5S-03

深層クラスタリングを用いた任意楽器パートの自動採譜

田中 啓太郎 † 中塚 貴之 † 錦見 亮 ‡ 吉井 和佳 ‡ 森島 繁生 ††
† 早稲田大学 ‡ 京都大学 大学院情報学研究科 †† 早稲田大学理工学術院総合研究所

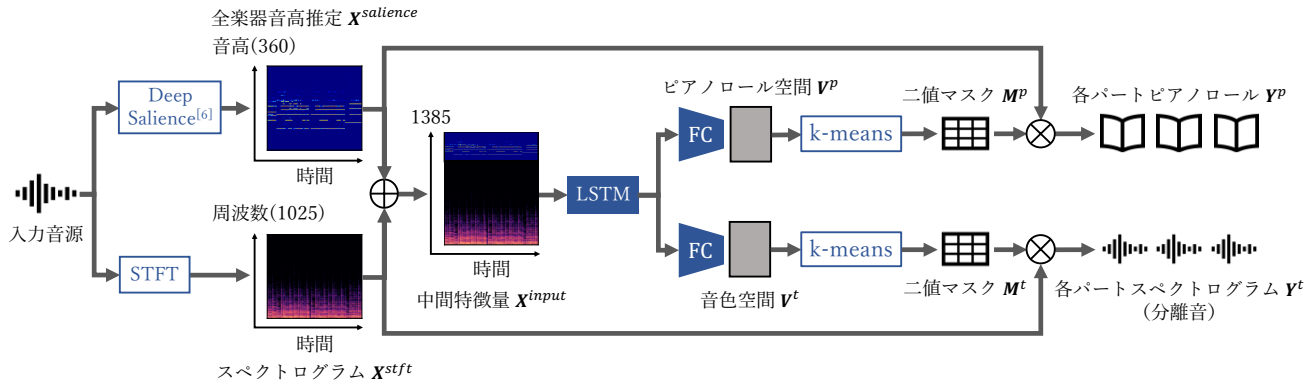

図 1: 複数の楽器パートの自動採譜と音源分離を同時に行うニューラルネットワーク

1. はじめに

本稿では，音楽音響信号に対して各時刻の音高を推定する問題を扱う．採譜の中でも特に，複数楽器の音が含まれる音楽音響信号に対し，各楽器パートについて採譜を行うことをパート譜採譜という．パート譜採譜は音楽情報の詳細な記録に必要不可欠な上に，実際の演奏時は通常パート譜を使用して演奏するため音楽においてパート譜採譜は重要である．人手でのパート譜採譜は多くの時間と労力を必要とするため，自動でのパート譜採譜の実現が求められている．

採譜を計算機で自動的に行う自動採譜は音楽情報処理の分野において盛んに研究されている．しかしながら，自動パート譜採譜は音高推定に加えて推定した各音が帰属する楽器パートを推定する必要があり，依然として挑戦的な課題である [1]．既存研究の多くは音楽音響信号に対し音高推定を行い，その推定結果を基にパート譜採譜を行う多段階処理のアプローチを取っている [2] [3]．このアプローチではそれぞれの処理における誤差が蓄積し，累積誤差が大きくなる問題がある．この問題は，End-to-end のアプローチによって推論モデル全体をまとめて最適化することで推定精度の改善が期待できる．

近年，深層学習を用いた End-to-end のアプローチで自動パート譜採譜を行う手法が提案されている．Wu らはセマンティックセグメンテーションを用いることで，各時間周波数ビンの多クラス分類による自動パート譜採譜を提案した [4]．しかし，Wu らの手法では採譜対象の楽器を学習時に指定し，指定した各楽器についての教師データ（正解音高データ）を用意する必要がある．現代では多様な楽器が作曲に使用されるため，入力の音楽音響信号に対応した教師あり楽器データを都度用意することは現実的ではない．また，EDM やポップスといった現代の音楽では，イコライザやエンハンサといった音楽音響信号の編集ツールによって，楽器音や環境音を独自に加工した音が作曲に使用される．このような理由から，パート譜採譜の需要が高い現代の音楽に対して予め推定対象の楽器を指定する手法は適切でない．

本研究では，深層クラスタリングによって任意楽器に対して自動パート譜採譜を行う枠組みを提案する．教師なし学習であるクラスタリングを用いることで，学習データに依存しないパート譜採譜を実現する．評価実験によって本手法のパート譜採譜に対する有用性を確認する．

2. 提案手法

2.1 問題設定

本稿では，(1) 任意の複数楽器音が含まれる音楽音響信号に対して自動パート譜採譜を実現し，(2)End-to-end のアプローチが多段階処理のアプローチに対して有意であることを確認する．ベースライン手法として，入力音源を楽器パートごとに音源分離し [5]，それぞれの分離音に対して多重音基本周波数推定 [6] を行う多段階処理のアプローチ（以下カスケードモデル）を採用する．提案手法では End-to-end のアプローチとして，多重音基本周波数推定結果の各時間周波数ビンに対する楽器パートごとの二値マスク推定を考える．これは，任意話者分離技術 [5] における各時間周波数ビンに対するマスク推定に着想を得たものである．

