時間周波数表現を用いた畳み込みニューラルネットワークに基づく演奏楽器推定 *

☆城間 佑樹 (都立大)，水野 賀文，高橋 祐 (ヤマハ)，
塩田 さやか (都立大)，近藤 多伸 (ヤマハ)，貴家 仁志 (都立大)

1　はじめに

演奏楽器推定は音楽情報検索の分野で活発に研究されている分類タスクの一つである [1–3]．特に近年では，音楽からの情報抽出のみならず，演奏音を自動で判別し適切なエフェクトを適用するなど様々なアプリケーションでの活用が期待されている．演奏音を分類するタスクは多数存在するが，入力音の観点からは演奏音が楽器単体のものか複数の楽器が混ざったものかという点で分けられる．さらに楽器単体の演奏音を扱う研究の中でも，演奏楽器がギターやフルートなどの調波楽器か，シンバルやドラムなどの非調波楽器かという点で分けられる場合もある [4, 5]．

演奏楽器推定を扱う先行研究では，入力音の短時間フーリエ変換 (Short-Time Fourier Transform; STFT) や定 Q 変換 (Constant Q Transform; CQT) などから得られた時間周波数表現を特徴量として用いており，メルスペクトログラムを用いた手法 [6] や CQT スペクトログラムを用いた手法 [7]，メル周波数ケプストラム係数 (Mel-Frequency Cepstral Coefficient; MFCC) を用いた手法 [8] などが提案されている．これらの特徴量を入力する分類器としては Support Vector Machine (SVM) を用いた手法 [9] や畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) を用いた手法 [10] などが提案されている．しかしこれらの既存手法は，時間周波数表現に対する適切な特徴量抽出の検討や表現力の高い分類器を用いた検討がされていない．

画像分類タスクでは，入力データに対して Local Binary Pattern (LBP) [11] や Histograms of Oriented Gradient (HOG) [11] といった画像特徴抽出法を適用し，画像内の特徴を強調することで性能向上することが知られている [12]．楽器音分類において，スペクトログラムのような時間周波数表現を直接入力特徴量として扱う場合，浮動小数点数で表される 2 次元配列の 1 チャネル画像とみなして扱うことが多い．しかし，従来の楽器音分類では時間周波数表現を直接入力するのみで，画像特徴抽出法による強調手法の活用などは検討されていなかった．そこで本研究では，入力される楽器音の時間周波数表現に対して画像特徴抽出法を用い，演奏楽器の局所的な特徴を強調することで，楽器音分類の性能向上を目指す．そのために，時間周波数表現に対して，画像特徴抽出法を適用した特徴量を分類器に入力して演奏楽器の推定を行う．また，特徴抽出の効果的な組み合わせについても調査する．実験では，ConvMixer [13] という画像分類において高い性能を持つ分類器を用いて，時間周波数表現と画像特徴抽出を組み合わせた演奏楽器分類を行った．実験結果より，従来法と比較して，画像特徴抽出法を含むアンサンブル手法において調波楽器および非調波楽器それぞれのエラー改善率が 89.3%，82.8%となったことを報告する．

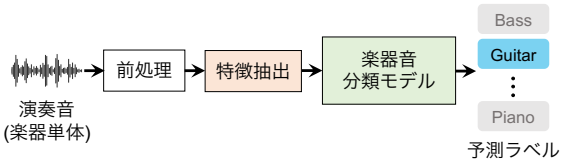

図 1　演奏楽器分類のフロー

2　演奏楽器分類

図 1 に演奏楽器分類のフローを示す．演奏楽器分類ではまず，入力音に正規化や無音除去などの前処理を適用した後，特徴抽出を行う．得られた特徴量を楽器音分類モデルに入力し，入力音の演奏楽器を推定するという流れになっている．先行研究では，特徴量や分類器を工夫することで分類精度を高めようとする手法が提案されている．例えば，メルスペクトログラムを CNN で分類する手法 [1] では，CNN の各畳み込み層の後ろにプーリング層を挿入することで，最終層にプーリング層を設けた従来法よりも高い性能となっている．しかし，非調波楽器では性能評価がなされていない．文献 [14] では入力音の音声波形を特徴抽出せず CNN に直接入力する End-to-End な枠組みで分類することで，手動で特徴抽出を行い決定木などで演奏楽器推定をする従来法よりも高い性能を達成しているが，正解率が 82%程度と分類性能が十分ではない．これらは入力音に複数の楽器音が混ざったものを扱っているが，一方で本研究で焦点としている楽器単体の演奏音の分類を行うタスクでは，ウェーブレット変換で計算したスカログラムを CNN に入力する手法 [10] がある．ウェーブレット変換を

