5S-02

ギターパートを対象とするエンドツーエンド音源分離の検討

尾関 日向 †　　酒向 慎司 ‡
名古屋工業大学 †　名古屋工業大学 ‡

1　はじめに

音源分離とは，複数の音源が混ざりあった音響信号から特定の要素を抽出する技術である．なかでも音楽音響信号を対象にした音源分離は，楽曲が含む各パートの音源を必要とする自動採譜技術にとって欠かせない要素技術である．従来はボーカルやドラムパートの分離が多く取り組まれてきたが，一般的に楽曲は他にも様々なパートを含んでおり，任意のパートが抽出できればより実用的な自動採譜や既存曲のリミックス，DJ プレイなどに広く活用できる．例えばギターはポピュラー音楽で頻繁に使用され演奏者も多く，先述の用途において需要が高い．

そこで本研究では最新のボーカル分離技術を適用したギターパート分離を試みる．特に，ステレオ音源の場合の定位情報，データセットの規模，パートが担う演奏上の役割の違いといった要素が分離精度に及ぼす影響を調査し，ギターにより適した分離手法を検討する．

2　音楽音響信号を対象とする音源分離

近年はディープラーニングを用いて分離音を推定する手法が数多く登場し，分離精度は飛躍的に向上している．ところが従来の音源分離の手法のほとんどは入力としてスペクトログラムを使用しており，位相情報を省略している．そこで位相を含め音声信号がもつあらゆる情報を考慮するため，スペクトログラムではなく波形を入力とする Wave-U-Net[1] が提案された．Wave-U-Net は全層畳み込みネットワークである U-Net を 1 次元時間領域に適応させたエンドツーエンドの学習モデルであり，最新のスペクトログラムベースの手法よりも優れた分離性能をもつことが示されている．本研究ではこの Wave-U-Net を用いる．

End-to-End Audio Source Separation applied to Guitar Extraction
† Hyuga Ozeki, Nagoya Institute of Technology
‡ Shinji Sako, Nagoya Institute of Technology

3　Wave-U-Net を用いたギターパート分離

過去のボーカル分離問題ではメインボーカルだけでなくコーラスも一括りに分離するような設定が一般的であった．今回はそれに倣い，楽曲が持つ全ての目的トラックをまとめて分離するようなモデルを作成する．モデル学習には Wave-U-Net を用い，ギターとボーカルに対しモノラル音源分離モデルをそれぞれ作成する（**実験 1**）．加えてステレオ入出力に対応したモデル，学習データサイズを 2 倍に拡張したモデルをそれぞれ作成し，分離精度の変化を比較する（**実験 2，3**）．また，ギターパート特有の分類として，主旋律を担当する「リードギター」，リズムに沿って伴奏を弾く「リズムギター」と呼ばれるの 2 つの演奏上の役割があり，音響的特徴も異なる．そこでリードギターとリズムギターそれぞれに対する分離モデルも作成し，分離精度との関係を調査する（**実験 4**）．

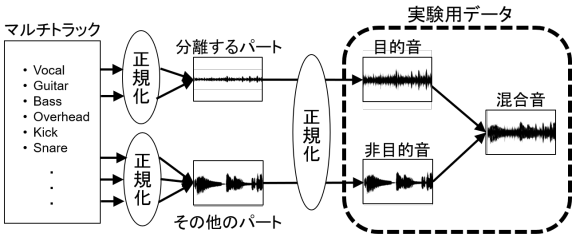

図 1　実験用データの作成

4　データセット

楽曲がパートごとに収録されたデータセットは数少なく，本実験では MedleyDB[2] からギターまたはボーカルを含む 120 曲分のマルチトラックを使用する．実験用データは図 1 のように作成する．まず楽曲ごとにマルチトラックを分離対象のトラックとそうでないトラックに分け，それぞれ混合して目的音，非目的音とする．さらにそれらを混合したものを分離前の元音源とみなす．混合の際にはトラックどうしの音量ピークが等しくなるようあらかじめ正規化する．実際の楽曲でこのような単純なミキシングは行われないが，本実験では考慮しない．さ

らに，**実験 3** のために異なる曲の目的音・非目的音を組み合わせて疑似的に曲数を 2 倍にしたデータセットも用意する．学習データとテストデータの比率は 3:2 とする．

