E-018

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

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，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．

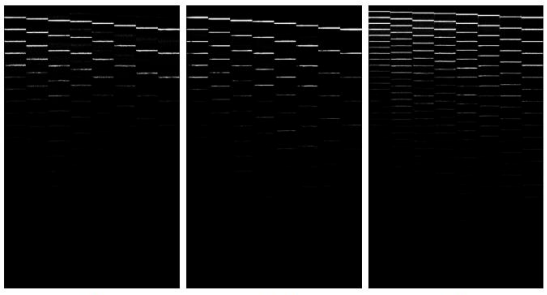

図 1　楽器による倍音の差

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[2]では CycleGAN[3]を用いてスペクトログラムを合成する手法（以下手法①と呼ぶ）と対象楽器の倍音を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が，どちらも 1 つの楽器を分離できたとは言い難い結果となった．

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，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．

手法①は 10 秒の音楽データをスペクトログラムに変換し学習，画像合成するものである．この手法①の欠点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たことから画像細部の情報を認識しにくくなるという点である．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一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する．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 行ずつ取り出して学習，変換させる手法を提案する．これを手法③とする．

一方で手法②は，機械学習を使用せずにホルンの各音の倍音との相関係数を見て，一定値以上の値が出たときに白黒変換した値を掛けることで分離するものである．手法②には分離した量が僅かであるという欠点がある．また，分離させる楽器の倍音の検出方法に課題がある．そのため，分離させる音楽データと特定の楽器の倍音の相関係数を見て分離させるか判断していた箇所を，EfficienetNet[4]を用いて特定の楽器の倍音が含まれているかどうか分類させる手法を検討する．これを手法④とする．

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音源分離の手法を提案することである．

Investigation of Sound Source Separation Method at Microtemporal Intervals Using CycleGAN and Harmonic Features

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

このために手法③と手法④を新たに設計した．まず手法③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方法で分離を行う．次に手法④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，それを用いて音源を分離させる．

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．

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．提案手法の概要を図 2 に，スペクトログラムの加工例を図 3 に示す．

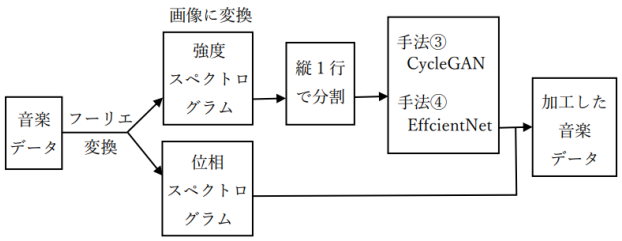

図 2　提案手法の概要

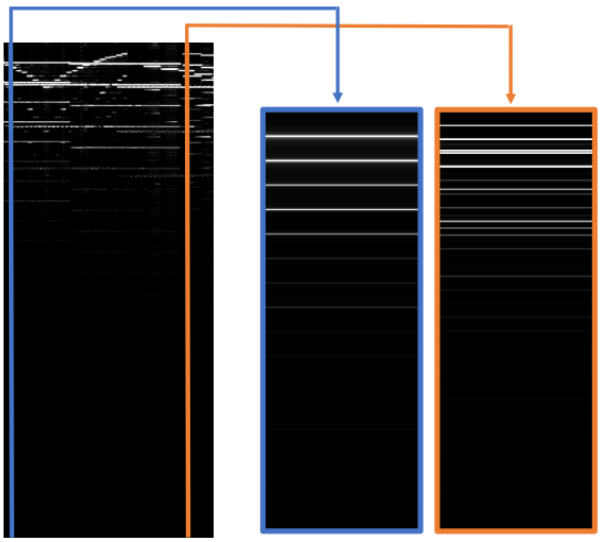

図 3　スペクトログラムの加工例

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．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 2 つの提案手法で分離させ，分離結果を比較する．

手法③　1 列ずつ画像合成する手法

CycleGAN の学習データセットにはフルート，クラリネット，ホルンの含まれた曲とフルート，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 22000 データ，test に 2200 データ用いて 200epoch で学習させた．学習後 CycleGAN で評価用データを変換し，画像のコントラストを調整してから音楽データを生成する．

手法④　倍音を機械学習で分類する手法

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．25 音全ての分類器は未実装であり，B♭，D，F の 3 音の分類器で実験を行う．1 音分の分類器に対して，train に 10000 データ，validation に 1875 データ，test に 625 データの学習データを用意し学習を行った．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分離させ，分離したい音の白黒変換した値を任意の列に掛けることで分離を行う．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 4 に示す．

