Vol.2020-MUS-128 No.11
2020/8/25

スペクトログラムとピッチグラムの深層クラスタリングに基づく複数楽器パート採譜

田中 啓太郎[1,a)]　中塚 貴之[1,b)]　錦見 亮[2,c)]　吉井 和佳[2,d)]　森島 繁生[3,e)]

概要：本稿では，任意の複数楽器で演奏された音楽音響信号に対し，各楽器パートのピアノロールを推定するための深層クラスタリングに基づく採譜手法について述べる．採譜対象の楽曲が常に特定の楽器で演奏されている場合，各ピアノロールを得るための直接的な方法は，深層ニューラルネットワーク (deep neural network, DNN) を用いて各楽器ごとのピッチグラム（音高サリエンシースペクトログラム）を推定する手法である．しかしながらこの手法には，事前指定外の楽器を含む楽曲を取り扱うことができないという致命的な限界がある．本研究では，楽器に依存しない音高推定器を用いてコンデンスピッチグラムを推定した後，深層球面クラスタリングによって指定した数の楽器パートに分離する．採譜精度向上のため，各楽器の音色特徴量と音高特徴量に基づくスペクトログラムとピッチグラムの同時クラスタリングを提案する．実験では，提案手法により事前指定外の未知楽器を含む楽曲に対しても，既知楽器のみで構成される楽曲とほぼ同程度かつ最先端の精度で採譜を行うことができることを確認した．

1. はじめに

音楽音響信号中の多重音基本周波数推定 (multi-pitch estimation, MPE) [6] は，音楽音響信号を自動的に楽譜化する自動採譜にとって基盤となる技術であり，音楽情報処理において重要な役割を担っている [2]．従来の MPE の手法は，主に一種類の楽器で構成された音楽音響信号の採譜を目的としていた．この単一楽器 MPE は深層学習によって精度が大幅に改善し，近年では単一楽器 MPE を拡張し一般化を進めたものとして複数楽器 MPE が研究されている．複数楽器 MPE とは，複数楽器で構成された音楽音響信号中の各楽器に対して，各時刻の演奏音を表すピッチグラム（音高サリエンシースペクトログラム）を推定するものである．複数楽器の場合，各ピッチグラムの帰属先となる楽器を推定する必要が追加される．この困難を軽減するため，先行研究 [15], [19] では対象楽器を少数の事前指定楽器に限定してきた．この条件下では，複数楽器 MPE においてクラス分類手法を適用することで，音楽音響信号から直接各ピッチグラムを推定する方法が考えられる．

1 早稲田大学大学院先進理工学研究科
2 京都大学大学院情報学研究科
3 早稲田大学理工学術院総合研究所
a) phys.keitaro1227@ruri.waseda.jp
b) t59nakatsuka@fuji.waseda.jp
c) nishikimi@sap.ist.i.kyoto-u.ac.jp
d) yoshii@kuis.kyoto-u.ac.jp
e) shigeo@waseda.jp

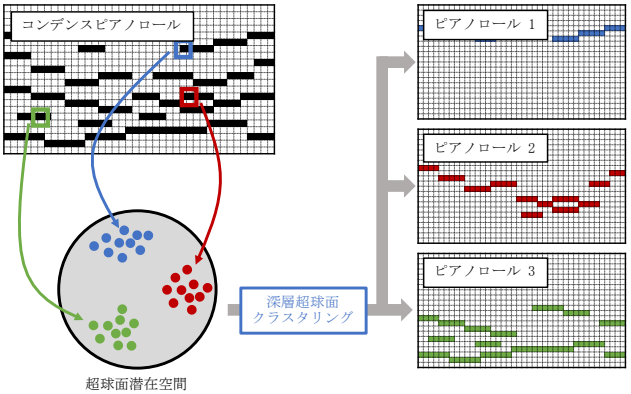

図 1　提案するクラスタリングによるパート譜採譜の概念図

これらのクラス分類に基づく手法は，特にクラシック音楽のように構成楽器がほぼ固定されている楽曲に対し，教師あり学習の枠組みにおいて成功を収めてきた．しかし，ポップスや EDM のように多くの種類の楽器が使用される現代の楽曲に対して質の高い自動採譜を達成するためには，採譜対象となる楽器に制約のないことが理想的である．

音声信号処理における話者分離では，類似した研究として任意話者の音声分離がなされている [12], [20]．深層ニューラルネットワーク (deep neural network, DNN) で対象の任意性を扱う際には，順列に関連した技術的問題が生じる．DNN では入力が与えられると，予め定義されたラベルに対応する出力次元へ決定論的に写像される．そのため，異なる出力次元間の順列並べ替えが許容されない．こ

© 2020 Information Processing Society of Japan

の順列問題を解決するため，任意話者に対する音声分離をクラス分類ではなくクラスタリング問題として捉える，深層クラスタリングと呼ばれる手法が提案されている [12]．この手法は類似度行列を使用することで，上記順列問題を回避すると同時に最適なクラスタリングを可能にする．

