1V-07

歌声音源を用いた深層学習による自動採譜の検討

千葉 綾乃† 小坂 哲夫†
山形大学大学院理工学研究科†

1 はじめに

本研究の目標は，歌声を含む楽曲音響信号から歌唱部を主旋律として演奏するピアノ楽譜を生成する自動採譜システムの構築である．本稿は，LSTM ベースの深層学習モデルを使用し，歌声採譜の検討を行う．従来，自動採譜研究[1-2]では，ピアノなど楽器に対する採譜は広く研究されており，一部の生成された楽譜は実用的なレベルにまで到達している．対して，歌声に対する自動採譜研究はあまり多くない，一方，ユーザが求める楽譜の多くは，単体楽器の音源のみからなる楽曲よりも，様々な楽器を含む音源の楽曲，特に歌声を含む音源である．本稿では以下 mix 音源と呼ぶ．本研究ではピアノ楽曲に対する自動採譜モデルを基にし，歌声の自動採譜を行う．Magenta[3]の Onsets and Frames[4]を使用し，歌声を学習させることで歌声採譜モデルとして活用する．また，歌声の抽出による性能を比較するために，音源分離手法を用いて作成した分離歌声音源を使用する．実験では，作成したモデルの基本的な能力の検証と音源の条件の違いによる性能の比較を行う．

2 提案手法の概要

本手法の流れを図 1 に示す．入力は歌声を含む mix 音源で，これを U-NET による音源分離法[5]を用いて歌声音源と伴奏音源に分離する．その後特徴量としてメルスペクトログラムを音源から抽出し，得られた特徴量を採譜モデルの入力とする．採譜モデルには，自動採譜モデル Onsets and Frame を用い双方向 LSTM[6]で採譜を行う．その後生成された midi 楽譜と正解 midi 楽譜を比較し，採譜率を算出する．

2.1 Onsets and Frames による自動採譜

Onsets and Frames はピアノを対象とした自動採譜モデルであり，音符検出タスクに，音符の発音時刻検出器とフレーム検出器の 2 つの検出器を用いている．また，発音時刻フレームの重要性に重点を置いており，発音時刻検出器の出力をフレーム検出器の追加入力として使用し，発音時刻検出器がそのフレーム内に発音時刻が存在していると判断した場合のみ音符を開始するようにモデルの最終出力を制限することで，高性能な採譜精度を示している．ただしピアノ以外の音源に対する性能は明らかではない．

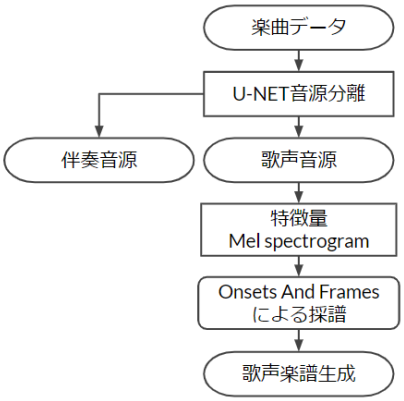

図 1 システム概要.

3 データセットの作成

本研究で使用するデータを表 1 に示す.

表 1 使用データ詳細.

データセット	学習曲数	評価曲数	元音源の状態
東北きりたん歌唱DB	47曲	3曲	歌声のみ
夏目悠李男性歌声DB	48曲	3曲	歌声のみ
No.7歌唱DB	48曲	3曲	歌声のみ
RWCポピュラーMDB	100曲	0曲	歌声+伴奏
RWC著作権切れMDB	0曲	9曲	歌声+伴奏
東北イタコ歌唱DB	0曲	9曲	歌声のみ
「波音リツ」歌唱DB	0曲	9曲	歌声のみ

データの内，RWC 音源は歌声と伴奏からなる mix 音源，その他は歌声のみの音源である．さらに mix 音源の種類を増やすため RWC の伴奏部分と歌声を混合し mix 音源としたものも評価データとして使用した．

4 実験・評価

モデルの評価は，[4]で定義された音符単位での F 値評価とフレーム(Frame)単位での F 値評価の 2 種類の指標を用いる．また，音符単位での評価では「消音時刻を考慮せず，発音時刻が ground-truth の±50ms 以内にある」場合(Note)と「消音時刻を考慮し，音長が ground-truth の 20%以内または 50ms 以内のいずれか大きい方」の場合(Note

A study of automatic music transcription using deep learning with singing sound source
†Ayano CHIBA and Tetsuo KOSAKA, Yamagata University

w/offset)の2種類の音符基準で測定される．

4.1 歌声用学習モデルの作成

事前実験として，[4]のピアノ用採譜モデルに対し，歌声音源で自動採譜を行った．しかし，[4]のピアノ音源に対する結果 Frame 評価 84.92%/Note 評価 86.44%と比較し，今回の採譜結果は Frame 評価 44.2%/Note 評価 21.82%と主旋律以外の音が多く挿入されており低い結果であった．原因として，ピアノと歌声の音響特徴が異なることが考えられる．そこで，歌声学習データを用いモデルを作成した．モデルは，東北きりたんと夏目悠李，No.7 学習用音源で学習した KN7 モデル，RWC 学習用音源から U-NET により分離した歌声を KN7 に追加した KN7+RVC モデルの2種類を作成した．