---

* Convolutional Neural Network-Based Musical Instrument Identification Using Time-Frequency Representations. by SHIROMA Yuki (Tokyo Metropolitan University), MIZUNO Yoshifumi, TAKAHASHI Yu (Yamaha Corporation), SHIOTA Sayaka (Tokyo Metropolitan University), KONDO Kazunobu (Yamaha Corporation), and KIYA, Hitoshi (Tokyo Metropolitan University)

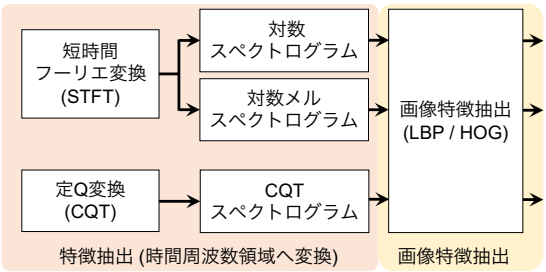

図 2　提案法における特徴抽出のフロー

用いることで FFT の窓長の設定によって生じる分類性能の変動に頑健となり，スペクトログラムを用いた従来法よりも高い性能を得られることが報告されている．しかし評価データに含まれているのは調波楽器のみであり，非調波楽器では性能評価がなされておらず汎化性能について課題がある．

3　提案法

提案法では演奏楽器分類における時間周波数表現を用いた適切な特徴量の検討，及び畳み込み層を含む深層学習に基づく分類器を用いた画像表現によるモデル化について検討する．図 2 に提案法における特徴抽出のフローを示す．提案法では前段で入力音を時間周波数表現へ変換し，後段で画像特徴抽出法を用いて局所的な特徴を抽出するものとなっている．

3.1　特徴抽出

特徴抽出ではまず，STFT または CQT を入力音に適用して時間周波数表現であるスペクトログラムまたは CQT スペクトログラムを得る．STFT は信号を一定時間ずつ分割しフーリエ変換することで時間周波数領域へ変換する手法である．スペクトログラムには絶対値と対数をかけた対数スペクトログラムを計算する．さらに，メルフィルタバンクを用いて対数メルスペクトログラムを求めることで次元圧縮と低域強調を行う場合も検討する．一方，CQT は楽器音や音楽の分析で多く用いられる時間周波数分析手法で音階に合わせて時間周波数領域へ変換する手法である．時間周波数領域で正規化を行っている．図 3 に各時間周波数表現の抽出例を示す。図 3(A) の対数スペクトログラムと (B) の対数メルスペクトログラムに着目すると，調波楽器の対数スペクトログラムには低域に調波構造が確認でき，低域強調を行うことで拡大されている様子が確認できる．さらに，スペクトログラムでは周波数の低域に音があることを確認できるが，(C) の CQT スペクトログラムでは表示している 7 オクターブの音階うち中央の音階に音があることがわかる．

