CE-007

拡散モデルと倍音の特徴を用いた音源分離手法の検討

川﨑　優生奈†　　田村　仁†
日本工業大学機械システム工学専攻†

1. はじめに

音楽情報処理では音源分離というテーマがあり，特にバンドを対象にした研究が世界的なコンテストが開催される程盛んに行われている．そこではボーカル，ギター，ドラム，その他の 4 つの分類分けに着目している．しかし，これら以外の楽器を対象にした研究は多くない．そこで，オーケストラや吹奏楽曲に注目し，本研究では木管アンサンブルから特定の楽器音を分離させることを目的とする．

音源分離の手法としては機械学習が多く用いられており，Open-Unmix[1]はその例である．また，楽器ごとに含まれる倍音には違いがあり，以下のような特徴が挙げられる．

1．金管楽器やダブルリードの楽器は全ての倍音を豊富に含む

2．木管楽器の中でもシングルリードの楽器は奇数倍の倍音を多く含むが偶数倍の倍音は少ない

3．フルートのようなエアリード楽器は倍音をほとんど含まずほかの楽器よりも倍音が少ない

本実験で用いた楽器のフルート，クラリネット，ホルンの音で B から 1 オクターブ上の B までの 8 音をスペクトログラムにして比較した画像を図 1 に示す．

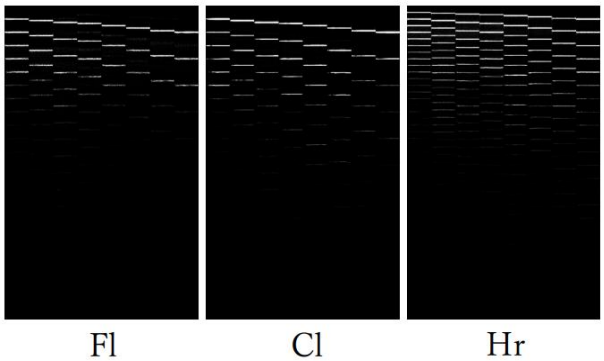

図 1　楽器による倍音の差

これを特徴とすることで音源分離が可能かどうか検討する．先行研究 1[2]では CycleGAN[3]を用いて微小時間ごとにスペクトログラムを合成する手法（以下手法①と呼ぶ）と，EfficientNet[4]を用いて特定の倍音が含まれているかどうか分類しスペクトログラムから除去する手法（以下手法②と呼ぶ）の 2 つの音源分離手法を提案した．さらに，先行研究 2[5]では生成した音楽データに混ざる雑音を抑制するために，手法②に対して画像ではなくバイナリファイルにすることで情報の劣化を防ぐ分離手法（以下手法③と呼ぶ）を提案した．しかし，どの手法も 1 つの楽器を分離できたとは言い難い結果となった．

スペクトログラムとは周波数分析を時間的に行い，色によって音の強さを表す，縦軸が周波数，横軸が時間のグラフである．縦軸を対数にしたスペクトログラムが一般的だが，対数軸だと倍音の特徴を捉えにくくなるという理由から線形のグラフを用いる．

CycleGAN とは画像合成の手法で，2 種類の画像間の変換方法を学習して，その結果を元に入力された画像に対応する画像を生成するというものである．

また，EfficientNet とは 2019 年に Google が発表した画像分類手法で，層を多くしすぎずに効率的なパラメータにすることで高精度の分類を可能にしたニューラルネットワークモデルである．

手法①は吹奏楽の演奏データをスペクトログラムに変換し，そのスペクトログラムを縦 1 ピクセルごとに切り分ける．その中から特定の楽器音を除去するように学習させた CycleGAN を用いてスペクトログラムを合成する手法で分離を行うものである．手法①では音楽データをスペクトログラムに変換したまま加工せずに学習させると，画像細部の情報を認識しにくくなるという欠点があったことを踏まえ，画像細部の情報を認識するためにスペクトログラムに加工を施した．CycleGAN は二次元の画像を合成する手法だが，スペクトログラムは横軸が時間のため音の連続性を除けば縦 1 列ずつが独立していると考えることができる．特定の音を除くのであれば瞬間的な音をスペクトログラムの 1 列で捉え，列ごとに処理を行うことで精度の高い分離を可能にすると推測した．これより，細部の変換を可能にするためにスペクトログラムを 1 列ずつ取り出して学習，変換させる手法を提案した．スペクトログラムを縦 1 ピクセルごとに分割した例を図 2 に示す．

