歌声音源分離による自動採譜の検討

○田﨑晃基，小坂哲夫（山形大学）

1　はじめに

音楽情報処理の代表的な研究として、自動採譜という技術がある。これは深層学習モデルを用いて楽曲を音響信号から楽譜へと変換する技術であり、音楽情報検索や音楽教育に応用されている。MT3[1]と呼ばれる自動採譜モデルでは、複数の楽器が混ざった状態の音源を入力とし、それぞれの楽器ごとの楽譜を生成するモデルを実現した。一部の楽器に対する採譜は実用的なレベルにまで達している。しかし、歌声音源に対する自動採譜の研究は難しいとされている。理由として、歌唱者の性別や声質などの個人差や、歌唱ジャンルによる歌い方の違いなどが考えられる。一方、一般的に多く耳にするのは歌声を含むバンド演奏であることが多く、楽譜化を求められるのもこの形態の楽曲が多い。

そこで我々の先行研究[2]では、一番需要の多いと考えられる歌声を含むバンド音源から、ピアノ楽譜への変換を試みた。手法としては混合音源から音源分離によって歌声と伴奏に分け、歌声を自動採譜した結果を右手、伴奏をコード認識した結果を左手としてピアノ両手楽譜を生成していた。しかし、音源分離の精度が低かったため、分離した歌声音源に対する十分な採譜結果を得られていなかった。

本研究では音源分離の学習に用いるモデル及びデータを変更し、音源分離の精度を向上させることで、歌声音源に対する採譜精度を向上させることを目的とする。結果では音源分離の精度と自動採譜の精度の比較を行い、音源分離の精度が自動採譜の精度に影響を及ぼしていることを示す。

2　提案手法の概要

本手法の流れを Fig1 に示す。本研究では歌声を含むバンド音源を混合音源と称し、そこから音源分離によって分離した歌声音源を分離歌声音源と称する。混合音源を入力とする音源分離には Band-Split RoPE Transformer(以下、BS-RoFormer)[3]を用いた。ここで混合音源から歌声音源のみを分離し、分離歌声音源が作成される。作成された分離歌声音源は Onsets-and-Frames[4]による自動採譜の入力となり、音源から楽譜へと変換される。また、今回は歌声に対する採譜精度の検討であるため、ピアノ左手楽譜については言及しない。以上がシステム一連の流れである。先行研究では音源分離部分に U-net[5]が用いられていたが、これをより精度が高いとされる BS-RoFormer に変更している。

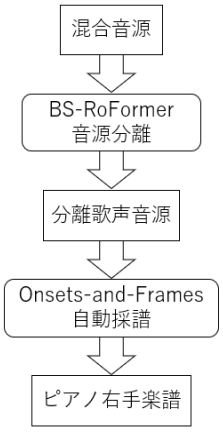

Fig 1 システム概要

2.1 U-net による音源分離

我々の先行研究では音源分離部分に U-net が用いられていた。U-net は生物医学画像分野にて提案された CNN モデルである。入力画像を複数回畳み込み、より深い情報にエンコードする。 その後、アップサンプリングによって元の画像サイズに復元される。また、エンコード時に畳み込みで失われる画像の位置情報をアップサンプリング時に結合するスキップ接続という手法を用いることで、ピクセルのずれを減らし、高品質な画像の復元が可能となっている。これを音源分離に利用し、音源分離時は、目標音源(本研究では歌声)に対するマスクスペクトログラムを作成し、混

*A examination of automatic music transcription using singing voice separation, by TASAKI, Kouki and KOSAKA, Tetsuo (Yamagata Univ.)