本稿では，任意の複数楽器で演奏された音楽音響信号に対し，深層クラスタリングに基づいて各楽器パートのピアノロールを推定する新たな手法を提案する．楽器に依存しない音高推定器を用いて，全楽器の演奏音をサリエンシーによって表すピッチグラム（コンデンスピッチグラム）を推定した後，深層球面クラスタリングによって指定した数の楽器パートに分離することで，各楽器パートのピアノロールである各パートピッチグラムを得る．クラスタリング時にピッチグラムに加えてスペクトログラムも考慮することで，楽曲に含まれる楽器の音色と音高双方の特徴に基づいた最適なクラスタリングが可能となる．また，MPE と音源分離の間には相補関係がある [9], [11] ため，スペクトログラムとピッチグラムの同時クラスタリングを行うことで採譜精度のさらなる向上を図る．

2. 関連研究

本章では，関連研究として複数楽器 MPE と，DNN において任意性を取り扱うための手法の概要を述べる．

2.1 複数楽器多重音基本周波数推定

自動採譜は長年研究されてきたが，その複雑さから依然として挑戦的な課題である [3]．中でも複数楽器 MPE は特に難しく，単一楽器 MPE と楽器パート推定を各推定音に対して同時に行う必要がある [2]．

複数楽器 MPE はストリームレベルでの採譜問題として一般に研究されており，推定音をグループ化することで各パートの連続的な音高遷移を得る．Duan ら [8] は，MPE の結果に対して制約付きクラスタリングを適用する手法を提案した．彼らの手法はあらゆる MPE アルゴリズム [10], [13], [18] と併用することができ，また各楽器のみで訓練された音源モデルを一切必要としない．この研究に追従して Arora ら [1] は同様のアプローチを取りつつ，MPE と音源依存の特徴量抽出に確率的潜在要素解析を，各楽器パートへのクラスタリングに隠れマルコフ確率場を用いた．以上の二手法は多様な楽器を扱うことができるが，アルゴリズムの特性上，各楽器は単音楽器である必要がある．

近年では，MPE と楽器認識をフレームレベルで同時に行うアプローチによって，複数楽器 MPE に取り組む研究がなされている．Wu ら [19] は DeepLabV3+ [5] と U-Net [17] の構造に基づく DNN モデルを提案し，複数楽器 MPE をスペクトログラム上のセマンティックセグメンテーション問題として捉えた．最新の研究では，ケルベロスネットワークが Manilow ら [15] によって提案されている．以上の手法に共通する弱点は，事前に指定した楽器群のみが採譜対象となっている点である．クラス分類に基づく手法を音源分離に適用するためには，各出力クラスと対象はともに明示的に表されなければならない．そのため，これらの手法を一般の場合に拡張することは困難である．

2.2 深層ニューラルネットワークにおける任意性

音楽音響信号から任意楽器のピアノロールを抽出するためには，楽器自体を特定せずに推定する必要がある．すなわち，各ピアノロールはピアノ，ギター，ヴァイオリンのように楽器が特定された状態ではなく，楽器 1，楽器 2，楽器 3 のように表現されなければならない．しかし，DNN に基づいたアプローチでは順列並べ替えの問題が生じる．特に各楽器パートのピアノロールが推定できても具体的な楽器名が不明であるような場合，直接に損失関数を計算することができない．

同様の問題が任意話者の話者分離において研究されている [12], [20]．対象となる話者の画像や動画がない限り各話者の推定分離音と正解分離音とを対応づけることができないため，同様の順列並べ替え問題が生じる．この問題を解決するため，近年では permutation invariant training(PIT) [20] や深層クラスタリング [12] とよばれる手法が提案されている．

PIT では推定分離音と正解分離音の全ての組み合わせに対する損失関数を計算し，その値が最小となる組み合わせのみでネットワークの最適化を行うことにより，順列並べ替え問題に取り組んでいる．実装が単純であり，同時に他の学習手法と組み合わせることができる一方，計算負荷が極めて高い．具体的には，N 個の音源が音声信号中に含まれている場合，$N!$ 通りの計算が行われる．

これに対し深層クラスタリングでは，出力の潜在表現と最適化によって順列並べ替え問題を回避している．X を時間周波数の要素数，D を潜在次元数とする $X \times D$ の行列 $\boldsymbol{A}$ が出力の潜在表現である時，類似度行列 $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ を計算する．同様に，$X \times N$ の行列 $\boldsymbol{B}$ が正解である時，類似度行列 $\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}$ を計算する．ただし，N は話者数を表す．最適化は二つの類似度行列間の距離を表す $||\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} - \boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}||_F^2$ の値を最小化するように行われる．いかなる $D \times D$ の順列行列 $\boldsymbol{P}$ に対しても $(\boldsymbol{A}\boldsymbol{P})(\boldsymbol{A}\boldsymbol{P})^{\mathrm{T}} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ が成立するため，順列並べ替え問題を回避している．加えて，最適化は変形後である $X \times X$ の行列において行われるため，対象となる音声ないし音響信号は任意数の音源を含むことができる．これらの長所に着目し，本研究では深層クラスタリングを用いる．