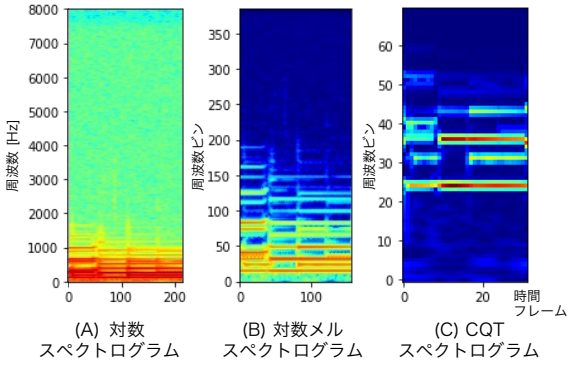

図 3　調波楽器音 (Keyboard) の時間周波数表現

3.2　画像特徴抽出

画像分類タスクでは，画像の特徴を強調するため，また頑健な特徴量を抽出するために HOG や LBP などの画像特徴抽出法が広く用いられている．そこで画像として扱われる時間周波数表現においても局所的な特徴を抽出するために画像特徴抽出法を適用する．画像特徴量としての HOG は影などによる明るさの変化に頑健であり，LBP は画像中の局所的な特徴を抽出することで特徴量の頑健性向上に貢献することが知られている．提案法では，HOG は時間周波数表現にエッジとして表れる調波楽器の調波構造や非調波楽器の打鍵音を頑健に捉えることを期待しており，LBP は音量の乱れや音の開始時刻の乱れに対して頑健に音響特徴を捉えることを期待している．

3.3　畳み込みニューラルネットワークの学習

提案法では時間周波数表現を入力データとするため，2 次元配列データを入力特徴量として想定している CNN を基としたネットワークへの入力を想定している．その中でも，近年の画像分類タスクにおいて高い識別性能を示している ConvMixer を分類器として用いる．ConvMixer は 2 次元データをパッチに分割した後，畳み込み層に通過させるものである．パッチに分割するため局所的な特徴を捉えやすく，時間周波数表現において局所的に楽器の特徴が現れる演奏楽器分類に適していると考えられる．また，ネットワークがパッチ分割と畳み込みのみで構成されているため学習が容易で，事前学習モデルや転移学習などを用いなくても高精度なモデルの学習が可能である．

4　演奏楽器推定実験

4.1　データセット

本実験では，演奏楽器推定のデータセットとして独自の楽器音データセットを使用した．データセットには楽器単体の演奏音がサンプリング周波数 44.1kHz で 18 時間 57 分間収録されている．演奏音は防音の施された環境でレコーディングエンジニアの監修の元の収録された．単音が断続的に鳴らしている音

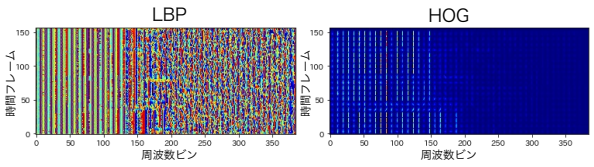

図 4　図 3(B) の対数メルスペクトログラムに画像特徴抽出法を適用した結果

と音楽を演奏している音が収録されている．楽器ラベルは 15 種類の調波楽器 (Acoustic Guitar，Electric Guitar，Acoustic Bass，Electric Bass，Piano. Choir，Flute，Keyboard，Male，Female，Strings，Trombone，Trumpet，Sax，Organ) と 7 種類の非調波楽器 (Cymbal，Snare，Percussion，Tom，Hi-Hat，Kick，Floor-Tom) がある．実験では，前処理として 16kHz にリサンプリングした後，音量が 40db 以下の区間を削除した．さらに，1 秒間ごとに分割した後，直流成分を除去することで音サンプルを作成した．作成したサンプルは学習用，評価用，テスト用データに分割した．調波楽器の学習データは 7.4 時間，評価用データは 1.8 時間テストデータは 2.3 時間となり，非調波楽器の学習用データは 3.0 時間，評価用データは 0.7 時間，テスト用データは 0.9 時間となった．クラスごとにサンプル数に違いがあり，調波楽器の最も多いクラスは Acoustic Guitar で 1.7 時間．最も少ないクラスは Organ で 0.3 時間である．非調波楽器の最も多いクラスは Cymbal で 1.3 時間，最も少ないクラスは Floor-Tom で 0.3 時間である．