2.2 提案モデル

図 1 に提案手法の概要図を示す．まず，入力音源に対して多重音基本周波数推定 [6] を行い，各時刻における全楽器の音高を推定した．推定結果は，対数周波数スペクトログラム $\boldsymbol{X}^{salience} = \{\boldsymbol{x}_1^{salience}, \ldots, \boldsymbol{x}_T^{salience}\} \in [0,1]^{T\times C}$ として得られる．ここで T は時間，C は定 Q 変換（Constant-Q Transform; CQT）における周波数である．本稿では，初期における学習の安定化のため音高推定部は独立に学習した．一方で，入力音源に対し短時間フーリエ変換（Short-Time Fourier Transform; STFT）することで周波数スペクトログラム $\boldsymbol{X}^{stft} = \{\boldsymbol{x}_1^{stft}, \ldots, \boldsymbol{x}_T^{stft}\} \in \mathbb{R}^{T\times F}$ を得た．ここで F は STFT における周波数である．2 つの周波数スペクトログラムを結合した特徴量を，中間特徴量 $\boldsymbol{X}^{input} = \{\boldsymbol{x}_1^{input}, \ldots, \boldsymbol{x}_T^{input}\} \in \mathbb{R}^{T\times(C+F)}$ と呼ぶ．本稿では中間特徴量 $\boldsymbol{X}^{input}$ から 2 つの潜在空

Automatic Transcription of Arbitrary Musical Instrument Parts Based on Deep Clustering: *Keitaro Tanaka*†, *Takayuki Nakatsuka*†, *Ryo Nishikimi*‡, *Kazuyoshi Yoshii*‡, *and Shigeo Morishima*†† (†*Waseda University*, ‡*Kyoto University*, ††*Waseda Research Institute for Science and Engineering*)

間 $\boldsymbol{V}^p \in \mathbb{R}^{TF\times D}$, $\boldsymbol{V}^t \in \mathbb{R}^{TC\times D'}$ へのマッピングを学習するために，任意話者分離技術 [5] に基づく推論モデルを構成した．ここで $\boldsymbol{V}^p, \boldsymbol{V}^t$ はそれぞれピアノロール空間，音色空間であり，D, D' は各空間の特徴量次元数である．それぞれの空間で k-means を行うことで，2 つの周波数スペクトログラム $\boldsymbol{X}^{salience}$, $\boldsymbol{X}^{stft}$ それぞれに対応する楽器パートごとの二値マスク $\boldsymbol{M}^p \in \{0,1\}^{T\times C\times N}$, $\boldsymbol{M}^t \in \{0,1\}^{T\times F\times N}$ が得られる．ここで N は楽器パート数である．

パート譜 $\{\boldsymbol{Y}_i^p\}_{i=1,\ldots,N}$ は，対数周波数スペクトログラム $\boldsymbol{X}^{salience}$ と楽器パートごとの二値マスク $\boldsymbol{M}_i^p$ を用いて，次式で計算される．

$$\boldsymbol{Y}_i^p = \boldsymbol{X}^{salience} \otimes \boldsymbol{M}_i^p \tag{1}$$

ここで $\otimes$ は行列同士の要素積である．また，分離音は次式で計算される楽器ごとの周波数スペクトログラム $\{\boldsymbol{Y}_i^t\}_{i=1,\ldots,N}$

$$\boldsymbol{Y}_i^t = \boldsymbol{X}^{stft} \otimes \boldsymbol{M}_i^t \tag{2}$$

に対し逆フーリエ変換することで得られる．

学習には以下の損失関数 $\mathcal{L}^{total}$ を用いた．

$$\mathcal{L}^{total} = \mathcal{L}^p + \mathcal{L}^t \tag{3}$$

$$\mathcal{L}^{p,t} = ||\boldsymbol{V}^{p,t^T}\boldsymbol{V}^{p,t} - \boldsymbol{Y}^{p,t^T}\boldsymbol{Y}^{p,t}||_F^2 \tag{4}$$

3. 評価実験

3.1 実験条件

楽曲の楽器パート数を $N = 3$ に固定し，Slakh2100 [7] から，174 曲をモデルの訓練，43 曲を検証に用いた．データセットの各楽曲を構成する各パートからランダムに選んだ 3 つのパートを合成してデータを作成した．ただし，打楽器や効果音などの採譜に音高情報が不要な楽器パートは除外した．データには単音楽器だけではなく，一度に複数音を出す楽器も含まれる．楽曲のサンプリング周波数は 44.1kHz である．CQT および STFT のシフト幅をともに 512，STFT の窓幅を 2048 とした．