3. 提案手法

本章では，クラスタリングに基づく任意楽器パートの採譜手法について述べる．提案手法の枠組みは大きく特徴量

© 2020 Information Processing Society of Japan

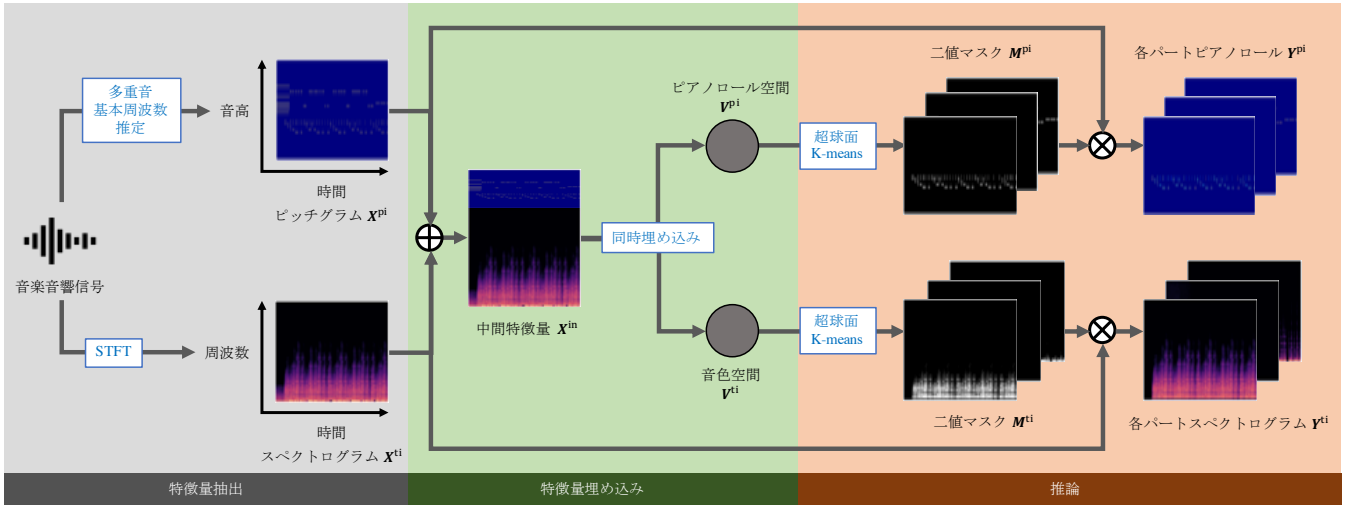

図 2　提案ネットワークの全体図

抽出部，特徴量埋め込み部，推論部の三部分に分かれている．学習初期の安定化のため特徴量抽出部と特徴量埋め込み部を独立に学習し，その後全体を通した学習により全体最適化を行う．

3.1　問題設定

$\boldsymbol{S} = \{\boldsymbol{s}_n \in \mathbb{R}^l\}_{n=1}^N$ を混合音音響信号の集合とする．ただし，$l = 44.1$ [kHz] $\times$ 10 [sec] は音響信号長を，N は混合音音響信号の数を表す．また各 $\boldsymbol{s}_n$ は 3 楽器で構成されていると仮定する．$\boldsymbol{Y}^{\mathrm{pi}} = \{\boldsymbol{y}_n^{\mathrm{pi}} \in [0,1]^{T\times C}\}_{n=1}^{N+1}$ をピッチグラムの集合とする．ただし，T は時間フレーム長を，C は定 Q 変換 (constant-Q transform, CQT) における周波数ビンの数を表す．$\boldsymbol{S}$ から $\boldsymbol{Y}^{\mathrm{pi}}$ への写像を行う DNNf を学習する．ここで，精度向上と学習安定化のために f に対して二種類の工夫を施す．$\boldsymbol{Y}^{\mathrm{ti}} = \{\boldsymbol{y}_n^{\mathrm{ti}} \in \mathbb{R}^{T\times F}\}_{n=1}^{N}$ を各ピアノロールに対応するスペクトログラムとする．ただし，F は短時間フーリエ変換 (short-time Fourier transform, STFT) における周波数ビンの数を表す．f について，$\boldsymbol{S}$ から $\boldsymbol{Y}^{\mathrm{pi}}$ への写像だけでなく，$\boldsymbol{Y}^{\mathrm{ti}}$ への写像も合わせて学習させることにより，採譜精度の向上を図る．f の学習を安定させるため，音高特徴量 $\boldsymbol{X}^{\mathrm{pi}} \in [0,1]^{T\times C}$ と音色特徴量 $\boldsymbol{X}^{\mathrm{ti}} \in \mathbb{R}^{T\times F}$ の二つを中間特徴量として用いる．f を二つのネットワークに分割し，それぞれを $\boldsymbol{S}$ から $\boldsymbol{X}^{\mathrm{pi}}$ への写像を行う特徴量抽出ネットワーク g，$\boldsymbol{X}^{\mathrm{pi}}$ と $\boldsymbol{X}^{\mathrm{ti}}$ からなる中間特徴量から $\boldsymbol{Y}^{\mathrm{pi}}$ と $\boldsymbol{Y}^{\mathrm{ti}}$ への写像を行う特徴量埋め込みネットワーク h とする．学習安定化のため，まず g と h を独立に学習し，その後 f $(= h \circ g)$ の学習によって全体最適化を行う．