Investigation of Sound Source Separation Method Using Diffusion Model and Harmonic Features

†Yukina Kawasaki，Hitoshi Tamura，Nippon institute of technology Graduate school Mechanical Systems Engineering Major

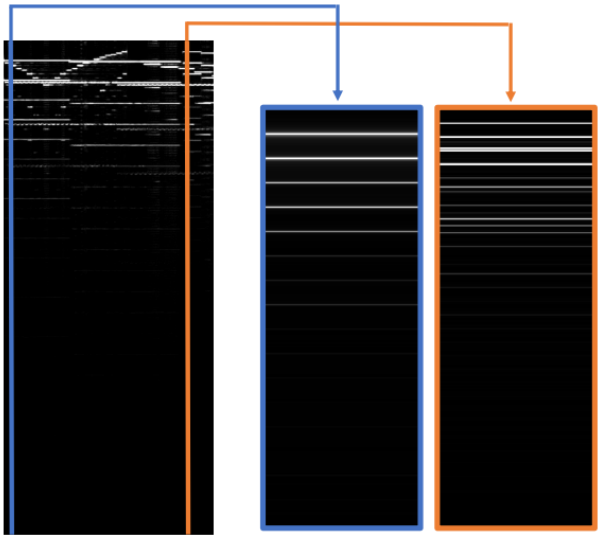

図 2　スペクトログラムの加工例

しかし，縦 1 列ごとに処理を加えた方が分離精度が良いとはいえ，CycleGAN のみでは分離させる楽器の特徴を捉えられていない結果となった．

一方で手法②は，倍音の特徴に着目するように EfficientNet で学習した分類器を用いて除去対象の楽器音が含まれているかどうかの判別器を構成し，分離させたい楽器音のスペクトログラムを白黒変換した値を任意の列に掛けることで分離を行うものである．

これら 2 手法の結果からは生成した音楽データに混ざる雑音が目立ち，原因として音楽データを画像にすることで情報量が落ちてしまうことや，縦 1 列で分割しているため音の連続性を考慮しなかったことから音がぶつ切りになっているということが考えられた．これに対し，情報量を保持したまま音源分離を行う手法③を提案した．

手法③は，手法②で音楽データを画像のスペクトログラムに変換していた箇所を全てバイナリデータにしたものである．吹奏楽の演奏データをフーリエ変換し，強度のデータをバイナリファイルとして保存する．その中から手法②のように特定の楽器音を倍音の特徴を用いて分離させる．結果としてバイナリデータにすることで情報の劣化自体を防ぐことはできたが，学習させるにはデータが重く，画像を用いた実験と比較して十分な学習ができずに分類の精度が低くなってしまった．

上記 3 手法の結果より，音源分離手法を改良するには雑音抑制のために音の連続性を考慮した方法の検討，さらに CycleGAN でない別の画像合成手法を用いた音源分離の検討が必要とされる．そこで，本研究では拡散モデルの画像合成手法に着目し，音源分離の精度を向上させることができるかどうか検討する．

2. 提案手法

本研究の目的は，倍音に着目した，吹奏楽曲を対象にした微小時間ごとの拡散モデルによる音源分離の手法を提案することである．また，予備実験として音の連続性を考慮した手法での音源分離を行う．

このために予備実験の手法④と本研究の提案手法⑤を新たに設計した．まず手法④では吹奏楽の演奏データをスペクトログラムに変換し，そのスペクトログラムを縦 1 ピクセルごとに切り分ける．その後倍音の特徴に着目するように EfficientNet で学習した分類器を用いて，除去対象の楽器音が含まれているかどうか決定する判別器を構成し，これを用いて分離させる．さらに分離後のスペクトログラムに対して音楽データに混ざる雑音を抑制するために，なめらかなスペクトログラムを生成するように学習させた CycleGAN で画像合成するという方法で音源分離を実現する．手法④の概要を図 3 に示す．

