Vol.2018-MUS-121 No.16
2018/11/21

階乗隠れセミマルコフモデルに基づく音楽音響信号に対するカバー譜生成

柴田 健太郎[1,a)]　錦見 亮[1,b)]　中村 栄太[1,c)]　深山 覚[2,d)]　後藤 真孝[2,e)]　糸山 克寿[3,f)]
吉井 和佳[1,4,g)]

概要：本稿ではポピュラー音楽の音響信号から原曲を再現するカバー譜を生成する手法について述べる．これは自動採譜の派生タスクであるが，ボーカル・ギター・ドラムスで構成されるバンドで演奏可能な楽譜（カバー譜）を推定するという点において異なる．歌声・ドラムスの分離及び採譜について既に様々な研究が行われており，高精度な楽譜の推定が達成されつつある．そこで本研究では，様々な楽器によって演奏されるその他の伴奏パートを，近似的に再構成するようなギターパート譜（リードギター，ベースギター，リズムギター）の推定を扱う．そのための手法として，３本の潜在チェインを持つ階乗型の隠れセミマルコフモデルに基づくベイジアンアプローチを提案する．一本のチェインはリズムギターのコード系列を表し，残りの２本のチェインはそれぞれリードギターとベースギターによって演奏される単旋律を表現する．入力の音響信号スペクトログラムはこれらのチェインによって生成される低ランクのスペクトログラムの和で近似される．推定対象の音響信号が与えられると潜在変数列及びパラメータはギブスサンプリングで統合的に推定される．RWC 研究用音楽データベースを用いた評価により，譜面採譜の有効性を示す．

1. はじめに

ポピュラー音楽のカバーのための採譜は，高度な音楽知識や経験を要する．ポピュラー音楽では通常，複数の楽器によって演奏され，同時に鳴る音を聞き分けるのは容易ではない．複数の音を聞き分けることができるミュージシャンにとっても，カバー譜の書き起こしにはまた別の技術が要求される．すなわち，原曲を限定的な種類の楽器で再構成する能力である．というのもカバー演奏は通常，ギターやキーボードといった典型的な少人数のバンド編成で行われるからである．計算機による歌声とドラムパートの分離や採譜は近年高精度で達成されつつある [7,12,14,20,24] のに対し，ポピュラー音楽の伴奏パートの採譜はまだあまり研究されていない．

本稿では，ポピュラー音楽の音楽音響信号から原曲を近似的に再現するカバー譜を生成する新しい問題を議論する．複数楽器の多重音からの採譜は，多重音高推定と楽器推定を行う必要があるため，難易度の高いタスクである．従来研究では各楽器の音色の特徴に着目することで複数の楽器の音源分離を行っている [2,10,23]. しかしながら，ポピュラー音楽においては同じ楽器（２本のギターなど）が同時に演奏されることが多いため，音色の特徴に基づくこれらの手法には限界がある．カバー譜の自動生成についても様々な研究が行われている [1,6] が，これらの手法はルールベースである，もしくは音響信号は考慮しない記号処理に基づくものである．このように現状，音響信号からの自動採譜とカバー譜の生成はそれぞれ別々に研究されているが，これらを同時に行うことでより高精度な結果が得られると考えられる．

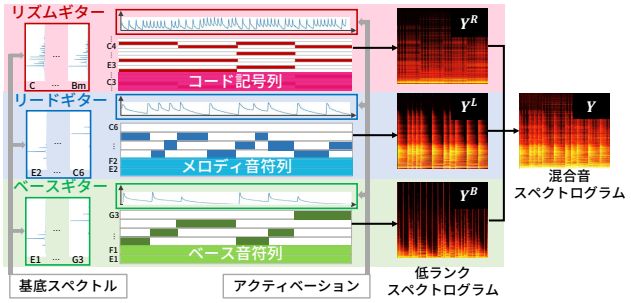

図 1　多重音響信号の伴奏パートの生成過程を表現する階乗型の隠れセミマルコフモデル

1　京都大学大学院情報学研究科
2　産業技術総合研究所 (AIST)
3　東京工業大学工学院
4　理化学研究所 革新知能統合研究センター (理研 AIP)
a)　shibata@sap.ist.i.kyoto-u.ac.jp
b)　nishikimi@sap.ist.i.kyoto-u.ac.jp
c)　enakamura@sap.ist.i.kyoto-u.ac.jp
d)　s.fukayama@aist.go.jp
e)　m.goto@aist.go.jp
f)　itoyama@ra.sc.e.titech.ac.jp
g)　yoshii@kuis.kyoto-u.ac.jp

表 1 ポピュラー音楽の伴奏を構成するパートの特徴

パート	音符	音域	リズム
リード	単音	中域 - 高域	複雑
ベース	単音	低域	複雑
ハーモニー	和音	広域	シンプル

そこで我々は図 1 に示す 3 つに大別される伴奏パートの機能に着目することで，音楽音響信号からカバー譜の生成を行う．これらのパートは典型的なバンドにおいては，リードギター・ベースギター・リズムギターにより演奏される．そこで本研究では，原曲の楽器編成によらず，3 本のギターで編成されるバンドで演奏可能かつ原曲を再現する楽譜の生成を目指す．

本稿では 3 本のマルコフチェインを持つ階乗型の隠れセミマルコフモデル (FHSMM) に基づく統計的手法を用いる（図 1）．マルコフチェインのうち一つはリズムギターのコード系列を表現し，残りの 2 つはコードが与えられた下でのリードギターのメロディ音高列，ベースギターのベース音高列を表現する．各ギターの低ランクのスペクトログラムはアクティベーションベクトルと基底スペクトル（和音基底もしくは単音基底）の積で表現される．ここで，時刻フレーム毎にひとつの基底のみがマルコフチェインによって選択され，アクティベートされる．混合音のスペクトログラムは各ギターのスペクトログラムの和によって表現される．この FHSMM は制約付きの非負値行列因子分解（NMF）として解釈可能である．観測データとして混合音のスペクトログラムが与えられると基底スペクトル，アクティベーションベクトル，潜在変数列はギブスサンプリングによって確率的に推定される．最終的に各ギターのパート譜はビタビアルゴリズムによって与えられる．

本研究の主な貢献は，同様な音色の複数楽器の音源分離と採譜を同時に行う枠組みの提案にある．また，従来の自動採譜のように忠実な楽譜の生成を目指すのではなく，複雑な音楽を近似的に採譜する新たな研究トピックの提案も本研究のもうひとつの貢献である．これは市販楽曲の実用的な自動採譜を可能にする手法開発の為の重要な一歩である．