3.2　特徴量抽出

特徴量抽出時は，入力の音楽音響信号からピッチグラムとスペクトログラムを得る．これは，各楽器の音高と音色に関する特徴が各楽器パートのピアノロール推定に重要であるためである．音高に関しては，楽器に依存しない音高推定器 [4] を用いて，入力された音楽音響信号に対し対数周波数スペクトログラムと同様の形式のコンデンスピッチグラムを計算した．なお，この推定器は調和定 Q 変換 (harmonic constant-Q transform, HCQT) を入力として，$\boldsymbol{X}^{\mathrm{pi}}$ で表されるコンデンスピッチグラムを出力する．

音色に関しては，信号から STFT を用いて計算された振幅スペクトログラム $\boldsymbol{X}^{\mathrm{ti}}$ を用いた．入力信号の総音量による影響を減らすため，スペクトログラムは各時間周波数ビンが平均 0，分散 1 となるように正規化している．

3.3　特徴量埋め込み

採譜と音源分離の同時学習を行う．これらは相補関係にあり，同時学習によって双方の精度向上につながることが知られている [9], [11]．得られた音高と音色の特徴を同時に学習するため，それぞれの特徴量を周波数軸にそって結合する．この入力特徴量 $\boldsymbol{X}^{\mathrm{in}} \in \mathbb{R}^{T\times(C+F)}$ は，ピアノロール空間 $\boldsymbol{V}^{\mathrm{pi}} \in [-1,1]^{TC\times D}$ と音色空間 $\boldsymbol{V}^{\mathrm{ti}} \in [-1,1]^{TF\times D'}$ へ写像される．**図 3** にその詳細を示す．図中の D と D' はそれぞれピアノロール空間と音色空間の潜在次元数を，H は双方向長短期記憶 (Bidirectional Long short-term memory, BLSTM) の隠れ層の数を表す．なお，L^2 正規化は各空間を D と D' 次元の超球面とするためのものである．

これら二つの潜在空間から二値マスクを作成し，ピッチグラムとスペクトログラムへ適用する．ここで，クラスタリングによってマスクを生成するためには，全ての時間周波数ビンが超球面上において理想的に埋め込まれている必要がある．すなわち，同じ音源に属するビン同士は近距離に，異なる音源に属するビン同士は遠距離になければならない．この理想的な埋め込みは，各空間の類似度行列 $\boldsymbol{V}^{\mathrm{pi,ti}}\boldsymbol{V}^{\mathrm{pi,ti}\,\mathrm{T}}$ を構成することによって実現できる．$\boldsymbol{V}^{\mathrm{pi,ti}}$ は L^2 正規化がなされているため，$TC \times TC$ また