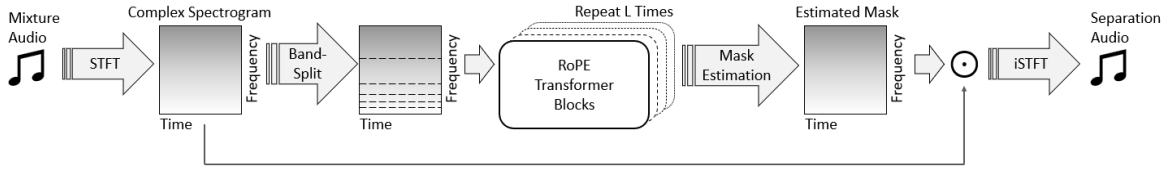

Fig 2 BS-RoFormer システム概要図

合音源のスペクトログラムにかけ合わせることで、音源分離を行う。

2.2 BS-RoFormer による音源分離

BS-RoFormer は音源を短時間フーリエ変換することで得られる複素スペクトログラムを入力として用いる Transformer モデルである。Fig2 は BS-RoFormer のシステム概要図である。特徴として帯域分割(Band Split)と回転位置埋め込み(Rotary Position Embedding)という 2 つの手法を用いている。帯域分割では入力された複素スペクトログラムを周波数方向に分割し、それぞれをモデリングすることで、分離精度を向上させている。また、回転位置埋め込みでは、学習時に正規化された行列の回転行列を用いることで学習の容易化と性能向上を実現している。

2.3 Onsets-and-Frames による自動採譜

Onsets-and-Frames は双方向 LSTM を用いた自動採譜モデルである。Fig3 は Onsets-and-Frames のシステム概要図であり、2 つの検出器から成っている。1 つはフレーム検出器である。これは入力として用いられるメルスペクトログラムをフレーム分割し、それぞれにおいて音の有無と音程を推定するものである。もう 1 つは発音時刻検出器である。基本的に音量が大きく推定しやすいとされている発音時刻に特化した検出器を用いることで、推定しやすい発音時刻をより精密に推定し、音符検出の精度の向上に役立てている。

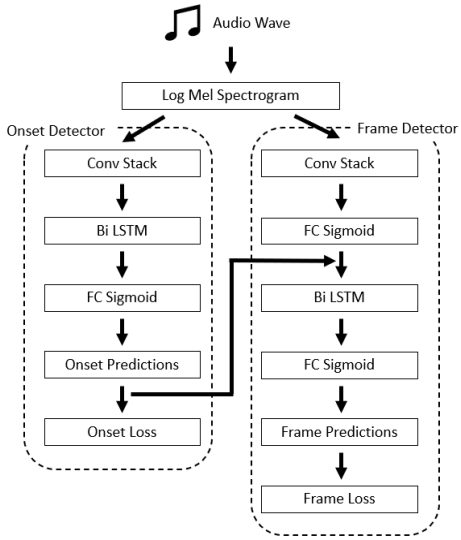

Fig 3 Onsets-and-Frames システム概要図

Onsets-and-Frames ではフレーム検出器の追加入力として発音時刻検出器の出力を用いることで、2 つの検出器の両方で音があると判断されないと音符が出力されないように制限されている。これにより高性能な採譜結果を示している。

3 実験条件

3.1 U-net 実験条件

U-net で使用したデータは Table 1 の通りである。

Table 1 U-net に用いたデータ[曲]

データセット	曲数(曲)
MUSDB18[6]	150
DSD100[7]	100
MedleyDB[8]	60
合計	310

各楽曲は学習の入力に用いられる際に、STFT(短時間フーリエ変換)によるスペクトログラムへの変換が行われる。音源分離時はマスクスペクトログラムをかけられた後、ISTFT(逆短時間フーリエ変換)によって再び音源へと戻される。音源分離後は分離歌声音源とそれ以外の分離音源が生成されるが、今回は分離歌声音源のみ用いている。

3.2 BS-RoFormer 実験条件

BS-RoFormer で使用したデータは Table 2 の通りである。先行研究と違うデータを用いているのは、用いた楽曲に被りが見つかったためである。そのため、被り楽曲を含む DSD100 と MedleyDB を外し、減ってしまった曲数を

Table 2 BS-RoFormer に用いたデータ[曲]