2. 関連研究

本章では自動採譜とカバー譜生成の関連研究について述べる．

2.1 自動採譜

自動採譜のための手法として主に NMF [10, 19, 22] と確率的潜在要素解析 (PLCA) [2] が用いられている．Vincent ら [22] は NMF ベースの周波数帯域の異なる楽器のテンプレートパターンの重み付き和で観測を表現するモデルを提案した．Grindlay ら [10] は音色構造を明示的にモデル化することで複数楽器の採譜へ NMF を拡張した．Benetos ら [2] は楽器スペクトルの時間方向の依存関係を扱うために音状態スペクトルテンプレートを用いた PLCA ベースのモデルを提案している．Vincent らは [23] 非線形独立部分空間解析 (ISA) と階乗型の隠れマルコフモデル (FHMM) に基づく生成モデルを提案している．我々のモデルは FHMM を用いているという点で彼らのモデルに類似するが，継続長を考慮しセミマルコフモデルに拡張している点と各楽器の役割に基づき遷移確率を学習している点で異なる．

近年では，ニューラルネットワークも自動採譜タスクに用いられており良い結果を収めている [3,18]．Sigtia ら [18] は ニューラル音響モデルとニューラル音楽文法モデルを含む End-to-End 型のピアノ採譜モデルを提案している．Bittner ら [3] は畳み込みニューラルネットワーク (CNN) を用いた複数楽器音響信号からの多重 F0 音高推定の手法を提案している．

2.2 カバー譜生成

自動カバー譜生成についても様々な試みがなされているが，楽譜の記号領域で編曲を行いカバー譜を生成するもの [6, 21] と音響信号も考慮してカバー譜を生成するもの [1, 15] の二つに大別できる．Tuohy ら [21] は遺伝的アルゴリズム（GA）を用いて演奏可能なギターのカバー譜を生成するシステムを提案している．Chiu ら [6] は原曲の楽譜の重要な音は残しつつ音符を削減することでピアノカバー譜を生成するシステムを提案している．Percival ら [15] はターゲットの楽曲の音響特徴量とクラシック音楽の楽譜コーパスから得られる確率モデルを用いることで弦楽四重奏編成のカバー譜を生成するシステムを提案している．Ariga ら [1] は原曲の音響特徴量を用いて難易度調節が可能なギターカバー譜を生成するシステムを提案している．

3. 提案手法

本章ではポピュラー音楽の音響信号からカバー譜を生成する提案手法について述べる．我々のモデルでは，統一的な枠組みで多重音高推定と楽器パート割り当てを統合的に行う．このモデルでは音響スペクトログラムの確率的生成モデルを定式化し，その逆問題を解く．つまり，与えられた音響スペクトログラムを観測データとして，モデルに含まれる確率変数を推定することで楽譜を推定する．提案する FHSMM はリードギター・リズムギター・ベースギターに対応する 3 組の潜在変数を持つ (図 1)．

3.1 問題設定

我々が取り組む問題を以下のように定義する．

入力: 伴奏音の振幅スペクトログラム $\mathbf{X} \in \mathbb{R}_{+}^{F \times T}$ と 16 分音符単位のビート時刻
出力: 3 つのパートの楽譜

ここで F は周波数ビン数，T は時刻フレーム数であり，出力楽譜の拍子は 4/4 であると仮定する．

3.2 確率的階乗モデル

我々は音響スペクトログラム $\mathbf{X} \in \mathbb{R}_+^{F\times T}$ を，リズム・リード・ベースギターで演奏されるハーモニー・リード・ベースパートのスペクトログラム $\mathbf{X}^{\mathsf{R}}, \mathbf{X}^{\mathsf{L}}, \mathbf{X}^{\mathsf{B}} \in \mathbb{R}_+^{F\times T}$ の和で表現する (図 1).

$$x_{ft} = x^{\mathsf{R}}_{ft} + x^{\mathsf{L}}_{ft} + x^{\mathsf{B}}_{ft}. \tag{1}$$

$\mathbf{X}^{\mathsf{R}}, \mathbf{X}^{\mathsf{L}}, \mathbf{X}^{\mathsf{B}}$ の生成過程はそれぞれリズム・リード・ベースギターモデルとして表される．リズムギターモデルはコード系列とオンセット時刻情報を潜在変数列として表現し，和音のスペクトログラム $\mathbf{X}^{\mathsf{R}}$ を出力する．リードとベースギターモデルは音符列とオンセット時刻情報を潜在変数列として表現し，単旋律スペクトログラム $\mathbf{X}^{\mathsf{L}}$ と $\mathbf{X}^{\mathsf{B}}$ を出力する．これらの二つのモデルは同様にモデル化されるが，使う音域が異なる．リードギターは比較的高音域，ベースギターは低音域の音高を用いる．

提案モデルは各時刻のスペクトログラムを基底ベクトルの重み付き和で表現する NMF の制約付き派生モデルであるとみなすことができる．各ギターパートにおいて，各時刻フレームのスペクトラムはマルコフチェインによって選択された基底ベクトルに重み付けたものとしてモデル化される．全体として各時刻フレームのスペクトラムは，リズム・リード・ベースパートに対応する 3 つの基底スペクトルの重み付き和でモデル化される．

3.2.1 リズムギターモデル

$\mathbf{X}^{\mathsf{R}}$ の生成過程を表現する隠れセミマルコフモデル (HSMM) を定式化する (図 2)．リードギターモデルの潜在変数列はコード記号列 $\mathbf{Z}^{\mathsf{R}} = \{z^{\mathsf{R}}_1, \cdots, z^{\mathsf{R}}_N\}$ とオンセットビート位置列 $\mathbf{U}^{\mathsf{R}} = \{u^{\mathsf{R}}_1, \cdots, u^{\mathsf{R}}_N\}$ で表される．ここで，N はコードの種類数であり，z^{R}_n は $\{\mathrm{C}, \cdots, \mathrm{B}\} \times \{\mathrm{major}, \mathrm{minor}\}$ の 24 種類のうち一つをとる．また，u^{R}_n は 0 から 15 の整数をとり，16 分音符単位の小節内の相対的なビート位置を表す（本稿では拍子は 4/4 に限定する）．

コード記号列 $\mathbf{Z}^{\mathsf{R}}$ は以下のようなマルコフモデルによって表現される．

$$p\big(z^{\mathsf{R}}_1 \big| \boldsymbol{\pi}^{\mathsf{R}}\big) = \pi^{\mathsf{R}}_{z^{\mathsf{R}}_1}, \tag{2}$$

$$p\big(z^{\mathsf{R}}_n | z^{\mathsf{R}}_{n-1}, \boldsymbol{\psi}^{\mathsf{R}}\big) = \psi^{\mathsf{R}}_{z^{\mathsf{R}}_{n-1}, z^{\mathsf{R}}_n}, \tag{3}$$