© 2020 Information Processing Society of Japan

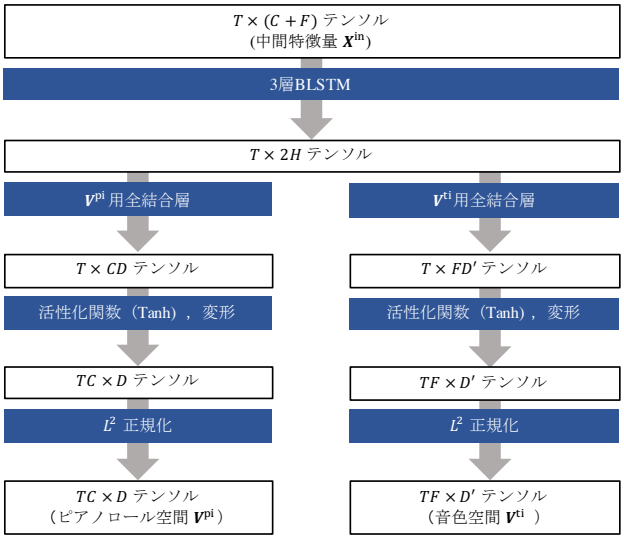

図 3 同時埋め込み部の詳細図

は $TF \times TF$ の行列 $\boldsymbol{V}^{\mathrm{pi,ti}}\boldsymbol{V}^{\mathrm{pi,ti}^{\mathrm{T}}}$ は，全ての時間周波数ビンに関するコサイン距離を示している．N を楽器パート数とし，$TC \times (N+1)$ の行列 $\hat{\boldsymbol{M}}^{\mathrm{pi}}$ と $TF \times N$ の行列 $\hat{\boldsymbol{M}}^{\mathrm{ti}}$ を正解マスクとする．各時間周波数ビンは一つの音源にのみ属すると仮定する．複数の音源が同一ビンを共有している場合には，スペクトログラムにおいて最大の値を持つ音源によって占有されるものとする．このとき $\hat{\boldsymbol{M}}^{\mathrm{pi,ti}}$ は二値を取り，割り当てられたビンの値は 1，割り当てられていないビンの値は 0 となる．これに伴い，類似度行列 $\hat{\boldsymbol{M}}^{\mathrm{pi,ti}}\hat{\boldsymbol{M}}^{\mathrm{pi,ti}^{\mathrm{T}}}$ もまた二値を取る．ネットワークの学習にあたっては，$\hat{\boldsymbol{M}}^{\mathrm{pi,ti}}\hat{\boldsymbol{M}}^{\mathrm{pi,ti}^{\mathrm{T}}}$ を $\boldsymbol{V}^{\mathrm{pi,ti}}\boldsymbol{V}^{\mathrm{pi,ti}^{\mathrm{T}}}$ の正解として用いることで，教師あり学習を行うことができる．

なお，$\hat{\boldsymbol{M}}^{\mathrm{pi}}$ において次元を一つ追加している．スペクトログラム $\boldsymbol{X}^{\mathrm{ti}}$ と異なりコンデンスピッチグラム $\boldsymbol{X}^{\mathrm{pi}}$ は推定値であるため，誤推定を含む．特に偽陽性に関しては正解となる帰属先の楽器が存在せず，例外として扱う必要がある．そのため，正解ピッチグラムにおける無音のビンの帰属先として新たに次元を設け，真陰性と偽陽性のビンを集めている．また，微小な値を持つビンにより学習の妨げを防ぐため，$\boldsymbol{X}^{\mathrm{ti}}$ のビンに関しては閾値以上のビンのみを考慮し，その他のビンについては全音源で共有する．

3.4 学習方法

音高推定器の学習は，式 (1) で表される損失関数の値の最小化によって行う．

$$\mathcal{L}_{DS} = -\hat{\boldsymbol{X}}^{\mathrm{pi}} \log(\boldsymbol{X}^{\mathrm{pi}}) - (1-\hat{\boldsymbol{X}}^{\mathrm{pi}}) \log(1-\boldsymbol{X}^{\mathrm{pi}}) \quad (1)$$

ここで，$\hat{\boldsymbol{X}}^{\mathrm{pi}}$ と $\boldsymbol{X}^{\mathrm{pi}}$ はそれぞれ正解コンデンスピッチグラムと推定コンデンスピッチグラムを表し，ともに 0 から 1 までの連続値をとる．同時埋め込み部の学習は，式 (2) で表される損失関数の値の最小化によって行う．なお，式中の $||\cdot||_F^2$ はフロベニウスノルムの平方である．

$$\mathcal{L}_{DC}{}^{p,t} = ||\boldsymbol{V}^{p,t}\boldsymbol{V}^{p,t^{\mathrm{T}}} - \hat{\boldsymbol{M}}^{\mathrm{pi,ti}}\hat{\boldsymbol{M}}^{\mathrm{pi,ti}^{\mathrm{T}}}||_F^2 \quad (2)$$

計算量を減らすため，実際には式 (2) の展開形を用いた．

$$\mathcal{L}_{DC}{}^{p,t} = ||\boldsymbol{V}^{p,t^{\mathrm{T}}}\boldsymbol{V}^{p,t}||_F^2 - 2||\boldsymbol{V}^{p,t^{\mathrm{T}}}\hat{\boldsymbol{M}}^{\mathrm{pi,ti}}||_F^2 + ||\hat{\boldsymbol{M}}^{\mathrm{pi,ti}^{\mathrm{T}}}\hat{\boldsymbol{M}}^{\mathrm{pi,ti}}||_F^2 \quad (3)$$

TC と TF の値は D と D' の値よりも極めて大きいため，式 (3) により類似度行列の直接計算を回避している．これら二種類の損失関数を用いて，全体の損失関数は式 (4) のように表される．α と β は各損失の重みパラメータである．

$$\mathcal{L}_{total} = \mathcal{L}_{DS} + \alpha\mathcal{L}_{DC}{}^{\mathrm{pi}} + \beta\mathcal{L}_{DC}{}^{\mathrm{ti}} \quad (4)$$

