E-035

エレクトリックギターパートの音源分離のための BLSTM による音源分離手法
BLSTM-Based Music Source Separation Method for Separating Two Electric Guitar Parts

山西 陽明† 村松 駿‡ 吉田 孝博†
Haruki Yamanishi Shun Muramatsu Takahiro Yoshida

1. はじめに

演奏動画編集，耳コピ，DJ プレイなどのリミックスなどの音楽制作や音響処理において，音源分離は品質向上や編集者の負担軽減のために重要で有用な技術である．現在の音源分離は，例えば，ボーカル，ベース，ドラム，これら以外のパートのように，各パートに分離する技術が主流であり，Moises[1]や Open-Unmix[2]，Wave-U-Net[3]などを用いたものなど，様々な手法が提案されるとともに，ソフトウェアも市販され，発展してきている．

しかし，リードパートとリズムパートのように 2 名のギタリストが含まれる楽曲において，各ギターパートまで分離できる音源分離技術は，必要とされているものの，いまだに発展途上である．この 2 つのギターパートの音源分離について，文献[3]の一部で挑戦されているが，分離性能が低いことが課題である．また，片パートは単音で演奏し，他パートは和音で演奏するという制約を設けたラベル付けにより分離を行う市販のソフトウェアも存在するが，この制約も実用上の課題となっている．

そこで本研究は，再帰型ニューラルネットワーク (RNN: Recurrent Neural Network)の一種である双方向長短期記憶 (BLSTM: Bidirectional Long Short Term Memory)を用いて，2 つのエレクトリックギターパートが含まれる音源において，それぞれのギターパートを分離できる手法を提案した．なお，複数種の楽器パートが含まれる楽曲から楽器毎のパートへ音源分離する技術は日々発展する手法や製品を利用できるため，本研究の提案手法は，楽曲からギターパートが抽出された後の音源に対して各ギターパートへ分離する処理を行う手法である．

本論文では，この提案手法の詳細を述べるとともに，2 つのエレクトリックギターパートを含む音源を用いた音源分離性能評価結果について示す．

2. 提案手法（2 つのエレクトリックギターパートの音源分離手法）

提案した分離手法の構造を図 1 に示す．図 1 に示すように，入力される分離前の楽曲の時系列音声信号は，短時間フーリエ変換 (Short Time Fourier Transform)により，時間-周波数領域での特徴量（スペクトログラム）を抽出し，BLSTM が 3 段直列接続された RNN に入力して音源分離処理を行う．RNN の出力は逆短時間フーリエ変換 (Inverse STFT)により音源分離後の各ギターパートの時系列音声信号を得る．BLSTM は，3 層を直列接続することによって精度の向上を目指した．勾配消失の問題解消や学習の安定化を図るために，1 層目入力から 3 層目出力でスキップ接続

† 東京理科大学大学院工学研究科 Graduate School of Engineering, Tokyo University of Science
‡ 東京大学大学院工学系研究科 Graduate School of Engineering, The University of Tokyo

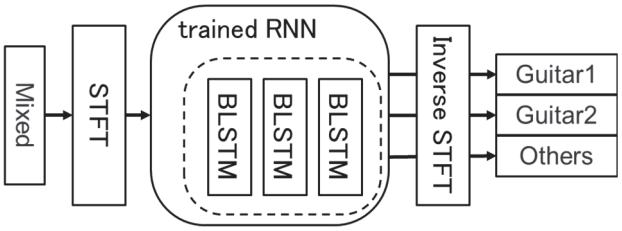

図 1 2 つのギターパートを分離する提案手法の構造

を行った．RNN のモデル学習は，複数種のギターパートが含まれるミックス音源を入力データ (Mixed)として，各ギターパートの音源 (Guitar 1, Guitar 2)を教師信号として行う．一般的に，多くのバンド演奏では，使用するギターやエフェクトの種類が各ギターパートで異なることから，本提案手法ではこの音色の違いを手掛かりに，譜面上の制約が不要なギターパートの音源分離を行っている．なお，本提案手法の前処理として，楽曲を楽器毎のパートへ分離する既存の技術[2]を用いるが，現在主流の手法のように，ギターパートが音源分離対象の楽器となっておらず，分離対象となっているボーカル，ベース，ドラム以外のパートが全て，その他のパートとして混合されて出力されてしまう手法と組み合わせざるを得ない場合も現状では存在する．その場合には，本提案手法において，RNN の学習を，複数種のギターパートと別の楽器が含まれるミックス音源を入力データとして，その他のパートに含まれる各ギターパートをそれぞれ教師信号として使用することで，前段の音源分離の制約に対応する．