ここで，π^{R}_a はコード a の初期確率，$\psi^{\mathsf{R}}_{a,b}$ はコード a からコード b への遷移確率である．

同様に，コードのオンセット列 $\mathbf{U}^{\mathsf{R}}$ は拍節マルコフモデル [11, 16] で以下のように表現される．

$$p\big(u^{\mathsf{R}}_n | u^{\mathsf{R}}_{n-1}, \boldsymbol{\phi}^{\mathsf{R}}\big) = \phi^{\mathsf{R}}_{u^{\mathsf{R}}_{n-1}, u^{\mathsf{R}}_n}, \tag{4}$$

ここで $\phi^{\mathsf{R}}_{a,b}$ はビート位置 a から b への遷移確率である．ただし，$a \geq b$ の時，そのコードは小節線をまたいで継続する．コードの最大継続長は 1 小節に制限されており，それを超える場合自己遷移で表現される．状態遷移の単位とビート時刻の単位は必ずしも一致しないのでこれはセミマルコフモデルであり，この拍節マルコフモデルは潜在状態の継続時間の分布を表現している．

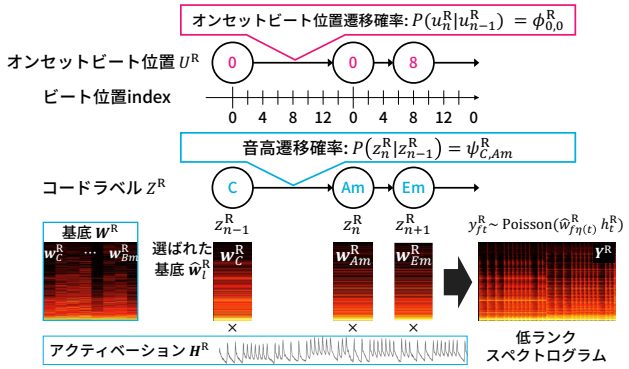

図 2 FHSMM に基づくリズムギターのスペクトログラムの生成過程.

リズムギターのスペクトログラム $\mathbf{X}^{\mathsf{R}}$ は基底スペクトル集合 $\mathbf{W}^{\mathsf{R}} = \{\mathbf{w}^{\mathsf{R}}_{\mathrm{C}}, \cdots, \mathbf{w}^{\mathsf{R}}_{\mathrm{Bm}}\} \in \mathbb{R}_+^{F\times 24}$ ($\mathbf{w}^{\mathsf{R}}_\chi$ はコード χ の基底スペクトルを表す) とアクティベーションベクトル $\mathbf{h}^{\mathsf{R}} \in \mathbb{R}_+^T$ を用いて以下のように生成される．

$$p\big(x^{\mathsf{R}}_{ft} | \mathbf{Z}^{\mathsf{R}}, \mathbf{U}^{\mathsf{R}}, \mathbf{W}^{\mathsf{R}}, h^{\mathsf{R}}_t\big) = \mathrm{Poisson}\Big(x^{\mathsf{R}}_{ft} | w^{\mathsf{R}}_{\eta^{\mathsf{R}}(t)f} h^{\mathsf{R}}_t\Big), \tag{5}$$

ここで，$\eta^{\mathsf{R}}(t)$ は時刻フレーム t が属するコードを表し，$\mathbf{Z}^{\mathsf{R}}$ と $\mathbf{U}^{\mathsf{R}}$ によって決定される．この定式化は KL-ダイバーシェンスに基づく NMF の確率モデル [19] から発想を得ている．ただし，各時刻フレーム t において $\eta^{\mathsf{R}}(t)$ によって指定された一つの基底ベクトルのみがアクティベートされ，和音スペクトログラム $\mathbf{X}^{\mathsf{R}}$ を表現している点が，我々のモデルの特徴である．

3.2.2 リード・ベースギターモデル

$\mathbf{Z}^{\mathsf{R}}$ と $\mathbf{U}^{\mathsf{R}}$ によって指定されたコード列の制約下での $\mathbf{X}^{\mathsf{L}}$ または $\mathbf{X}^{\mathsf{B}}$ の生成過程を表現する条件付き HSMM を定式化する．簡単のため，“$*$” は “L”(リードギター) もしくは “B”(ベースギター) を表すものとする．HSMM の潜在変数列は音高列 $\mathbf{Z}^* = \{z^*_1, \cdots, z^*_{M^*}\}$ とオンセットビート位置列 $\mathbf{U}^* = \{u^*_1, \cdots, u^*_{M^*}\}$ で表される．ここで M^* はリード・ベースパートの音符数であり，z^{L}_m は $\{\mathrm{E3}, \cdots, \mathrm{C6}\}$ の 33 種類の音高の内の一つ，z^{B}_m は $\{\mathrm{E1}, \cdots, \mathrm{G3}\}$ の 28 種類の音高のうち一つを取り，u^*_m は 0 から 15 の整数をとる．

音高列 $\mathbf{Z}^*$ は以下のように条件付きマルコフモデルによって表現される．

$$p\big(z^*_1 | \mathbf{Z}^{\mathsf{R}}, \mathbf{U}^{\mathsf{R}}, \boldsymbol{\pi}^*\big) = \pi^*_{z^*_1, z^{\mathsf{R}}_1}, \tag{6}$$

$$p\big(z^*_m | z^*_{m-1}, \mathbf{Z}^{\mathsf{R}}, \mathbf{U}^{\mathsf{R}}, \boldsymbol{\psi}^*\big) = \psi^*_{z^{\mathsf{R}}_{\rho^*(m)}, z^*_{m-1}, z^*_m}, \tag{7}$$

ここで，$\rho^*(m)$ は音符 z^*_m の属するフレームのリズムギターによって推定されたコード，$\pi^*_{c,a}$ はコード c の下での音高 a の初期確率，$\psi^*_{c,a,b}$ はコード c の下での音高 a から b への遷移確率を表す．

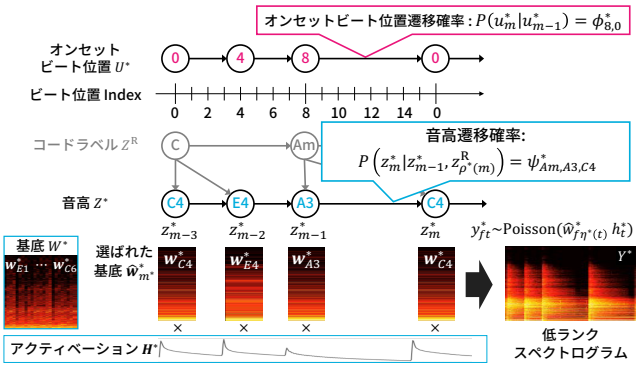

図 3 HSMM に基づくリードまたはベースギタースペクトログラムの生成過程.