4.2　実験条件

特徴抽出では音声波形の時間周波数領域の変換法として，STFT と CQT を用いる．STFT の窓長は 1024 点，遷移幅は 512 点とした．また，対数メルスペクトログラムを計算する際のメル次元は 385 とし，MFCC を計算する際のビン数は 40 とした．CQT の遷移幅は 512 点として，ビン数は 84，オクターブ数は 6 とした．画像特徴抽出法では LBP と HOG を用いた．LBP は 8 近傍でバイナリ化を行った．HOG は 2 次元データ内の局所的な勾配を抽出したものであり，視覚的な形状を抽出できる特徴量である．HOG は角度分解能 20°，1 セルにおけるピクセルサイズには 8x8、1 ブロックにおけるセルサイズには 3x3 の値を用いた．調波楽器の演奏音から抽出した対数メルスペクトログラムに画像特徴抽出法を適用した結果を図 4 に示す．どちらの手法も調波構造の部分が強調されているが、それぞれ強調した結果の傾向が異なることがわかる．

分類器には 64 層の畳み込み層とパッチ分割を含む ConvMixer を使用した．分類器の学習のエポック数は 100，最適化関数は Adam [15] を用いた．学習率は 0.0001 または 0.001 とした．データ拡張として Spec Augment [16] を用いた．Spec Augment における Time Wrap のパラメータを $w = 5$，時間マスクの最大幅 $\tau = 5$，周波数マスクの最大幅 $\nu = 13$ とした．

表 1　演奏楽器分類正解率 [%]

時間周波数表現	画像特徴抽出法	調波楽器	非調波楽器
対数スペクトログラム	-	96.89	**99.29**
	HOG	82.44	89.85
	LBP	95.03	98.72
対数メルスペクトログラム	-	**97.25**	98.99
	HOG	92.55	97.48
	LBP	91.14	95.99
CQT スペクトログラム	-	96.77	98.72
	HOG	91.17	96.50
	LBP	94.91	98.01
従来法 (MFCC + SVM)		86.90	98.99
従来法 (スカログラム + CNN)		86.27	97.39

本実験では従来法として楽器単体の演奏音を用いて演奏楽器推定を行っている 2 つの手法を用いた．1 つ目は MFCC を SVM で分類する手法 [9] で 2 つ目はスカログラムを 19 層 CNN で分類する手法 [10] で，それぞれ SVM と CNN を分類器と用いた先行研究で最先端の手法である．

4.3　実験結果

演奏楽器分類実験の結果を表 1 に示す．まず，調波楽器では対数メルスペクトログラムを，非調波楽器では対数スペクトログラムを用いた場合に最も高性能であることがわかる．このことから，演奏音の時間周波数表現を用いた ConvMixer で学習することで，従来法である SVM や CNN よりも高精度で演奏楽器を推定できることがわかる．次に，調波楽器と非調波楽器それぞれの傾向について述べる．調波楽器では，対数メルスペクトログラムの方が対数スペクトログラムや CQT より性能が高く，非調波楽器では対数スペクトログラムのほうが対数メルスペクトログラムや CQT より性能が高いことがわかる．この結果から，調波楽器では調波構造が低域強調により捉えやすくなり，精度向上に繋がったと考えられる．一方，非調波楽器では低域強調よりも高い周波数分解能が特徴を捉えるのに有効であったと考えられる．さらに，画像特徴抽出法の分類精度について確認すると，各時間周波数表現において画像特徴抽出法を適用したものと適用していないものを比較すると，画像特徴抽出法なしが最も高性能となっている．これは図 4 からも分かる通り，画像特徴抽出法では画像強調によって分類に必要な情報も一部欠落してしまっている可能性があると考えられる．しかし，それぞれ

表 2　アンサンブルを行った提案法の分類正解率 [%]

手法	調波	非調波
対数スペクトログラム (なし，LBP)	97.73	99.52
対数メルスペクトログラム (なし，HOG，LBP)	97.72	99.38
CQT スペクトログラム (なし，LBP)	97.65	98.99
すべてのアンサンブル (上記 3 つ)	**98.53**	**99.55**