学習初期の安定化のため，まず音高推定器と同時埋め込みネットワークを式 (1) と式 (3) によってそれぞれ学習させた．その後，式 (4) によってネットワーク全体の最適化を行った．なお，全ての学習には Adam [14] を用いた．

3.5 推論

推論時は，学習済み潜在空間 $\boldsymbol{V}^{\mathrm{pi}}$，$\boldsymbol{V}^{\mathrm{ti}}$ から $\boldsymbol{X}^{\mathrm{pi}}$ と $\boldsymbol{X}^{\mathrm{ti}}$ にそれぞれに対する二値マスク $\{\boldsymbol{M}_i^{\mathrm{pi}}\}_{i=1,\ldots,N+1}$ と $\{\boldsymbol{M}_j^{\mathrm{ti}}\}_{j=1,\ldots,N}$ を生成する．これは埋め込み特徴量のクラスタリングによって行われる．二つの潜在空間はともに超球面状であるため，超球面上の距離（コサイン距離）に基づく超球面クラスタリング [7] を適用する．各楽器パートおよび無音パートのピアノロール $\{\boldsymbol{Y}_i^{\mathrm{pi}}\}_{i=1,\ldots,N+1}$ は，式 (5) によって求められる．

$$\boldsymbol{Y}_i^{\mathrm{pi}} = \boldsymbol{X}^{\mathrm{pi}} \otimes \boldsymbol{M}_i^{\mathrm{pi}} \quad (5)$$

加えて，各楽器パートのスペクトログラム $\{\boldsymbol{Y}_j^{\mathrm{ti}}\}_{j=1,\ldots,N}$ が式 (6) により得られる．これは，逆短時間フーリエ変換によって各楽器パートの分離音へと変換することができる．

$$\boldsymbol{Y}_j^{\mathrm{ti}} = \boldsymbol{X}^{\mathrm{ti}} \otimes \boldsymbol{M}_j^{\mathrm{ti}} \quad (6)$$

式 (5) および式 (6) において，$\otimes$ は行列の要素積を表す．

4. 評価実験

4.1 使用データ

評価実験には Slakh2100-orig [16] を使用した．このデータセットは訓練データ 1500 曲，検証データ 375 曲，評価データ 225 曲を含む．各楽曲は複数楽器で演奏されており，データセットは MIDI データを伴う混合音および分離音で構成されている．含まれる楽器は 12 種類（ピアノ，ベース，ギター，ドラム，弦楽器，シンセパッド，リード，金管楽器，オルガン，パイプ，シンセリード，クロマティックパーカッション）であるが，このうち採譜時に音高情報が重要でないドラムとクロマティックパーカッションは除外した．任意楽器に対する採譜精度を確かめるため，残る 10 楽器のうちオルガン，パイプ，シンセリードを除く 7 楽

© 2020 Information Processing Society of Japan

器のみを用いて学習を行った．評価時には，学習済みの 7 楽器のみを含む閉条件と全楽器を含む開条件を用意し，それぞれ評価を行った．

学習データは各楽曲を 10 秒ごとに切り出した上で，正解コンデンスピッチグラムは構成楽器の MIDI データを，混合音は各楽器の演奏音を，それぞれ重ね合わせることで用意した．混合された MIDI データは二値化した後, [4] に倣いガウシアンフィルタを適用した．楽曲の演奏音はモノラル，サンプリングレートは 44.1kHz である．STFT の窓幅は 2048（約 50 [ms]），シフト幅は STFT，HCQT ともに 512（約 11 [ms]）とした．HCQT は 32.7Hz（C1）から 6 オクターブにかけて，5 倍音まで計算を行った．全体のデータ量は，訓練データ 11 時間，検証データ 3 時間，評価データ 6 時間（開閉両条件とも）となった．

4.2 実験条件

混合音中の各楽器パートに対する，フレームレベルでの採譜精度を評価した．実験に際し，楽器数は 3 楽器に固定した．評価尺度は評価尺度は式 (7) で表される適合率 (P) と再現率 (R)，F 値 (F) を用いた．ただし，TP と FP，FN は，それぞれ真陽性，偽陽性，偽陰性のビンの数である．

$$P = \frac{TP}{TP+FP},\ R = \frac{TP}{TP+FN},\ F = \frac{2PR}{P+R} \qquad (7)$$

ある特定の楽器のピッチグラムビンにおいて，二値で表される推定結果と正解とが一致し，かつ正しくパート割り当てが行われた時に限り，推定が正しく行われたものとした．

既存のクラス分類手法との比較を行うため，[19] を 8 クラス（上記 7 種類の既知楽器と非楽器クラス）の出力に変えて追実装した．パート推定について提案手法のクラスタリングアプローチと，既存のクラス分類アプローチとの公平な比較評価のため，実験条件を以下のように設定した．