同様に，音符のオンセットビート位置列 $\mathbf{U}^*$ は以下のような拍節マルコフモデルによって表現される.

$$p\left(u^*_m|u^*_{m-1}, \boldsymbol{\phi}^*\right) = \phi^*_{u^*_{m-1},u^*_m}, \tag{8}$$

ここで $\phi^*_{a,b}$ はビート位置 a から b への遷移確率である.

リード・ベースギターパートのスペクトログラム $\mathbf{X}^*$ は基底スペクトル集合 $\mathbf{W}^*$ とアクティベーションベクトル $\mathbf{h}^* \in \mathbb{R}^T_+$ を用いて以下のように生成される.

$$p\left(x^*_{ft}|\mathbf{Z}^*, \mathbf{U}^*, \mathbf{W}^*, h^*_t\right) = \mathrm{Poisson}\left(x^*_{ft}|w^*_{\eta^*(t)f} h^*_t\right), \tag{9}$$

ここで, $\eta^*(t)$ はフレーム t が属する音符を表し, $\mathbf{Z}^*$ と $\mathbf{U}^*$ によって決定される．ただし, $\mathbf{W}^*$ は $\mathbf{W}^{\mathsf{L}} = \{\mathbf{w}^{\mathsf{L}}_{\mathrm{E3}}, \cdots, \mathbf{w}^{\mathsf{L}}_{\mathrm{C6}}\} \in \mathbb{R}^{F\times 33}_+$ または $\mathbf{W}^{\mathsf{B}} = \{\mathbf{w}^{\mathsf{B}}_{\mathrm{E1}}, \cdots, \mathbf{w}^{\mathsf{B}}_{\mathrm{G3}}\} \in \mathbb{R}^{F\times 28}_+$によって与えられる．ここで $\mathbf{w}^*_\chi$ は音符 χ の基底ベクトルを表す.

3.3 ベイジアン定式化

3.2.1 章と 3.2.2 章の 3 つのサブモデルを統合することで，以下のようなセミ-ベイジアンモデルを定式化する.

$$\begin{aligned} &p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\Theta}) \\ &= p(\mathbf{X}|\mathbf{Z}, \mathbf{U}, \mathbf{W}, \mathbf{H}) p(\mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\psi}) p(\mathbf{U}|\boldsymbol{\phi}) p(\mathbf{W}) p(\mathbf{H}), \end{aligned} \tag{10}$$

ここで，$\boldsymbol{\Theta} = \{\pi^{\mathsf{R}}, \pi^{\mathsf{L}}, \pi^{\mathsf{B}}, \psi^{\mathsf{R}}, \psi^{\mathsf{L}}, \psi^{\mathsf{B}}, \phi^{\mathsf{R}}, \phi^{\mathsf{L}}, \phi^{\mathsf{B}}\}$ は事前学習される一連のパラメータ，$\mathbf{Y} = \{\mathbf{Z}, \mathbf{U}, \mathbf{W}, \mathbf{H}\}$ は (観測スペクトログラム $\mathbf{X}$ に合わせてその都度推定される) 確率変数である．ここで $\mathbf{Z} = \{\mathbf{Z}^{\mathsf{R}}, \mathbf{Z}^{\mathsf{L}}, \mathbf{Z}^{\mathsf{B}}\}$ であり，$\mathbf{U}, \mathbf{W}, \mathbf{H}$ も同様に決定される.

事前分布 $p(\mathbf{W})$ と $p(\mathbf{H})$ を置くことで定式化が完結する．以下のようにガンマ事前分布 $\mathbf{W}$ を定義する.

$$w^{\mathsf{R}}_{kf} \sim \mathcal{G}\left(a_{w^{\mathsf{R}}_{kf}}, b_{w^{\mathsf{R}}_{kf}}\right), \tag{11}$$

$$w^*_{if} \sim \mathcal{G}\left(a_{w^*_{if}}, b_{w^*_{if}}\right), \tag{12}$$

ここで $k \in \{\mathrm{C}, \cdots, \mathrm{Bm}\}$ はコードを表し, $i \in \{\mathrm{E3}, \cdots, \mathrm{C6}\}$ または $\{\mathrm{E1}, \cdots, \mathrm{G3}\}$ は音高を表し，a_w と b_w はハイパーパラメータであり，形状母数と尺度母数の逆数である．ポピュラー音楽におけるハーモニーパートの曖昧さを表現するために，リードギターの基底ベクトル $\mathbf{W}^{\mathsf{R}}$ に事前分布を弱く置く (つまり，$a_{w^{\mathsf{R}}_f}$ と $b_{w^{\mathsf{R}}_f}$ を $a_{w^*_f}$ と $b_{w^*_f}$ よりも小さく設定する)．*i.e.*, $\mathbf{W}$ の事前分布の期待値が和音テンプレートや単音テンプレートと一致するようにハイパーパラメータはあらかじめ設定される．同様に，$\mathbf{H}$ の事前分布は以下のように設定する.

$$h^{\mathsf{R}}_t \sim \mathcal{G}\left(a_{h^{\mathsf{R}}_t}, b_{h^{\mathsf{R}}_t}\right), \tag{13}$$

$$h^*_t \sim \mathcal{G}\left(a_{h^*_t}, b_{h^*_t}\right), \tag{14}$$

ここで，a_{h_t} と b_{h_t} はハイパーパラメータである.

3.4 ベイズ推論

音楽のスペクトログラム $\mathbf{X}$ が与えられると，ベイズ則より事後分布を計算できる.

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta}) = p(\mathbf{X}, \mathbf{Y}|\boldsymbol{\Theta}) / p(\mathbf{X}|\boldsymbol{\Theta}). \tag{15}$$

この事後分布は解析的に求められないので，ギブスサンプリングを交互に繰り返し行うことで潜在変数 $\mathbf{Z}$ と $\mathbf{U}$，基底スペクトル $\mathbf{W}$，アクティベーション $\mathbf{H}$ をサンプリングできる．つまり，$\mathbf{G} \subset \mathbf{Y}$ となるようなサンプルを条件付き事後分布 $p(\mathbf{G}|\mathbf{Y}_{\neg\mathbf{G}}, \mathbf{X}, \boldsymbol{\Theta})$ から得られる．ここで $\mathbf{Y}_{\neg\mathbf{G}}$ は $\mathbf{Y}$ から $\mathbf{G}$ を除いた補集合を表す．簡単のため今後 $\boldsymbol{\Theta}$ の依存関係は明記しない.