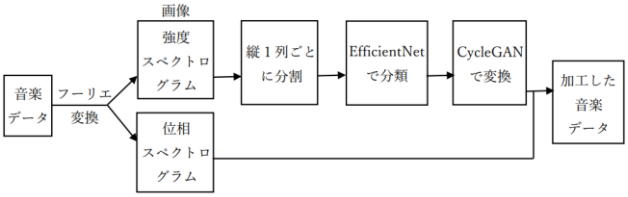

図 3　手法④の概要

次に手法⑤は，はじめに手法④と同様に吹奏楽の演奏データをスペクトログラムに変換し，縦 1 列ごとに分割する．その後，CycleGAN を用いて画像合成していた箇所に代わりに拡散モデルの画像合成手法を用いて，分離させる対象の楽器音が含まれていないスペクトログラムを生成するという方法で音源分離を行う．ノイズに 3 楽器含まれている曲を加えて，分離対象以外の 2 楽器のみの曲を学習させることで，3 楽器含まれる曲を入力すると 2 楽器のみの曲を出力するようにし，これによって音源分離を行う手法を検討する．手法⑤の概要を図 4 に示す．

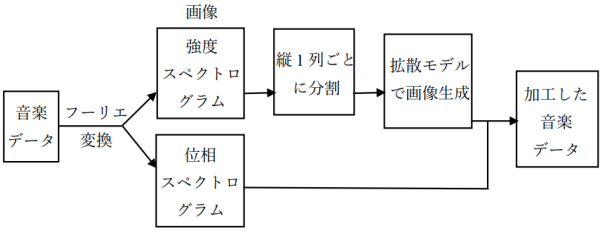

図 4　手法⑤の概要

以上の手法で音源分離の性能を高めることができるか実験で確かめる．

3. 評価実験

3. 1 実験方法

本実験では，フルート，クラリネット，ホルンの 3 種類の楽器の演奏からホルンの音を分離させる．本抄録では実験時間の都合上，予備実験の結果のみを記載し，手法⑤での分離結果は発表で述べる．

学習用と評価用のデータセットには，Apple 社の音楽制作ソフトウェア GarageBand で打ち込んだ音源を利用した．音楽データに対してフーリエ変換を行い，音楽データの強度と位相の値を取得する．そのうち強度のデータを 1 ピクセル毎に読み取って画像にプロットしスペクトログラムを作成する．2 節で説明した手法でスペクトログラムを加工した後，元の音楽データの位相データと合わせ，フーリエ逆変換により音楽データを生成する．

評価用のデータセットは後に性能を確かめるため，3 種類の楽器が含まれている曲と対になるようなフルート，クラリネットのみ含まれている曲を 1 データ 10 秒で 500 曲用意した．3 種類の楽器が含まれているデータを分離させ，その分離結果を比較する．

手法④　画像合成と分類器を合わせた手法

EfficientNet で分類するホルンの音は半音含む 2 オクターブ分の 25 音とした．ホルンの 1 音に対して，train に 10000 データ，validation に 1875 データ，test に 625 データのスペクトログラムを用意し学習を行い，25 音に分類する多クラス分類器を作成した．評価データを分類した後，含まれていると判断された音を該当箇所から，除去対象の音の白黒変換した値を任意の列に掛けることで分離させる．白黒変換する値の元データに使用したホルンの 25 音分のスペクトログラムを図 5 に示す．