- クラスタリングアプローチでは，全体の F 値が最高となるように各クラスタを各楽器に割り当てる．
- 閉条件でのクラス分類アプローチでは，パート推定結果を用いて直接各楽器への割り当てを行う．
- 開条件でのクラス分類アプローチでは，全体の F 値が最高となるように各クラスを各楽器に再度割り当てる．

4.3 実験結果

表 1 に実験結果を示す．開条件下での未知楽器の採譜精度において，提案手法はクラス分類に基づく手法 [19] を上回った．また，既存手法では未知楽器の採譜精度が既知楽器の採譜精度よりも大幅に減少したのに対し，提案手法においては両者が比肩する結果となった．さらに，提案手法は閉条件か開条件かを問わず，既知楽器の採譜においても既存手法と同等の精度であった．

図 4 の提案手法による採譜結果例から，提案手法が音高推定と楽器パート割り当てに成功していることが分かる．その一方失敗例では，3 秒付近においてピアノロール 2 に含まれるべき音がピアノロール 1 に含まれているほか，多くの誤推定が行われている．

5. おわりに

本稿では，深層球面クラスタリングに基づく任意の複数楽器パートに対する採譜手法について述べた．提案手法ではピッチグラムとスペクトログラムの同時クラスタリングを通して，音楽音響信号中の音色と音高の特徴が同時に考慮された採譜が行われるとともに，実験により訓練データに含まれない楽器を含む楽曲のパート譜採譜が可能であることを確かめた．

提案手法では各楽器パートのピアノロールに加えて分離音も得ることができるが，両者間の対応づけは手動で行う必要がある．この自動化を可能にするアルゴリズムを考案し，より統一的に音高情報と音色情報を考慮することが，最も興味深い今後の方針として考えられる．

謝辞 本研究の一部は，JST ACCEL No. JPMJAC1602，JSPS 科研費 No. JP16H01744 および JP19H04137 の支援を受けた．

参考文献

[1] Arora, V. and Behera, L.: Multiple F0 Estimation and Source Clustering of Polyphonic Music Audio Using PLCA and HMRFs, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 23, No. 2, pp. 278–287 (2015).

[2] Benetos, E., Dixon, S., Duan, Z. and Ewert, S.: Automatic Music Transcription: An Overview, *IEEE Signal Processing Magazine*, Vol. 36, No. 1, pp. 20–30 (2019).

[3] Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H. and Klapuri, A.: Automatic Music Transcription: Challenges and Future Directions, *Journal of Intelligent Information Systems*, Vol. 41 (2013).

[4] Bittner, R. M., McFee, B., Salamon, J., Li, P. and Bello, J. P.: Deep Salience Representations for F_0 Estimation in Polyphonic Music, *18th Int. Soc. for Music Info. Retrieval Conf.*, pp. 63–70 (2017).

[5] Chen, L. C., Papandreou, G., Schroff, F. and Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation, *eprint arXiv:1706.05587* (2017).

[6] Christensen, M. G., Stoica, P., Jakobsson, A. and Jensen, S. H.: Multi-pitch estimation, *Signal Processing*, Vol. 88, No. 4, pp. 972–983 (2008).

[7] Dhillon, I. S., Fan, J. and Guan, Y.: Efficient Clustering of Very Large Document Collections, *Data Mining for Scientific and Engineering Applications*, Vol. 2 (2001).

[8] Duan, Z., Han, J. and Pardo, B.: Multi-pitch Streaming of Harmonic Sound Mixtures, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 22, No. 1, pp. 138–150 (2014).

[9] Duan, Z. and Pardo, B.: Soundprism: An Online System for Score-Informed Source Separation of Music Audio, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 5, No. 6, pp. 1205–1215 (2011).

[10] Duan, Z., Pardo, B. and C.Zhang: Multiple Fundamental Frequency Estimation by Modeling Spectral Peaks

© 2020 Information Processing Society of Japan

表 1　クラス分類アプローチと提案手法による実験結果

	閉条件						開条件					
	[19]			提案手法			[19]			提案手法		
楽器	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
ピアノ	51.28	46.50	**45.87**	62.02	39.61	44.07	52.51	48.04	**47.37**	61.87	38.90	43.64
ベース	73.75	58.79	**64.04**	39.72	50.78	42.24	74.27	59.66	**64.67**	40.59	51.88	43.23
ギター	46.64	36.72	37.69	52.91	35.45	**39.46**	44.59	37.12	37.25	53.45	36.50	**40.32**
弦楽器	55.27	56.79	**52.74**	66.35	48.74	52.40	53.21	56.97	**52.05**	65.31	48.40	52.04
シンセパッド	43.72	44.80	**42.07**	49.65	35.12	38.70	44.42	46.89	**43.91**	51.99	36.58	40.81
リード	28.53	33.90	29.27	29.87	37.37	**31.53**	26.92	31.72	27.53	28.87	35.46	**30.04**
金管楽器	35.24	25.12	24.50	37.10	30.23	**29.53**	37.66	25.67	25.89	36.78	30.64	**30.26**
オルガン	—	—	—	—	—	—	20.14	19.01	16.89	36.62	28.57	**29.11**
パイプ	—	—	—	—	—	—	22.62	27.13	23.02	38.37	39.49	**35.22**
シンセリード	—	—	—	—	—	—	20.58	17.44	17.59	29.41	25.11	**24.98**