3.4.1 潜在変数 $\mathbf{Z}^{\mathsf{R}}$ と $\mathbf{U}^{\mathsf{R}}$ の更新

$p(\mathbf{Z}^{\mathsf{R}}, \mathbf{U}^{\mathsf{R}}|\mathbf{Y}_{\neg\mathbf{Z}^{\mathsf{R}},\mathbf{U}^{\mathsf{R}}}, \mathbf{X})$ から $\mathbf{Z}^{\mathsf{R}}$ と $\mathbf{U}^{\mathsf{R}}$ をサンプルするために，効率的なフォワードフィルタリング-バックワードワンプリングアルゴリズムを用いる．フォワードフィルタリングでは，フォワードメッセージ $\alpha^{\mathsf{R}}(z^{\mathsf{R}}_n, u^{\mathsf{R}}_n)$ が系列の先頭から再帰的に以下のように計算される.

$$\alpha^{\mathsf{R}}\left(z^{\mathsf{R}}_1, u^{\mathsf{R}}_1\right) = p\left(z^{\mathsf{R}}_1\right) = \pi^{\mathsf{R}}_{z^{\mathsf{R}}_1}, \tag{16}$$

$$\begin{aligned} \alpha^{\mathsf{R}}\left(z^{\mathsf{R}}_n, u^{\mathsf{R}}_n\right) &= p\left(x_{\tau^{\mathsf{R}}(n-1):\tau^{\mathsf{R}}(n)+1}|z^{\mathsf{R}}_n, u^{\mathsf{R}}_n\right) \\ &\sum_{z^{\mathsf{R}}_{n-1}} \sum_{u^{\mathsf{R}}_{n-1}} \psi^{\mathsf{R}}_{z^{\mathsf{R}}_n, z^{\mathsf{R}}_{n-1}} \phi^{\mathsf{R}}_{u^{\mathsf{R}}_n, u^{\mathsf{R}}_{n-1}} \alpha\left(z^{\mathsf{R}}_{n-1}, u^{\mathsf{R}}_{n-1}\right), \end{aligned} \tag{17}$$

ここで，$\tau^{\mathsf{R}}(n)$ は n 番目のコードの最後のフレームを表し，$x_{a:b}$ は $\{x_a, \cdots, x_b\}$ を表す.

バックワードサンプリングでは，以下のようにバックワードメッセージ $\gamma^{\mathsf{R}}(z^{\mathsf{R}}_n, u^{\mathsf{R}}_n)$ 再帰的に計算しながら，z^{R}_n と u^{R}_n を系列の終端から逆順に再帰的にサンプリングを行う.

$$\gamma^{\mathsf{R}}\left(z^{\mathsf{R}}_N, u^{\mathsf{R}}_N\right) = p\left(z^{\mathsf{R}}_N, u^{\mathsf{R}}_N|\mathbf{X}\right) \propto \alpha^{\mathsf{R}}\left(z^{\mathsf{R}}_N, u^{\mathsf{R}}_N\right), \tag{18}$$

$$\begin{aligned} \gamma^{\mathsf{R}}\left(z^{\mathsf{R}}_n, u^{\mathsf{R}}_n\right) &= p\left(z^{\mathsf{R}}_n, u^{\mathsf{R}}_n|z^{\mathsf{R}}_{n+1:N}, u^{\mathsf{R}}_{n+1:N}, \mathbf{X}\right) \\ &\propto \psi^{\mathsf{R}}_{z^{\mathsf{R}}_n, z^{\mathsf{R}}_{n+1}} \phi^{\mathsf{R}}_{u^{\mathsf{R}}_n, u^{\mathsf{R}}_{n+1}} \alpha\left(z^{\mathsf{R}}_n, u^{\mathsf{R}}_n\right). \end{aligned} \tag{19}$$

最終的に譜面を生成する際には z^{R}_n と u^{R}_n のサンプリングをビタビアルゴリズムにすることで，各時刻ステップの $\gamma^{\mathsf{R}}(z^{\mathsf{R}}_n, u^{\mathsf{R}}_n)$ を最大化する最尤な $\mathbf{Z}^{\mathsf{R}}$ と $\mathbf{U}^{\mathsf{R}}$ を得ることができる.

3.4.2 潜在変数 Z* と U* の更新

同様にフォワードフィルタリング-バックワードサンプリングアルゴリズムで $p(\mathbf{Z}^*,\mathbf{U}^*|\mathbf{Y}_{\neg\mathbf{Z}^*,\mathbf{U}^*},\mathbf{X})$ から $\mathbf{Z}^*$ と $\mathbf{U}^*$ をサンプルすることができる．フォワードフィルタリングでは，フォワードメッセージ $\alpha^*(z_n^*,u_n^*)$ が系列の先頭から再帰的に以下のように計算される．

$$\alpha^*(z_1^*,u_1^*) = p\big(z_1^*|z_1^{\mathsf{R}}\big) = \pi^*_{z_1^*,z_1^{\mathsf{R}}}, \tag{20}$$

$$\alpha^*(z_m^*,u_m^*) = p\big(x_{\tau^*(m-1):\tau^*(m)+1}|z_m^*,u_m^*\big) \sum_{z_{m-1}^*}\sum_{u_{m-1}^*} \psi^*_{z^{\mathsf{R}}_{\rho^*(m)},z_{m-1}^*,z_m^*}\phi^*_{u_m^*,u_{m-1}^*}\alpha\left(z_{m-1}^*,u_{m-1}^*\right). \tag{21}$$

バックワードサンプリングでは，バックワードメッセージ $\gamma^*(z_n^{\mathsf{R}},u_n^{\mathsf{R}})$ を再帰的に以下のように計算することで z_n^* と u_n^* を再帰的に系列の終端から逆順にサンプルすることができる．

$$\gamma^*(z_{M^*}^*,u_{M^*}^*) = p(z_{M^*}^*,u_{M^*}^*|\mathbf{X}) \propto \alpha^*(z_{M^*}^*,u_{M^*}^*), \tag{22}$$

$$\gamma^*(z_m^*,u_m^*) = p\left(z_m^*,u_m^*|z_{m+1:M^*}^*,u_{m+1:M}^*,\mathbf{X}\right) \propto \psi^*_{z^{\mathsf{R}}_{\rho^*(m)},z_{m+1}^*,z_m^*}\phi^*_{u_m^*,u_{m+1}^*}\alpha\left(z_m^*,u_m^*\right). \tag{23}$$