データセット	曲数
MUSDB18	100
MoisesDB[9]	235
合計	335

カバーするために新しく MoisesDB を用いた。各個別音源は曲ごとに歌声音源の無音区間に合わせた処理を施している。

学習時には「異なる曲の異なる秒数から始まる 8 秒を無作為に抽出する作業」を各ステムで行い合成することで混合音源を作る、ランダムミックスというデータ作成法を使用している。これは、あえて音楽的に破綻している音源を作り出し学習させることで、難関な音源分離を可能とし、分離精度の向上に役立つとされている。ランダムミックスの有用性は参考文献[10]で報告されている。

音源分離の評価には Mir-eval[11]で算出される SDR を用いている。算出式を(1)に示す。

$$\mathrm{SDR} = 10\log_{10}\frac{\| s \|}{\| s - \hat{s} \|}\ [dB] \quad \cdots \ (1)$$

ここで、s は目標音響信号であり、$\hat{s}$は分離音源の音響信号である。これは分離された音源が目標音源にどの程度近いのかを数値で示したものであり、数値が高いほど分離精度が高いことを示している。本研究ではこの指標を用いて音源分離の精度を示す。

3.3 Onsets-and-Frames 実験条件

Onsets-and-Frames で使用したデータは Table 3 のとおりである。

Table 3 Onsets-and-Frames に用いたデータ[曲]

データセット	学習	検証	評価
東北きりたん歌唱 DB[14]	42	5	3
夏目悠李男性歌唱 DB[15]	43	5	3
No.7 歌唱 DB[16]	43	5	3
合計	128	15	9

本研究では、WAV と MIDI の整合性が取れるように DAW ソフトでの無音区間処理を行っている。また評価楽曲は各歌唱データベースの楽曲に RWC 著作権切れ音楽データベース[12]の伴奏を合成させることで混合音源を作成し、それらを音源分離することで分離歌声音源を作成している。

採譜の評価は[13]で定義されたフレーム単位での F 値評価と、音符単位での F 値評価を行う。フレーム評価では 10ms のフレーム分割を行い、それぞれのフレームで予測と正解を比較し評価を行う。音符単位での評価は 2 種類ある。1 つは Note 評価であり「音符の発音時刻が正解の±50ms 以内にありかつ音程があっている」場合のみ予測が正しいとする。もう 1 つは Note w/offset 評価であり Note 評価の条件に加え消音時刻を考慮し、「音符の長さが正解の 20%以内、または 50ms 以内のいずれか大きい方」の場合のみ予測を正しいとする。本研究ではこの 3 つの評価指標を用いて、歌声自動採譜の精度を示す。

4 実験

4.1 実験結果

先行研究と提案手法の結果の比較を Table 4 に示す。これは評価データ 9 曲に対する平均値である。また、ベースラインは評価データに分離していない歌声音源を用いた場合の数値である。そのため SDR は算出されない。

Table 4 自動採譜結果[%]

	SDR[dB]	Frame	Note	Note w/offset
ベースライン	-	81.61	81.71	54.88
先行研究	11.25	67.34	62.98	34.79
提案手法	**20.85**	**80.24**	**78.57**	**52.64**

最初に音源分離の結果であるが、先行研究での歌声音源分離結果の SDR が 11.25 dB だったのに対し、提案手法での結果は 20.85 dB であった。そのため、音源分離の精度向上は達成できたと考えられる。次に自動採譜の結果であるが、すべての評価指標において先行研究よりも提案手法のほうが大幅に精度が向上していることがわかる。ベースラインと比較しても、提案手法のほうがかなり近い値まで採譜精度が向上していることがわかる。よって、これらのことから分離歌声音源に対す

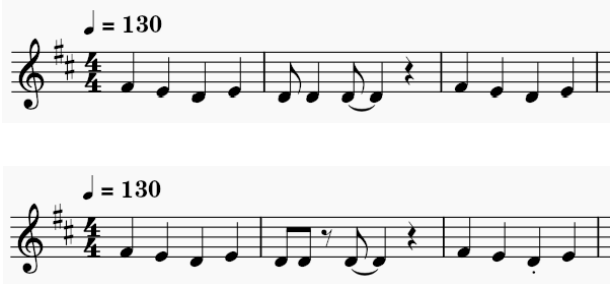

Fig 4 正解楽譜(上)と提案手法の採譜結果(下)の比較

る採譜精度には分離精度が大きく関係し、SDR 評価による分離精度の値を向上させることが採譜精度の向上につながることが示された。また、Fig4 は正解の楽譜と本研究の提案手法による採譜結果の比較である。評価楽曲の中から一曲冒頭部分のみ示している。見比べてみると、すべての音符の発音時刻や音程があっていることがわかる。