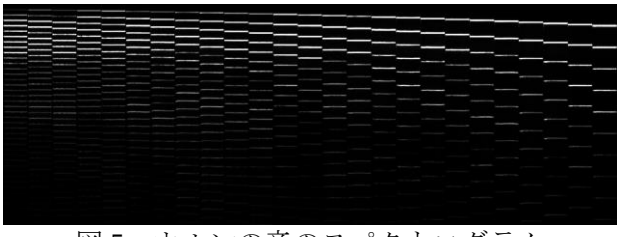

図 5　ホルンの音のスペクトログラム

CycleGAN の学習データセットには，手法②を用いてフルート，クラリネット，ホルンの含まれた曲からホルンを分離させた後のスペクトログラムと，フルート，クラリネットのみが含まれた曲の 2 つのグループを用意した．それぞれ train に 1800 データ，test に 200 データ用いて 200epoch で学習させた．分類器を用いた分離手法でスペクトログラムの加工を行った後，なめらかな画像になるよう学習させた CycleGAN で評価用データを変換し，音楽データを生成する．

3. 2 評価指標

評価尺度として，信号対歪み比(Signal to Distortion Ratio：SDR)を用いる．生成した信号が目的とする信号に対しどれほど歪んでいるかを評価し，値が大きいほど時間波形の歪みが小さいことを示す．SDR の求め方を式 1 に記す．

$$SDR = 10\log_{10}\frac{\sum_n |s(n)|^2}{\sum_n |s(n) - \hat{s}(n)|^2} \qquad \text{(式 1)}$$

$s(n)$：目的の信号　$\hat{s}(n)$：処理後の信号

3. 3 実験結果

評価データの分離結果の一例を図 6 に示す．「元画像」は分離前の 3 楽器含まれた音源のスペクトログラム，「正解画像」はフルートとクラリネットのみの音源のスペクトログラムである．

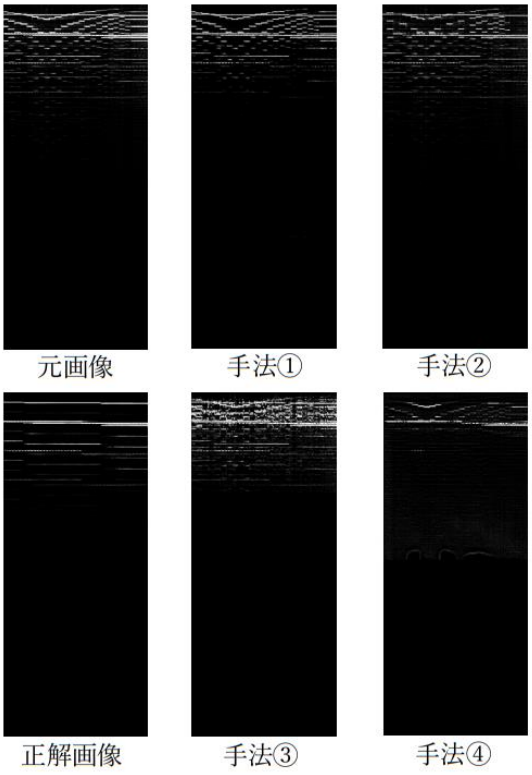

図 6　分離結果

さらに，手法①から④で評価実験を行い，SDR を比較し性能を確かめる．評価用のデータ 500 曲の SDR の平均を表 1 に，各手法での分離量を比較した画像を図 7 に示す． 図 7 の「正解画像」は分離対象のホルンの音源のスペクトログラムである．