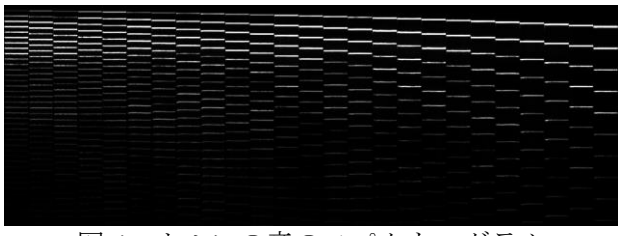

図 4　ホルンの音のスペクトログラム

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．

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

$s(n)$：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 5 に示す．(a)は分離前の元音源，(b)は手法③の結果，(c)は手法④の結果，(d)はフルートとクラリネットの正解音源のスペクトログラムである．

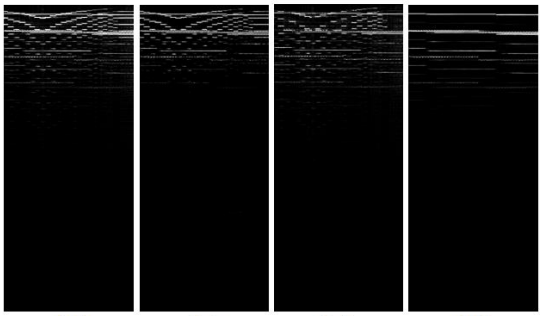

図 5　分離結果

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．500 曲のうち手法③での結果の方が高かったデータは 268 曲，手法④での結果の方が高かったデータは 232 曲であった．

また，手法④での分離精度を測定するため，ホルンの音源に手法④で学習した 3 音のみが含まれたデータを用いて評価データを 10 曲用意した．評価データ 10 曲の SDR の平均を表 2 に，各手法での分離量を比較した画像を図 8 に示す．

表 1　500 曲分の SDR 平均比較

	SDR平均(db)
①	6.84
②	3.12
③	5.92
④	5.61

表 2　3 音のみのデータ 10 曲分の SDR 平均比較

	SDR平均(db)
①	6.42
②	2.26
③	5.34
④	2.38

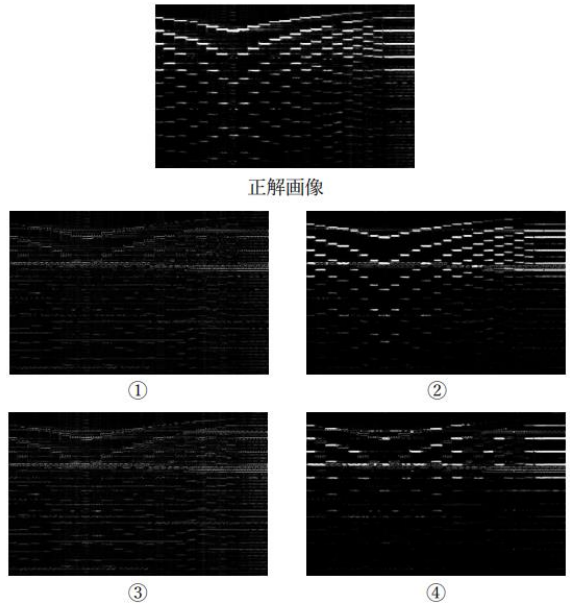

図 6　各手法での分離量比較

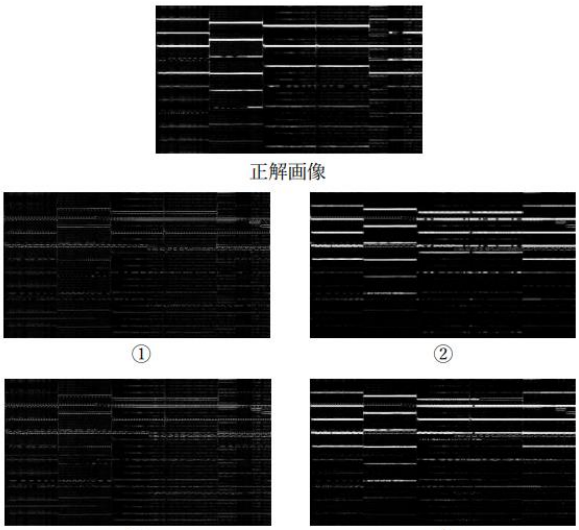

図 7　各手法での 3 音分データの分離量比較

4．考察

画像合成を用いた手法①と手法③では，表 1，表 2 の SDR の結果より手法①の方が高い性能が出ていることから，スペクトログラムに加工を行わない方が高い性能になることが分かる．1 列ずつ分割して別々に加工をしたため音の繋がりが再現できていないのではないかと考える．