4.2 考察

なぜ分離精度が向上したことで採譜精度が向上したのかについて考えられる要因がある。それは、分離精度が低いと他楽器の音の入り込みが多く認識の阻害要素になり、正しく音符が検出できないためである。実際に予測された音符数を算出すると、Grand Truth は 1742 個に対し、先行研究での予測結果では 1600 個、提案手法での予測結果では 1660 個であった。よって音源分離の精度が高いと音符検出のタスクが行いやすくなり、正しく音符を予測できるため採譜精度が向上したと考えられる。

5　まとめ

本研究では先行研究から歌声音源分離部分における実験条件の変更を行い、歌声分離精度の向上による自動採譜精度の向上を検討した。結果として、自動採譜における 3 つの評価指標すべてにおいて、先行研究よりも大幅に精度が向上した。今後の予定として、本研究では評価楽曲における歌唱者をクローズで行っているため、オープン歌唱者に対する評価実験を行いたいと考えている。また、自動採譜モデルの性能向上も検討し、分離歌声音源に対してより精度の高い自動採譜を行うことのできるモデルの作成を目指す。

参考文献

[1] Josh Gardner, *et al.*, “MT3: Multi-Task Multitrack Music Transcription”, arXiv preprint arXiv:2111.03017v4, 2022

[2] 千葉綾乃, 他, “歌声音源を用いた深層学習による自動採譜の検討”, 情報処理学会第 86 回全国大会, 2024

[3] Wei-Tsung Lu, *et al.*, “Music source separation with Band-Split Rope Transformer”, arXiv preprint arXiv:2309.02612v2, 2023

[4] Curtis Hawthorne, *et al*, ”Onsets and Frames: Dual-Objective Piano Transcription ”,arXiv preprint arXiv:1710.11153, 2018.

[5] Andreas Jansson, *et al.*, “ Singing Voice Separation with Deep U-Net Convolutional Networks”, Proceedings of the 18th ISMIR Conference, pp. 23-27, 2017.

[6] Zafer Rafii, *et al.*, “MUSDB18-hq”, https://sigsep.github.io/datasets/musdb.html

[7] Liutkus, *et al.*, “The 2016 Signal Separation Evaluation Campaign” Latent Variable Analysis and Signal Separation, Proceedings of 12th International Conference, LVA/ICA, pp. 25-28, 2015.

[8] R. Bittner, *et al.*, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research", 15th International Society for Music Information Retrieval Conference, 2014.

[9] Igor Pereira, *et al.*, “MoisesDB: A Dataset for source separation beyond 4-stems”, arXiv preprint arXiv:2307.15913v1, 2023

[10] Chang-Bin Jeon, *et al.*, “Why does music source separation benefit from cacophony?”, arXiv preprint arXiv:2402.18407v1, 2024

[11] Colin Raffel, *et al.*, “mir eval: A transparent implementation of common mir metrics.” In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR. Citeseer, 2014.

[12] “RWC 研究用音楽データベース, ”国立研究開発法人産業技術総合研究所, https://staff.aist.go.jp/m.goto/RWC MDB/index-j.html

[13] Mert Bay, et al., “Evaluation of multiple-f0 estimation and tracking systems.” In ISMIR, pages 315-320, 2009.

[14] SSS 合同会社, 森勢将雅:研究者向け音声合成検証用東北きりたん歌唱データベース, access: 2023-11-30, https://zunko.jp/kiridev/login.php

[15] 歌声 DB 制作:アマノケイ, 音声提供者: 霧野蒼太, access: 2023-11-30, https://ksdcm1ng.wixsite.com/njksofficial

[16] No.7 製作委員会, 研究者向け音声合成検証用 No.7 歌唱データベース, access: 2023 11-30,