表 1　500 曲分の SDR 平均比較

	SDR平均(db)
手法①	12.39
手法②	14.81
手法③	13.96
手法④	11.35

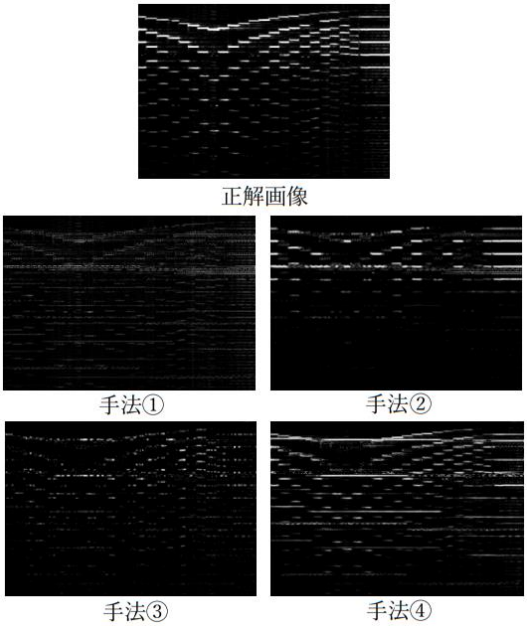

図 7　各手法での分離量比較

4. 考察

CycleGAN を用いた画像合成での分離手法①と EfficientNet と CycleGAN を組み合わせた分離手法④を比較する．表 1 より，SDR の評価は手法④が最も低く，提案手法が従来手法に劣っていると読み取られる．しかし，図 7 の分離量を比較すると，分離させる対象のホルンの正解画像と比べ，手法①はおおよそ分離できていないように見える．一方で手法④の分離量は正解画像と比べて多く，必要以上に分離してしまっていることがわかる．これより，特定の楽器音を認識できていない手法①よりも手法④の SDR の方が低い原因は，手法④で分離させる楽器以外の音も抜けてしまっており，全体の音量が小さくなっていることであると考えられる．これは手法②，手法③と比較した際にも同様のことがいえる．

また，EfficientNet を用いた分類器での分離手法②と手法④を見ると，表 1 より，SDR の評価では手法②が 4 手法のうち最も高い精度を示している．しかし，図 7 から分離している箇所は一部分であることがわかる．よって，定性的に見ると手法③は分離させる目的の楽器音が最も抜けており，4 手法で最も正解に近い結果であると推測できる．必要以上に分離されている箇所については，分類器の正解率が 96.3%と高いことから，分離方法に課題が残されていると考察する．

以上より，今後は手法④の分離方法に対して，用意しているテンプレートの幅を広げ余裕を持たせるというような工夫することで音源分離の精度の向上が見込めると推測する．また，手法①で用いている画像合成手法を拡散モデルにすることで分離の精度を高めることができるか検討する．

5. おわりに

本研究では，倍音に着目した吹奏楽音源での微小時間ごとの拡散モデルでの音源分離手法，また，画像合成と分類器を組み合わせた音源分離手法を提案した．CycleGAN を用いた画像加工と倍音を特徴とした EfficientNet での画像分類の 2 つの手法を合わせて実験を行い，微小時間ごとの処理，また倍音を機械学習で分類する手法が音源分離において有用であることを示した．今後は拡散モデルを実験に取り入れ，スペクトログラムの加工方法を工夫することで音源分離の性能向上を目指す．

参考文献

[1]F.-R. Stöter， S. Uhlich， A. Liutkus， and Y. Mitsufuji，”OpenUnmix - A Reference Implementation for Music Source Separation”， Journal of Open Source Software， vol. 4， no. 41， p. 1667， 2019.

[2] 川﨑 優生奈，田村 仁，CycleGAN と倍音の特徴を用いた微小時間ごとの音源分離手法の検討，第 22 回情報科学技術フォーラム講演論文集第 2 分冊，pp．319-322，2023.

[3]Jun-Yan Zhu，Taesung Park，Phillip Isola，Alexei A. Efros，”Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks”，IEEE Conference on computer vision．pp. 2223-2232，2017.

[4]M. Tan and Q. Le，”EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”，Los Angeles USA: Proceedings of Machine Learning Research，vol. 2019，pp. 6105-6114，2019.

[5] 川﨑 優生奈，田村 仁，倍音の特徴を用いた音源分離手法における雑音抑制の検討，第 85 回全国大会講演論文集，Vol. 1，pp. 557-558，2023.

[6]久野 文菜，大場 隆史，中園 歩，谷口 航平，畑中 衛，林 広幸，濱川 礼，End-to-End 学習を利用したスペクトログラム生成による楽器音抽出手法の提案，情報処理学会研究報告，Vol. 2019-EC-51，No. 18，pp. 1-2，2019.

[7]Phillip Isola，Jun-Yan Zhu，Tinghui Zhou，Alexei A. Efros，”Image-to-Image Translation with Conditional Adversarial Networks”，IEEE Conference on Computer Vision and Pattern Recognition(CVPR)．pp. 5967-5976，2018.