3.4.1 章と同様に譜面を最終的に生成する際には，ビタビアルゴリズムで最尤な $\mathbf{Z}^*$ と $\mathbf{U}^*$ を推定することができる．

3.4.3 基底スペクトル W とアクティベーション H の更新

ベイジアン NMF [5] の推論と同様に，$\mathbf{W}$ と $\mathbf{H}$ は条件付き事後分布 $p(\mathbf{W},\mathbf{H}|\mathbf{Z},\mathbf{U},\mathbf{X})$ から直接サンプリング可能である．最新の $\mathbf{W}$ と $\mathbf{H}$ のサンプル値を使って得られる補助変数 λ_{ft} を以下のように定義する．

$$\lambda^*_{ft} = \frac{\hat{w}^*_{f\eta(t)}h^*_t}{\hat{w}^{\mathsf{R}}_{f\eta^{\mathsf{R}}(t)}h^{\mathsf{R}}_t + \hat{w}^{\mathsf{L}}_{f\eta^{\mathsf{L}}(t)}h^{\mathsf{L}}_t + \hat{w}^{\mathsf{B}}_{f\eta^{\mathsf{B}}(t)}h^{\mathsf{B}}_t}, \tag{24}$$

$\boldsymbol{\lambda}$ を用いることで，$\mathbf{W}$ を以下のようにサンプリングできる．

$$w^*_{if} \sim \mathcal{G}\left(\alpha^*_{w^*_{if}} + \sum_t x_{ft}\lambda^*_{ft},\ \gamma^*_{w^*_{if}} + \sum_t h^*_t\right) \tag{25}$$

同様に，$\mathbf{H}$ も以下のようにサンプリングできる．

$$h^*_t \sim \mathcal{G}\left(\alpha^*_{h^*_t} + \sum_f x_{ft}\lambda^*_{ft},\ \gamma^*_{h^*_t} + \sum_f w^*_f\right) \tag{26}$$

$\mathbf{W}^{\mathsf{R}}$ と $\mathbf{H}^{\mathsf{R}}$ についても同様にサンプル可能である．

3.5 楽譜生成

$\mathbf{Z},\mathbf{U},\mathbf{H}$ を用いることで 3 つのギターのパート譜を生成できる．$\mathbf{Z},\mathbf{U},\mathbf{W},\mathbf{H}$ のサンプリングを充分な回数行った後にビタビ復号を行うことで，近似的に $\mathbf{Z}$ と $\mathbf{U}$ の最大事後確率推定が可能となる．事後確率を最大化する $\mathbf{H}$ は条件付き事後分布から得られる．

$\mathbf{Z}^*$ と $\mathbf{U}^*$ がそれぞれ音高とオンセットビート位置を表すので，リードギターとベースギターのパート譜はこれらの潜在変数から生成できる．ただし，同音の繰り返しは自己遷移で表現される（例えば，$z^{\mathsf{L}}_{n-1} = z^{\mathsf{L}}_n = \mathrm{C3}$）．一方で，リズムギターの実際のオンセット・音価はコードシンボル $\mathbf{Z}^{\mathsf{R}}$ と継続長 $\mathbf{U}^{\mathsf{R}}$ からは直接は決定できない．そこで，音楽的に自然な小節内のリズムパターンを得るために，一小節単位のリズムパターンのテンプレートの辞書（16 次元バイナリベクトル）を楽譜データから作成し，コサイン距離に基づくテンプレートマッチングを行う (図 4)．アクティベーション $\mathbf{H}^{\mathsf{R}}$ の大体のピークの位置を検出して最も近いパターンを割り当てることで，リズムパターンの推定を行う．

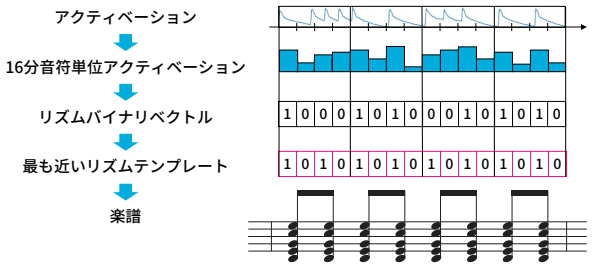

図 4 テンプレートマッチングに基づくリズムパターン推定．

4. 評価実験

提案モデルを評価するために多重音高推定とカバー譜生成の二つの観点から比較実験を行った．最初に，リズム・リード・ベースギターの 3 つのパートのみから成るボーカルやドラムを含まない音響信号に対して，多重音高推定と楽器割り当てを行い精度を評価した．次に，様々な楽器で演奏される実際のポピュラー音楽に対してカバー譜生成を行い精度を評価した．

4.1 実験条件

オクターブあたりのビン数が 96，シフト幅 10ms，周波数帯域が 32.7 Hz (C1) から 8372.0 Hz (C9) の対数振幅スペクトログラムを定 Q 変換 [17] によって得た．ビート時刻は予めビートトラッキング手法 [4] によって得られたものを用いた．事前分布 $\mathbf{W}^{\mathsf{R}}$ のハイパーパラメータ $\mathbf{a}^{\mathsf{R}}_w,\mathbf{b}^{\mathsf{R}}_w$ はアコースティックギターでローポジションの 24 種類のコードを弾いた MIDI シンセサイザで演奏した際のスペクトルから決定した．事前分布 $\mathbf{W}^*$ のハイパーパラメータ $\mathbf{a}^*_w,\mathbf{b}^*_w$ は同様に E3 から C6 の 33 種類，E1 から G3 の 28 種類の音高を MIDI シンセサイザで演奏した際のスペクトルから決定した．アクティベーションのハイパーパラメータは経験的に $a_{h^{\mathsf{R}}_t} = a_{h^*_t} = 1, b_{h^{\mathsf{R}}_t} = b_{h^*_t} = 1$ とした．初期確率は $\boldsymbol{\pi}^{\mathsf{L}} = 1/33, \boldsymbol{\pi}^{\mathsf{B}} = 1/28, \boldsymbol{\pi}^{\mathsf{R}} = 1/24$ とした．また ϕ と ψ はビートルズを含む 7 アーティストの 132 曲の楽譜から事前学習した．モデルの精度を以下に示す 16 部音符単位の再現率，適合率，F 値で評価した．

表 2 多重音高推定・楽器割当の評価結果.