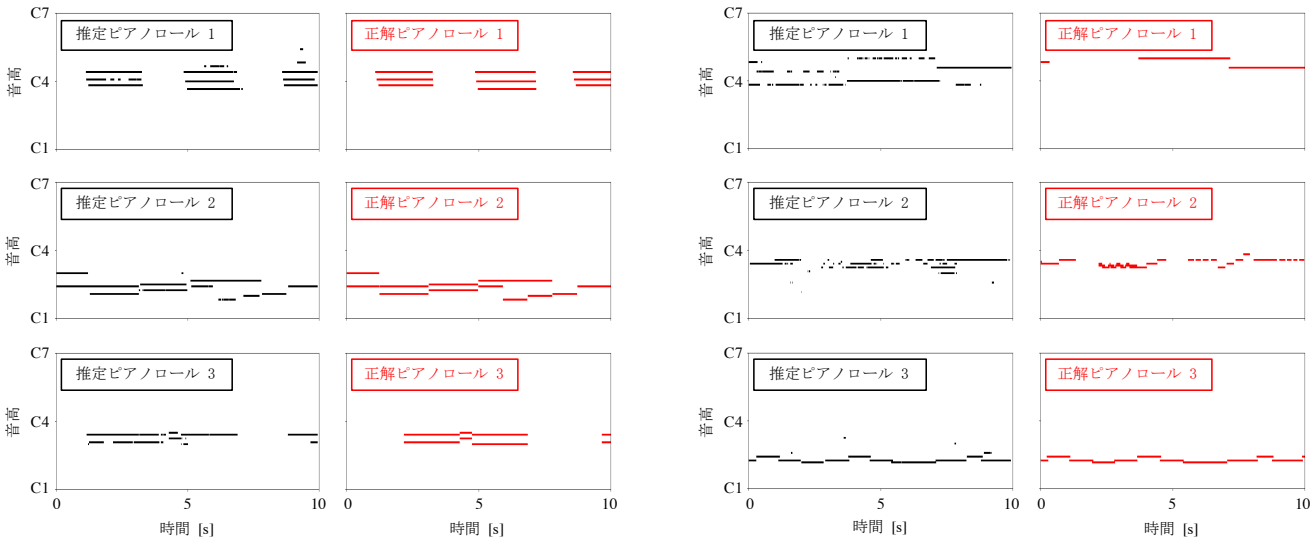

図 4　提案手法によるパート譜採譜結果の成功例（左，Track01879）と失敗例（右，Track01878）

and Non-peak Regions, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 18, No. 8, pp. 2121–2133 (2010).

[11] Ewert, S., Pardo, B., Muller, M. and Plumbley, M. D.: Score-Informed Source Separation for Musical Audio Recordings: An overview, *IEEE Signal Processing Magazine*, Vol. 31, No. 3, pp. 116–124 (2014).

[12] Hershey, J. R., Chen, Z., Le Roux, J. and Watanabe, S.: Deep clustering: Discriminative embeddings for segmentation and separation, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35 (2016).

[13] Jin, Z. and Wang, D.: HMM-based Multipitch Tracking for Noisy and Reverberant Speech, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 5, pp. 1091–1102 (201).

[14] Kingma, D. and Ba, J.: Adam: A Method for Stochastic Optimization, *International Conference on Learning Representations* (2014).

[15] Manilow, E., Seetharaman, P. and Pardo, B.: Simultaneous Separation and Transcription of Mixtures with Multiple Polyphonic and Percussive Instruments, *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 771–775 (2020).

[16] Manilow, E., Wichern, G., Seetharaman, P. and Le Roux, J.: Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity, *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (2019).

[17] Ronneberger, O., Fischer, P. and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, *Medical Image Computing and Computer-Assisted Intervention*, Vol. 9351, pp. 234–241 (2015).

[18] Wu, M., Wang, D. and Brown, G. J.: A Multipitch Tracking Algorithm for Noisy Speech, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 11, No. 3, pp. 229–241 (2003).

[19] Wu, Y., Chen, B. and Su, L.: Polyphonic Music Transcription with Semantic Segmentation, *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 166–170 (2019).

[20] Yu, D., Kolbæk, M., Tan, Z. and Jensen, J.: Permutation Invariant Training of Deep Models for Speaker-independent Multi-talker Speech Separation, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245 (2017).

© 2020 Information Processing Society of Japan