提案法のパート譜採譜精度をカスケードモデルと比較評価した．評価尺度は推定ピアノロールの正解ピアノロールに対する F 値を用いた．F 値は推定結果の正解に対する再現率 R_{ec} と適合率 P_{rec} の調和平均であり，以下の式で計算される．

$$F_{score} = \frac{2R_{ec}P_{rec}}{R_{ec} + P_{rec}} \tag{5}$$

再現率 R_{ec} と適合率 P_{rec} はそれぞれ，真陽性 TP，偽陰性 FN，偽陽性 FP を用いて $R_{ec} = TP/(FN+TP)$, $P_{rec} = TP/(FP+TP)$ と表される．Slakh2100 [7] から，訓練と検証に未使用の 50 曲を評価に用いた．

3.2 実験結果

結果を図 2 に示す．丸印は外れ値を表し，箱ひげ図は外れ値を除くデータの下位 25%値，中央値，上位 25%値を表す．カスケードモデルと提案手法それぞれの F 値の平均値は 0.423 と 0.571 であり，この数値からも提案手法の有効性が確認できる．同一楽曲の同一箇所（約 30 秒分）に対する楽器パートごと（(a)〜(c)）の採譜結果例を図 3 に示す．上 3 つが推定結果，下 3 つが正解を表す．パート譜採譜は大体できており，特に (a) は採譜結果と正解がほぼ完全に一致している．一方，10 秒以降において (c) に採譜されるべき音が (b) に採譜されているように，誤った推定が行われていることも見て取れる．また，(b) と (c) は C2 から C3 にかけて同時刻の共有音がいくつかある．しかしそれらの音は (b) と (c) いずれかにのみ採譜されている．これは，提案手法では各時間周波数ビンが一つのパートにのみ割り当てられ，複数パート間で共有できないことが原因である．

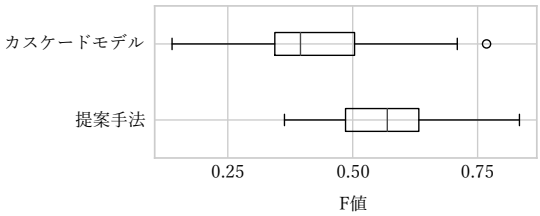

図 2: カスケードモデルとの採譜精度比較

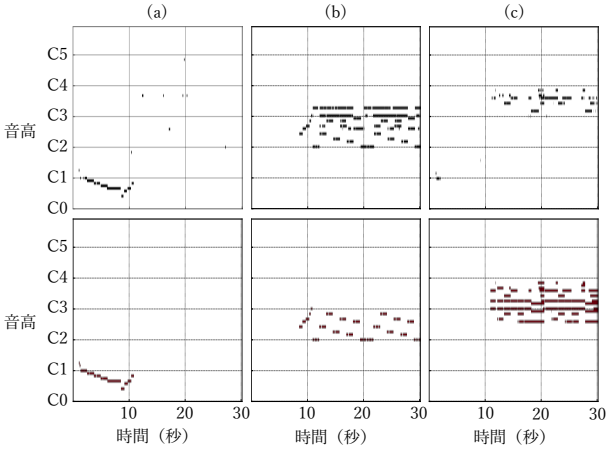

図 3: 提案手法によるパート譜採譜結果の例

4. おわりに

本稿では，深層クラスタリングを用いた任意楽器に対する自動パート譜採譜を提案した．評価実験の結果，入力音源に対するパート譜採譜の精度がカスケードモデルよりも上昇し，本手法の有効性が確認された．今後は，音高推定部も含めたネットワーク全体の最適化に取り組むほか，未知の楽器に対する有効性を他手法との比較により行う．さらに，現在固定している楽器数を自動で予測可能な枠組みへの拡張を目指す．

謝辞 本研究は，JST ACCEL (JPMJAC1602) および JSPS 科研費 (JP19H04137) の補助を受けた．

参考文献

[1] E. Benetos *et al.*: “Automatic Music Transcription: An Overview,” *IEEE Signal Processing Magazine*, 20–30, 2019.

[2] Z. Duan *et al.*: “Multi-pitch Streaming of Harmonic Sound Mixtures,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 130–150, 2014.

[3] V. Arora *et al.*: “Multiple F0 Estimation and Source Clustering of Polyphonic Music Audio Using PLCA and HMRFs,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 278–287, 2015.

[4] Y. Wu *et al.*: “Polyphonic Music Transcription with Semantic Segmentation,” *ICASSP*, 166–170, 2019.

[5] J.R. Hershey *et al.*: “Deep Clustering: Discriminative Embeddings for Segmentation and Separation,” *ICASSP*, 31–35, 2016.

[6] R.M. Bittner *et al.*: “Deep Salience Representations for F_0 Estimation in Polyphonic Music,” *ISMIR*, 63–70, 2017.

[7] E. Manilow *et al.*: “Cutting Music Source Separation Some Slakh: A Dataset to Study the Impact of Training Data Quality and Quantity,” *WASPAA*, 45–49, 2019.