4.2 分離音源に対する評価

歌声のみの評価データおよび mix 音源から歌声を分離した評価データの比較を行う．モデルとしては KN7 および KN7+RVC の両モデルを使用する．まず，東北きりたんと夏目悠李，No.7 評価用音源 9 曲を用い，歌声のみの評価音源 KN7_vocal を用意した．さらに KN7_vocal に RWC 評価音源の伴奏を人工的に加算し mix 音源を作成し評価データとした(KN7_mix)．こちらに対しては歌声を U-NET で分離する．実験結果を表2に示す．

表2 学習モデル・評価音源毎の評価結果．

学習モデル	評価音源	Frame	Note	Note w/offset
KN7	KN7_vocal	81.38 %	80.94 %	53.55 %
KN7	KN7_mix	69.66 %	60.01 %	33.42 %
KN7+RVC	KN7_vocal	**81.63 %**	**81.71 %**	**55.72 %**
KN7+RVC	KN7_mix	60.72 %	51.91 %	30.26 %

結果，ピアノ音源を学習で用いた場合よりも大幅に性能が向上した．また，同一モデルに対する評価では，mix 音源よりも歌声のみで構成された音源の方が高い性能を示し，また学習モデルに関して，mix 音源に対する評価を行う場合，歌声分離を行ったデータを含めない方が良い結果となることが確認できた．以上より分離した歌声でも採譜は可能であるが歌声のみの音源よりも性能が低下することが分かった．

4.3 歌唱者および音量に関する検討

前節の実験は，音源データはオープンであるが，歌唱者に対してはクローズな条件であった．歌唱者オープンの場合の評価を行うため，東北イタコ(Itako)と波音リツ音源(Namie)を評価データとして用いた．また前節で分離した歌声での性能低下の理由を検討するため，歌声と伴奏の音量差による性能の変化を調査した．今回は歌声対雑音の SN 比が-5.0dB，0dB，5.0dB の mix 音源を人工的に加算することで作成し，これに加え歌声単体での評価を SN 比=∞として比較した．認識モデルとしては KN7 を用いた．結果を表3に示す．

歌唱者に関する検討では歌唱者により性能が大きく変動することが分かった．Namie は歌唱者クローズに近接する結果だったことから，オープン/クローズより話者による影響が大きいことが分かった．また SN 比が性能に大きく影響することが確認できた．作成した mix 音源を確認すると，調整後も伴奏の音が残っていたり，伴奏が歌声よりも大きいタイミングがあった．これにより，伴奏の音を拾ってしまい余計な音符が挿入されたり，分離時により歌声が欠落し十分な認識ができなかったと考えられる．

表3 歌唱者・SN 比評価音源毎の評価結果．

評価音源	SN比(dB)	Frame	Note	Note w/offset
Itako	∞	**60.45 %**	**44.26 %**	**29.36 %**
Itako mix	5	54.00 %	40.62 %	23.61 %
Itako mix	0	50.02 %	38.32 %	20.33 %
Itako mix	-5	31.06 %	26.48 %	10.86 %
Namine	∞	**79.36 %**	**73.86 %**	**52.94 %**
Namine mix	5	71.15 %	63.64 %	39.14 %
Namine mix	0	58.48 %	51.37 %	27.33 %
Namine mix	-5	37.97 %	31.62 %	13.42 %

5 おわりに

本稿では，mix 音源に対する歌声自動採譜の検討を行った．実験の結果，分離音源を用いた採譜は歌声のみの音源に対する採譜より難しいことが確認でき，SN 比に大きく影響されることが分かった．今後の検討課題として，音源分離の性能向上が上げられる．

参考文献

[1]K. Shibata, et al., “Non-local musical statistics as guides for audio-to-score piano transcription,” Information Sciences, Vol.566, pp.262-280, 2021.
[2]S. Bock, et al., “Polyphonic piano note transcription with recurrent neural networks,” Proc. Of ICASSP2012, pp.121-124, 2012.
[3]”Magenta”, https://magenta.tensorflow.org
[4]C. Hawthorne, et al., ”Onsets and frames: dual-objective piano transcription”, arXiv preprint arXiv:1710.11153, 2018.
[5]A. Jansson , et al., “Singing voice separation with deep U-Net convolutional networks,” Proc. of ISMIR2017, pp. 23-27, 2017.
[6]M. Schuster and K.K. Paliwal, “Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), pp. 2673-2681, 1997.