3. 分離精度評価方法

本研究では，提案手法における 2 つのギターパートの音源分離精度を評価するとともに，ギターパートの特徴と分離性能との関係も考察する．

使用したデータセットは，MoisesDB[4]，MedleyDB[5]，MedleyDB 2.0[6]である．サンプリング周波数は 44.1 kHz，量子化ビット数は 16 bit である．MoisesDB では，全 240 曲のうち，ディストーションエフェクトが異なる 2 種類のギターの音色（clean electric guitar と distorted electric guitar）のパートが楽曲中に含まれる 32 曲，MedleyDB および MedleyDB 2.0 では全 199 曲のうち，2 種類のギターパートが含まれる 19 曲を使用した．さらに，学習データを増やして汎化性能を向上させるため，2 種類のギターパートを同様のサンプリング周波数と量子化ビット数で録音した，10 曲の自作データも用意した．以上のデータセットを，約 20 秒間の音源に分割し，合計 514 区間の音源について，学習データと評価データの割合を 2 対 1 として交差検証（クロスバリデーション）する．

なお，提案手法の前処理として組み合わせる，楽曲を楽器毎のパートへ分離する手法との親和性を考慮し，今回の

評価では，すでに vocals, bass, drums, others の 4 パートへ完全な音源分離が行えているとの前提のもとで評価する．そのため，提案手法に入力するミックス音源は，vocals, bass, drums のパートを除いた残りすべてのパートが含まれる音源を使用する．そのため，学習データの教師信号は，図 1 中の Guitar 1 としてクリーン（歪み無し）のエレキギターを，Guitar 2 としてディストーションエフェクト（歪み有り）のエレキギターパートを使用する．

提案手法におけるパラメータとして，Prop. 1 と Prop. 2 でそれぞれ，STFT のフレーム長を 2048 と 4096，STFT のフレームシフトを 512 と 1024，BLSTM の隠れ層サイズを 512 と 2048 の計 2 種類で検証を行った．

使用する評価指標は，音源分離で一般に広く用いられる SDR (Signal-to-Distortion Ratio) [7]を採用する．こちらは，値が大きいほど分離精度が良いとされる．真の音源に許容される歪が加わった成分をs_{target}，他の音源からの干渉成分をe_{interf}，センサーノイズ成分をe_{noise}，音源分離によって生じた人工的な歪成分e_{artif}とすると，SDR は以下の式(1)で表される．

$$\text{SDR} = 10\log_{10}\left(\frac{s_{\text{target}}}{e_{\text{interf}} + e_{\text{noise}} + e_{\text{artif}}}\right)^2 \text{[dB]} \qquad (1)$$

4. 音源分離精度評価結果と考察

4.1 評価結果

図 2 に，本研究で得られた clean パートと distorted パートの音源分離結果における提案手法と従来手法の SDR の平均値を示す．なお，図 2 中に従来手法 (Prev.)として示した SDR 値は，文献[3]に記載された値を引用している．この従来手法は，入力を音源波形として，全層畳み込みネットワークである U-Net を 1 次元時間領域に適用したエンドツーエンドの学習モデルである Wave-U-Net を用いている．学習と評価に MedleyDB を使用しているが，内訳は不明なため，clean パートと distorted パート間の差異などを相対比較するための参考値として掲載した．

この図より，提案手法でも従来手法でも，Clean パートよりも Distorted パートの方が分離性能が高いことが分かる．なお，従来手法では Clean パートと Distorted パートの SDR の差が約 8.5 dB と大きかったが，提案手法では，パート間の差が約 3.0 dB となり，ギターの種類にかかわらず，分離性能を維持できていることが分かる．

