $\mathbf{Z}$ が推定される）ような値となるように留意した.

2.4 確率モデルを用いた音符列の訂正

誤りを含む音符列 $\mathbf{X}$ が与えられたとき，訂正結果 $\mathbf{Z}$ を得る方法について述べる．まず，訂正結果 $\mathbf{Z}$ の音符数 M を固定した場合について考える．全ての調 k, ρ に対して，$P(\mathbf{Z}, \mathbf{Y}, \mathbf{X} | k, \rho)$ を最大にする $\mathbf{Z}$ と $\mathbf{Y}$ を求める．具体的には，n を 1 から N まで順に，全ての z_{y_n} と y_n に対して，$P(z_1, \ldots, z_{y_n}, y_1, \ldots, y_n, x_1, \ldots, x_n | k, \rho)$ を最大にする $(z_1, \ldots, z_{y_n - 1}, y_1, \ldots, y_{n-1})$ を求めればよい．$P(k, \rho, \mathbf{Z}, \mathbf{Y}, \mathbf{X})$ が最大となる調 k, ρ を選択することで，訂正結果となる $\mathbf{Z}$ を得る．次に，$\mathbf{Z}$ の最適な音符数を求めるには，音符数に関する事前分布 $P(M)$ を用いる．確率 $P(M)$ と，各 M に対して求めた $P(k, \rho, \mathbf{Z}, \mathbf{Y}, \mathbf{X})$ の積を最大化することで，音符数の最適値が求まる.

3. 評価実験

3.1 実験条件

実験には，RWC ポピュラー音楽データベース [4] に含まれる 4/4 拍子かつ曲中で調が変化しない 5 曲 (No. 2, 8, 17, 51, 100) を使用した．これらの音響信号に階層隠れセミマルコフモデルに基づく採譜手法 [1]（同音連打は認識できないことに注意）を適用し，誤りを含む音符列 $\mathbf{X}$ を得た．言語モデルの学習には，ビートルズの楽曲と文献 [2] で使用された演歌データを用いた．挿入・削除モデル，置換モデルのパラメータは，$\eta_0 = 0.8$，$\eta_1 = 0.2$，$\psi_{i,i} = 0.1$，$\psi_{i,i+1} = 0.8$，$\psi_{i,i+2} = 0.1$，$\sigma_o^2 = 0.4$，$\sigma_p^2 = 0.2$ とした.

採譜結果の訂正は，楽曲の調が既知であるとして行った．また訂正前と訂正後の音符数は同じである ($N = M$) と仮定した．得られた音符列の評価は，正解の音符系列と比較することで行った．評価尺度として，文献 [5] で提案されているピッチ誤り率，挿入誤り率，削除誤り率，オンセット誤り率，オフセット誤り率の平均を用いた.

3.2 実験結果

表 1 に採譜結果と訂正結果の評価値を示す．訂正を行うことで正解に近づく場合と遠ざかる場合があることが分かる．図 2 に No. 100 の訂正結果の一部を示す．(1) では，音階に含まれない音高が正しく訂正されている．(2) では正解とは異なるが，不自然なリズムが訂正されている．以上の結果から，提案手法を用いることで，誤りを含む採譜結果のみから正解楽譜を推定することができるとは限らないが，音楽的に自然な音符列に訂正できることは確かめられた.

4. おわりに

本稿では，音楽言語モデルと誤り付与モデルに基づく，誤りを含む歌声採譜結果の訂正手法を提案した．実験の結果，提案手法は採譜精度の改善自体には必ずしも寄与しないが，音楽的に不自然な誤りを訂正する能力を持つことが確認できた．真の音符列をある程度再現したうえで，音楽的に妥当な音符列を目指すことは，自動採譜技術を実用化するうえで重要な方向性であると我々は考えている．今後は，マルコフモデルより強力な深層言語モデルを用いたり，訂正後の音符列の長さを効率的に最適化するアルゴリズムを導出する予定である.

謝辞 本研究の一部は，JST ACCEL No.JPMJAC1602 および科研費 No. 19H04137, No. 19K20340, No. 16H01744 の支援を受けた.

参考文献

[1] R. Nishikimi *et al.*: “Scale- and Rhythm-Aware Musical Note Estimation for Vocal F0 Trajectories Based on a Semi-Tatum-Synchronous Hierarchical Hidden Semi-Markov Model,” *ISMIR*, 376–382, 2017.

[2] E. Nakamura *et al.*: “Unsupervised Melody Style Conversion,” *ICASSP*, 196–200, 2019.

[3] J. Shifrin *et al.*: “HMM-based musical query retrieval,” *ACM*, 295–300, 2002.

[4] M. Goto *et al.*: “RWC Music Database: Popular, Classical and Jazz Music Databases,” *ISMIR*, 287–288, 2002.

[5] E. Nakamura *et al.*: “Towards Complete Polyphonic Music Transcription: Integrating Multi-Pitch Detection and Rhythm Quantization,” *ICASSP*, 101–105, 2018.