手法	パート	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
FHSMM	リード	27.7 (46.5)	31.5 (50.6)	26.0 (44.7)
	ベース	72.2 (78.2)	70.8 (76.7)	73.6 (79.8)
	リズム	52.0 (77.4)	45.2 (76.4)	62.5 (78.4)
NMF	リード	16.1 (33.6)	20.2 (39.4)	14.0 (30.9)
	ベース	26.3 (77.2)	25.6 (75.6)	27.0 (78.8)
	リズム	42.8 (65.2)	37.4 (67.0)	51.0 (63.6)

*カッコ内の数はオクターブ誤りを許容した再現率，適合率，F 値.

$$\mathcal{P} = \frac{\mathsf{N}_{\mathrm{tp}}}{\mathsf{N}_{\mathrm{sys}}},\ \mathcal{R} = \frac{\mathsf{N}_{\mathrm{tp}}}{\mathsf{N}_{\mathrm{ref}}},\ \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}}, \tag{27}$$

ここで N_{tp} は正しく推定された音符数，$\mathsf{N}_{\mathrm{sys}}$ は推定された音符数，$\mathsf{N}_{\mathrm{ref}}$ は正解楽譜の音符の総数である．

4.2 複数楽器多重音高推定

提案手法の多重音高推定・楽器割り当ての精度を定量的に評価した．リズム・リード・ベースギターが明瞭である３曲のポピュラー音楽 (RWC-MDB-P-2001 No. 12, No. 51, No. 70) を RWC 音楽データベース [9] から選び，実験に用いた．３つのパートのパート別音源を足し合わせることで３つの楽器のみから成る音源を作成し実験に用いた．比較のために 85 個の基底 (リズムギター基底 24 個, リードギター基底 33 個，ベースギター基底 28 個) を持つ教師付き NMF で，それぞれのパートで各フレームごとに最も大きなアクティベーションの基底が選択されるものを実装し，同様に精度を評価した．

表 2 に示す通り，提案手法が比較手法よりも精度が高かった．図 5 に示す通り，提案手法ではメロディパート，ベースパートに正しく追従できており，楽器割り当てを正しく行えている．提案手法のリードギター推定結果の F 値が 27.7% と低いが，オクターブ誤りを許容すると 46.5% まで上がり，これは比較手法の 33.6% を大きく上回る．比較手法に比べ提案法の優位性は示されたものの，依然として精度は低く，改善の余地は多い．局所解を抜け出し，このようなオクターブ誤りを減らすには，$\mathbf{Z}$ と $\mathbf{U}$ のサンプリングにオクターブ誤りを考慮した提案分布を持つマルコフ連鎖モンテカルロ (MCMC) 法を用いることが考えられる．

4.3 カバー譜生成

提案手法によるカバー譜生成の精度を評価した．RWC 研究用音楽データベースのポピュラー音楽 100 曲のうち，拍子が 4/4 でありビートトラッキングが正しく行われた 84 曲を実験に用いた．前処理として，調波打楽器音分離 (HPSS) [7] と歌声分離 [12] を行い，伴奏音のみを抜き出した．再現率，適合率，F 値は式 (27) で計算される．ただし，伴奏パートは必ずしもギターのみならず様々な楽器で演奏されており，それを３つのパートのみで近似するので，再現率，適合率，F 値は最も高い場合でも 100 になるとは限らない．

表 3 84 曲のポピュラー音楽 (RWC-MDB-P-2001) に対するカバー譜推定結果.

評価尺度	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
正解	46.0	44.1	50.0
オクターブ誤りを許容した正解	73.0	70.9	76.5

生成した結果の一例を図 7 に示す．生成した楽譜の再現率，適合率，F 値を表 3 に，84 曲の F 値の分布を図 6 に示す．図 6 から分かるように，F 値は曲によって差が大きい．結果を分析したところ，多くのパートから成る楽曲 (6 パート以上等) では F 値は低くなる傾向があった．このような楽曲に対処するためには，余分な音を吸収するノイズモデルが必要である．同様に，対象楽曲の構成楽器が極端に少ない場合 (弾き語り等) でも F 値が低くなる傾向であった．我々のモデルは 3 つのパートを仮定しているので，このような楽曲を扱えないのはモデルの限界といえる．一方で，図 7 に示すように，3 つのパートのみで曲の大部分を再現できている生成結果も見られた．このような生成結果は，我々のモデルの能力を表しているといえる．しかしながら，まだ改善の余地も大いにある．例えば，特にリードギターで音楽的に不自然なオンセットや音価がみられた．この問題の原因の一つは，現在のモデルで休符を扱っていないことである．ノイズモデルの導入で休符の生成を可能にすることで，より自然な生成結果が期待できる．

5. おわりに

本稿では多重音響信号からのカバー譜を生成する枠組みを提案し，原曲を近似的に再構成する自動採譜の新しい研究トピックについて述べた．我々のモデルはリズムギターのコード列，リードギターのメロディ，およびベースギターのベースラインに対応する３本のマルコフチェインを持つ FHSMM で構成される．楽曲のスペクトログラムはそれらのマルコフチェインから独立に生成される低ランクのスペクトログラムの和で表現される．ポピュラー音楽のような複雑な音響信号からのカバー譜生成において，本手法が有効であることを実験により示した．

本手法にはまだ改善の余地が多く残されている．主にリードギターが休符を表現できるようにするために，ノイズモデルの導入が必要だと考えられる．また，一次マルコフモデルでは音符の遷移を十分に表現できないため，音楽的に不自然な楽譜の生成を避けられなかった．考えられる解決策として，変分オートエンコーダ (VAE) [13] や敵対