の分類精度はある程度高いことから時間周波数表現とは異なる情報を抽出できていると考えられる．

次に各モデルのアンサンブルについて調査する．表 2 に提案法を用いてアンサンブルを行った結果を示す．表 2 には各時間周波数表現ごとに画像特徴抽出なし，HOG，LBP の 3 つモデルの組み合わせ 4 種類を評価し，最も性能が良かった組み合わせを示している．各時間周波数表現ごとに表 2 のアンサンブルを行った結果と表 1 の単独の特徴量を用いた提案法を比較すると，どの時間周波数表現においてもアンサンブルをすることでより高い性能となっている．このことから，画像特徴抽出法を用いて時間周波数表現の局所的な特徴を強調することで分類に貢献する特徴量を抽出できていると考えられる．表 2 では，異なる時間周波数表現も含めたアンサンブルの結果も示す．他のアンサンブルよりも高い性能となっており，このことから様々な特徴量を用いることで高精度に分類できると考えられる．

5　まとめ

本論文では，時間周波数表現を用いた畳み込みニューラルネットワークに基づく演奏楽器推定手法を提案した．調波楽器では低域強調を行うことで調波構造が強調されるため対数メルスペクトログラムが入力特徴量として適していることがわかった．非調波楽器では対数メルスペクトログラムが最も高い性能となり細かい周波数情報が特徴を捉えることに貢献した．また，画像特徴抽出法は時間周波数表現中の局所的な特徴を抽出することができ，得られた特徴量はアンサンブルを用いることで分類正解率向上に寄与することがわかった．今後の課題として，雑音や残響などへの頑健性の調査や，学習データに含まれない楽器への性能向上などが挙げられる．

謝辞

本研究の一部は JSPS 科研費 JP20H00613 及び ROIS-DS-JOINT (021RP2022) の助成を受けたものである．

参考文献

[1] A. Solanki *et al.*, “Music instrument recognition using deep convolutional neural networks,” *International Journal of Information Technology*, pp. 1–10, 2019.

[2] S. Prabavathy *et al.*, “An enhanced musical instrument classification using deep convolutional neural network,” *Int. J. Recent Technol. Eng.(IJRTE)*, vol. 8, no. 4, pp. 8772–8774, 2019.

[3] K. Racharla *et al.*, “Predominant musical instrument classification based on spectral features,” in *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 617–622, 2020.

[4] N. Gajhede *et al.*, “Convolutional neural networks with batch normalization for classifying hi-hat, snare, and bass percussion sound samples,” *In Proc. of The Audio Mostly*, pp. 111–115, 2016.

[5] H. Yun-Ning *et al.*, “Frame-level instrument recognition by timbre and pitch,” *In Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, pp. 135–142, 2018.

[6] Y. Han *et al.*, “Deep convolutional neural networks for predominant instrument recognition in polyphonic music,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 208–221, 2017.

[7] L. V. *et al.*, “Deep convolutional networks on the pitch spiral for musical instrument recognition,”

[8] 室谷良平ら, “Resonator 型くし形フィルタを用いた演奏楽器推定手法,” in **第 *79* 回音楽情報科学研究会**, 2008.

[9] K. Arimoto, “Identification of drum overhead-microphone tracks in multi-track recordings,” in *2nd AES Workshop on Intelligent Music Production*, 2016.

[10] A. Dutta *et al.*, “Cnn based musical instrument identification using time-frequency localized fures,” *Internet Technology Letters*, vol. 5, no. 1, p. e191, 2022.

[11] G. Sharma *et al.*, “Trends in audio signal feature extraction methods,” *Applied Acoustics*, vol. 158, p. 107020, 2020.

[12] X. Wei *et al.*, “Convolutional neural networks and local binary patterns for hyperspectral image classification,” *European Journal of Remote Sensing*, vol. 52, no. 1, pp. 448–462, 2019.

[13] A. Trockman *et al.*, “Patches are all you need?,” *arXiv preprint*, 2022.

[14] L. P. *et al.*, “Automatic instrument recognition in polyphonic music using convolutional neural networks,” *arXiv preprint*, 2015.

[15] D. P. Kingma *et al.*, “Adam: A method for stochastic optimization,” *In Proc. of International Conference on Learning Representations (ICLR)*, 2015.

[16] D. S. Park *et al.*, “SpecAugment: A simple data augmentation method for automatic speech recognition,” *In Proc. of Interspeech 2019*, pp. 2613–2617, 2019.