1V-08

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

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．

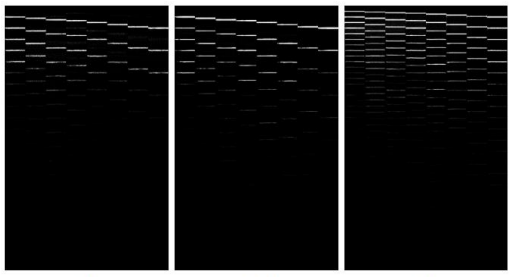


図 1　楽器による倍音の差

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 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し除去する手法（以下手法②と呼ぶ）の 2 つの音源分離手法を提案した．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．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．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ものである．次に手法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，抜き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ものである．しかし，双方とも生成した音楽データに混ざる雑音が目立つ結果となった．そこで，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が原因であると考え、本研究ではまず手法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抑えられないか検討する．これを手法③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音源分離の手法における雑音抑制を提案することである．このために手法③を新たに設計した．手法②で画像を用いていた箇所を全てバイナリファイルにした手法である．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．その中から特定の楽器音を倍音の特徴を用いて分離させる．提案手法の概要を図 2 に示す．

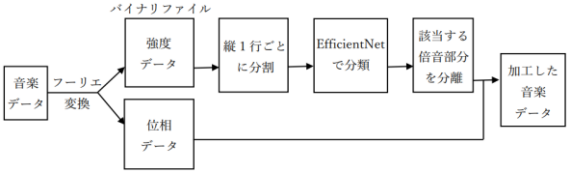


図 2　提案手法の概要

Investigation of Noise Suppression in Sound Source Separation Method Using Harmonic Features

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

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．また評価用のデータとして，3 種類の楽器が含まれる曲と対になるフルート，クラリネットのみの曲を 1 データ 10 秒で 500 曲用意した．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，train50 データ，validation10 データ，test5 データ用意して学習を行った．特定したホルンの音を引き，分離させる．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し，そのうち強度のデータをバイナリファイルで保存する．分類し任意の音を分離させた後，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．

3. 2 評価指標

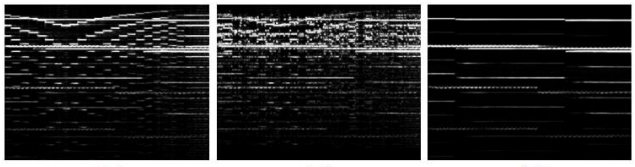
評価尺度として，信号対歪み比(Signal to Distortion Ratio：SDR)を用いる．生成した信号が目的とする信号に対しどれほど歪んでいるかを評価し，値が大きいほど時間波形の歪みが小さいことを示す．SDR の求め方を式 1 に記す．

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

$s(n)$：目的の信号　$\hat{s}(n)$：処理後の信号

3. 3 実験結果

分離結果を図 3 に示す．(a)は分離前の元音源，(b)は手法③の結果，(c)は正解音源のスペクトログラムである．



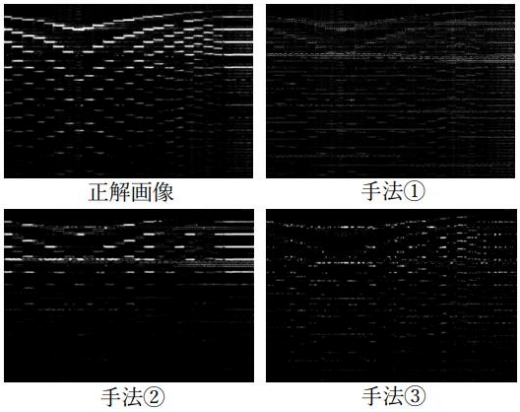
(a)　(b)　(c)

図 3　分離結果

また，手法①から③での評価用のデータ 500 曲の SDR の平均を表 1 に，分離量を比較した画像を図 4 に示す．

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



正解画像　手法①　手法②　手法③

図 4　各手法での分離量

4. 考察

表 1 の SDR の結果からは手法②が最も良い精度を示していることが分かる．図 4 の分離量からも正解画像に最も近いのは手法②だと言える．手法③は双方の結果で手法②に劣っているが，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るも学習させるにはデータが重く，本実験では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と考えられる．これより，音楽データの画像化は大きな問題はなく，倍音の特徴を捉えることと画像全体の加工を機械学習で合わせて行うことで精度の向上が見込めるのではないかと推測する．

5. おわりに

本研究では，倍音の特徴を用いた音源分離手法における雑音抑制方法を提案した．しかし，画像化を避けたことによる結果は芳しくなかった．今後は CycleGAN と EfficientNet を用いた従来手法を組み合わせ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.