また，倍音に着目した手法②と手法④では，表 1，表 2 の SDR の結果より手法④の方が高精度の分離であるということが読み取られる．これより，特定の楽器の倍音が含まれているかどうか機械学習を用いて判断することは有用であると言える．

次に画像合成を用いた手法①と倍音を機械学習で分類する手法④を比較する．表 1，表 2 の SDR

の結果からは手法①の方が精度が良いと言える．しかし，図 6，図 7 より分離させるべきホルンの音は正解画像の通りで，手法①の結果は波形が大きく異なることが分かる．一方で手法④の結果は学習済みの 3 音分の分離結果の図 7 から正解画像と近しい波形の画像が生成されていることがわかる．手法④は 3 音分の分類器しかないため図 6 の結果では正解画像と一致している箇所が手法②よりも少ないが，以下の図 8 に示すように学習した 3 音を分離しており，分離させるべき音を特定してスペクトログラムを加工する処理が各所で行われている．これより，未実装の音の分類器を作成すれば図 6 のデータでも手法②のように分離する可能性が高いと予測される．

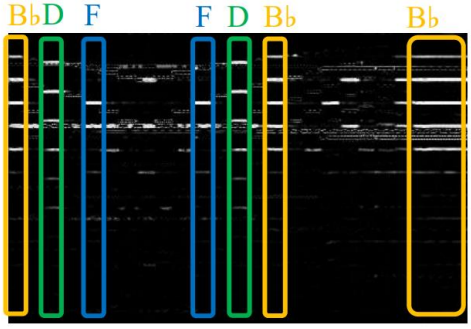

図 8　手法④の詳細な分離量

また，手法①の結果を見ると明度の低い分離結果画像が生成されている．一般に GAN のような画像合成手法では，損失関数が目標画像と合成画像の差分を取っている．このため今回のスペクトログラムのような線画状の画像を目標とすると，平均的に見て差分を小さく抑え込むように全体的に線がぼやけた曖昧な画像を生成しやすいという欠点がある．これは手法③の結果にも言えることである．

手法②と手法④の分離量の結果が正解画像と近いにもかかわらず SDR の結果が低かった原因として，対象楽器の倍音以外の倍音にも反応してしまい，必要以上に分離したことから音楽データの音量が小さくなったことが考えられる．

さらに，図 6 より手法②は最も正解画像に近いスペクトログラムを生成したが，表 1 より SDR は他の 3 手法と比べて著しく低い結果となった．図 9 は手法②の分離結果を拡大した画像である．

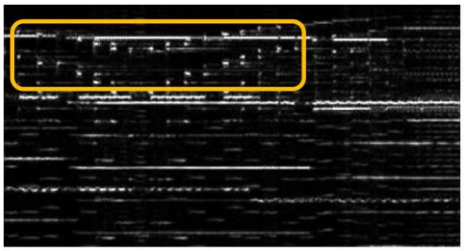

図 9　手法②の分離結果

図 9 から，同じ音でも倍音に多少のずれがあると認識できず，分離しきれずに一部が残っている画像になる．これがノイズになってしまうと読み取られる．そのため SDR の低さはノイズの多さにも起因すると予測される．

以上より，手法④の分類器を用いて分類した後，機械学習を用いて分離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．

今後は上記の改良案に加え，未実装の音の分類器を作成する，現状 1 音ずつの分類器を用いているところを 25 音の多クラス分類器にすることで分類の精度を高める．また，画像合成する方での問題点を 25 音別の CycleGAN を作ることによって改善できるか実験を行う必要がある．

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 つの手法で実験を行い，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．今後は分類器をホルンの全ての音に対応させ，スペクトログラムの加工方法にも機械学習を取り入れ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 を用いた音源分離の倍音成分に着目した手法の検討，情報処理学会第 85 回全国大会講演論文集(2)，pp. 557-558，2023.

[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]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.

[6]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.