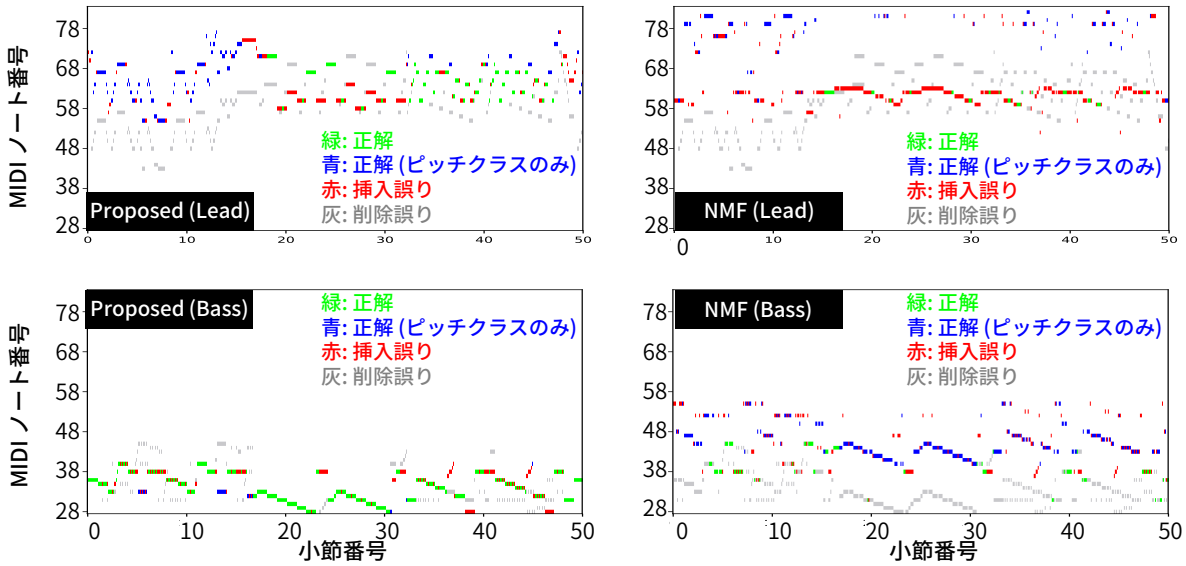

図 5 RWC-MDB-P-2001 No.12. のリード・ベースギターのピアノロール

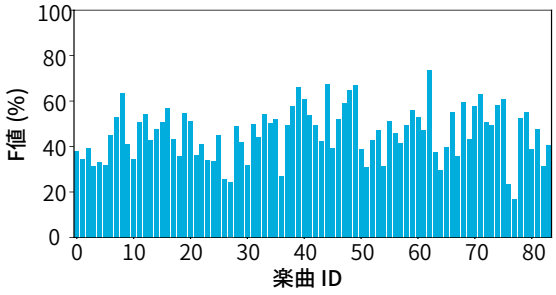

図 6 カバー譜生成結果の F 値の分布.

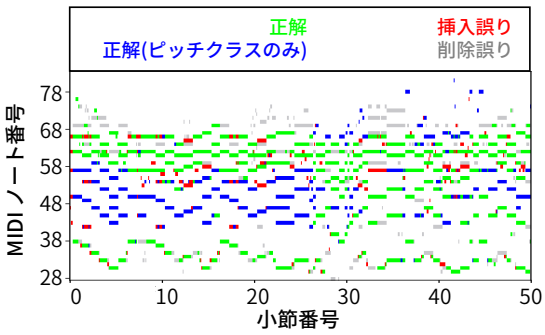

図 7 RWC-MDB-P-2001 No.51/Piece ID 45. の生成結果
コンデンスピアノロール

的生成ネットワーク (GAN) [8] のような深層生成モデルの導入が考えられる．さらに，ノンパラメトリックベイズのような手法を取り入れることで，３本のギターに限らず任意の数のパート譜を生成できるシステムへの拡張も考えられる．将来的に，自動ドラム採譜や歌声採譜と組あわせることでボーカル・ギター・ドラムから成るバンド編成のカバー演奏の為の楽譜を生成するシステムの開発を目指している．

6. 謝辞

本研究の一部は，JSPS 科研費 No. 26700020，No. 16H01744，JSPS 特別研究員奨励費 No. 16J05486，および JST ACCEL No. JPMJAC1602 の支援を受けた．

参考文献

[1] Shunya Ariga, Satoru Fukayama, and Masataka Goto. Song2Guitar: A difficulty-aware arrangement system for generating guitar solo covers from polyphonic audio of popular music. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 568–574, 2017.

[2] Emmanouil Benetos, Roland Badeau, Tillman Weyde, and Gaël Richard. Template adaptation for improving automatic music transcription. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 175–180, 2014.

[3] Rachel M Bittner, Brian McFee, Justin Salamon, Peter Li, and Juan P Bello. Deep salience representations for f0 estimation in polyphonic music. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 23–27, 2017.

[4] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. Madmom: A new python audio and music signal processing library. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1174–1178. ACM, 2016.

[5] Ali Taylan Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational intelligence and neuroscience*, 2009.

[6] Shih-Chuan Chiu, Man-Kwan Shan, and Jiun-Long Huang. Automatic system for the arrangement of piano reductions. In *IEEE International Symposium on Multimedia (ISM)*, pages 459–464, March 2009.

[7] Derry Fitzgerald. Harmonic/percussive separation using median filtering. In *International Conference on Digital Audio Effects (DAFX)*, pages 1–4, 2010.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in*

Neural Information Processing Systems, pages 2672–2680, 2014.

[9] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical and jazz music databases. In *International Conference on Music Information Retrieval (ISMIR)*, pages 287–288, 2002.

[10] Graham Grindlay and Daniel PW Ellis. Transcribing multi-instrument polyphonic music with hierarchical eigeninstruments. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1159–1169, 2011.

[11] M. Hamanaka, M. Goto, H. Asoh, and N. Otsu. A learning-based quantization: Unsupervised estimation of the model parameters. In *Proc. International Computer Music Association*, pages 369–372, 2003.

[12] Andreas Jansson, Eric J Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep U-Net convolutional networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 323–332, 2017.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Emilio Molina, Lorenzo J Tardón, Ana M Barbancho, and Isabel Barbancho. Sipth: Singing transcription based on hysteresis defined on the pitch-time curve. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(2):252–263, 2015.

[15] Graham Percival, Satoru Fukayama, and Masataka Goto. Song2Quartet: A system for generating string quartet cover songs from polyphonic audio of popular music. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 114–120, 2015.

[16] C. Raphael. A hybrid graphical model for rhythmic parsing. *Artificial Intelligence*, 137:217–238, 2002.

[17] Christian Schörkhuber and Anssi Klapuri. Constant-Q transform toolbox for music processing. In *Sound and Music Computing Conference, Barcelona, Spain*, pages 3–64, 2010.

[18] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(5):927–939, 2016.

[19] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180. IEEE, 2003.

[20] Carl Southall, Ryan Stables, and Jason Hockman. Automatic drum transcription using bi-directional recurrent neural networks. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 591–597, 2016.

[21] Daniel R Tuohy and Walter D Potter. GA-based music arranging for guitar. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 1065–1070. IEEE, 2006.

[22] Emmanuel Vincent, Nancy Bertin, and Roland Badeau. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):528–537, 2010.

[23] Emmanuel Vincent and Xavier Rodet. Music transcription with isa and hmm. In *International Conference on Independent Component Analysis and Signal Separation*, pages 1197–1204. Springer, 2004.

[24] Chih-Wei Wu and Alexander Lerch. Drum transcription using partially fixed non-negative matrix factorization with template adaptation. In *International Society for Music Information Retrieval Conference (ISMIR)*, pages 257–263, 2015.