5 実験

5.1 評価方法

推定した分離音源を評価するために一般に広く用いられる SDR(Signal-to-Distortion Ratio)[3] を採用する．この値が高いほど分離精度が良いと言える．

$$\mathrm{SDR} := 10\log_{10}\frac{\|s_\mathrm{t}\|^2}{\|e_\mathrm{i}+e_\mathrm{a}\|^2}\ [\mathrm{dB}]$$

s_t : 推定信号内の目的音成分
e_i : 推定信号内の非目的音成分
e_a : 推定信号内のその他の成分

5.2 分離モデルの作成

実験 1〜4 に対応する以下の分離モデルを作成した．

M1G, M1V
: それぞれギター，ボーカルに対するモノラル音源分離モデル

M2G, M2V
: G, V をステレオ音源に対応させたモデル

M3G, M3V
: G, V の学習データサイズを 2 倍にしたモデル

M4L, M4R
: それぞれリードギター，リズムギターに対するモノラル音源分離モデル

5.3 実験結果と考察

作成した各モデルの分離精度を表 1 に示す．ただし Tar., Oth. はそれぞれ目的音，非目的音を指し，分離精度は推定音源ごとに算出された SDR の平均である．

	M1G	M2G	M3G	M1V	M2V	M3V	M4L	M4R
Tar.	1.68	2.08	2.39	5.36	4.67	5.79	2.51	-5.99
Oth.	3.92	3.95	4.64	8.58	8.24	9.27	5.59	5.16

表 1　各モデルの分離精度 (SDR(dB))

モノラル分離とステレオ分離の精度を比較すると，ギター分離においては SDR に大きな変化はなく定位情報が分離に貢献していないことがわかる．ボーカル分離においては過去の Wave-U-Net を使った実験 [1] では定位情報を与えることで非目的音の分離精度が大きく向上したが，今回の実験ではそのような結果は見られなかった．これは使用したデータセットの違いによるものと考えられる．

データセットを疑似的に拡張したモデルではボーカルに比べギターの分離精度が大きく向上し，学習時の損失関数の収束にも改善が見られた．ギター分離がボーカルの場合よりも大きなサイズの学習データを必要とした理由として，ギターのほうが音色のバリエーションが多いために様々なタイプの楽曲を学習しなければならないことや，曲中の演奏時間が短いケースがあることが挙げられる．

リードギターとリズムギターの比較ではリードギターのほうが分離精度が高い結果となり，同じギターという楽器でも演奏上の役割によって分離の難易度に差があることが確認できた．今回モデル作成時のハイパーパラメータは過去のボーカル分離実験 [1] 時の設定を流用したため，メロディ構造を含み音響構造の変化がボーカルにより近いリードギターにおいて高い分離精度が出たと考えられる．リズムギターにはリードギターほど頻繁な音響的変化がないとすれば，畳み込みの入力サイズを大きくして時間スケールを長くするなどの調整でよりリズムギターに適したモデルを作ることができるだろう．

また，音源が部分的に静かであったり無音の場合，SDR が極端に低くなることが従来研究において指摘されている [1]．今回の実験でもいくつかの曲で同様の結果が確認され，それらの曲の中には分離パートの演奏時間が極端に短いものがあった．よって音楽のようなスパース性のある信号の分離精度の評価指標には検討の余地があると言える．

6 まとめ

本研究では既存のボーカル音源分離手法を用いてギターパート分離を試みた．性能の高い分離モデルを作るためにはギター特有の幅広い音色や演奏上の役割の違いなどを踏まえた手法を考案したり，より大きなサイズのデータセットを用意する必要がある．ギター以外の楽器についても同様に検討を行わなければならない．また，現在用いられている分離音源の評価指標には音楽音響信号を対象とする場合において，改善の余地がある．

参考文献

[1] Daniel S. *et al.* : “Wave-U-Net: a multi-scale neural network for end-to-end audio source separation,” 2018
[2] Rachel B. *et al.* : “MedleyDB: a multitrack dataset for anotation-intensive MIR research,” 2014
[3] Emmanuel V. *et al.* : “Performance measurement in blind audio source separation,” 2006