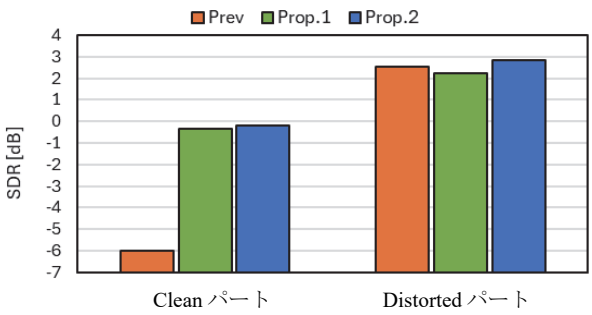

図 2　各パートの音源分離結果

4.2 考察

提案手法において，ギターパートの特徴と分離性能との関係を考察するため，評価に使用する音源を「ギター以外の楽器を含む」，「ギターのみの音源」，「Distorted ギターが単音引きする」，「Distorted ギターが和音引きをする」のように使用する音源をカテゴリ分けし，カテゴリ毎の SDR を算出した．

表 1 に，各カテゴリの音源における Clean パート，Distorted パートの SDR を示す．これより，他の楽器を含む音源の方が，ギターのみの音源よりも約 1.3 dB ほど低い結果となっている．ゆえに，他楽器とギターパートを区別できるようなモデル構築が必要であると考える．また，Distorted ギターが和音で弾いている場合が，単音で弾いている場合よりも約 0.7 dB ほど低い．歪の強さでラベル付けをしたのにも関わらず，このような結果になっているのは，Distorted ギターのデータセット自体が単音で演奏している場合が多かったために，和音で弾いた場合に分離精度が振るわない結果になったと考える．よって，Distorted パートの和音弾きのデータセットの拡充が必要であると考える．

5. まとめ

本研究では，2 つのエレクトリックギターパートの音源分離のための，BLSTM による音源分離手法を検討した．提案手法では，2 つの音色のエレクトリックギターにおいて，両パートとも同等の分離性能を得ることができた．しかし，音源のカテゴリ毎に SDR を評価することで，他の楽器が含まれる場合や，ギターパートが和音で弾かれている場合の性能低下が確認された．そのため今後は，ギターパートの特性に左右されたい分離性能を得るため，学習データとして和音で演奏された音源を増やしたり，RNN の構成を改良してゆく．

参考文献

[1] Music AI Inc., “Moises App: The ultimate practice tool for musicians,” The Musician's App, https://moises.ai/products/moises-app/, (accessed May. 2025)
[2] F.-R. Stöter, A. Liutkus and N. Ito, “Open-Unmix - A reference implementation for music source separation,” 14th Int’l conference on latent variable analysis and signal separation, pp. 293–305, 2018.
[3] 尾関日向, 酒向慎司, “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,” 情報処理学会第 82 回全国大会講演論文集, 1 号, no.5s-02, pp.363-364, 2020.
[4] I. Pereira, F. Araújo, F. Korzeniowski and R. Vogl, “Moisesdb: A dataset for source separation beyond 4-stems,” 2023, arXiv:2307.15913, https://arxiv.org/abs/2307.15913
[5] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," Proc. of 15th ISMIR Conf, 2014.
[6] R. Bittner, J. Wilkins, H. Yip and J. P. Bello, "MedleyDB 2.0: New Data and a System for Sustainable Data Collection," Proc. of the 17th ISMIR Conf., Late Breaking and Demo Papers, 2016.
[7] E. Vincent, R. Gribonval and C. Févotte, “Performance measurement in blind audio source separation,” IEEE Trans. ASLP, vol. 14, no. 4, pp. 1462-1469, 2006.

表 1　音源カテゴリ毎の SDR

音色	Clean	Clean	Distorted	Distorted
楽器構成	他の楽器含む	ギターのみ	単音	和音
SDR [dB]	0.67	2.01	1.63	0.93