ボーカル音源を対象とした楽曲のコード認識の検討

○北堀和佳，小坂哲夫（山形大学）

1　はじめに

近年，音楽視聴サービスや動画視聴サービス等の浸透によって，音楽に触れる機会が著しく増えている．また，それらのサービスにより，ユーザーは楽曲を気軽に投稿できるようになった．一方，投稿される楽曲の幅が広がることから，様々な楽曲に対応可能なシステムが必要とされる．

これまで，楽曲の雰囲気を担うとされるコードの研究が行われてきた．コードとは，音楽の三要素（メロディ，ハーモニー，リズム）のうちのハーモニーに当たる複数音の協和である．深層学習を用いてコードの類似性を分析することで楽曲の分類が可能であると考えられる．先行研究において，DNN を用いたコード認識[1]では，通常音源だけでなく，打楽器分離やボーカル分離を組み合わせたコード認識精度向上を検討している．文献[2]ではそれまで RNN や LSTM，CNN を組み合わせたコード認識から bi-transformer に変更することで，高い精度のコード認識が報告がされている．しかし，この文献で使用されている音源データは，複数トラックを一つにまとめた MIX 音源でのみの学習，評価であったため，単一楽器のメロディからのコード認識は行われていなかった．単一楽器やボーカルのメロディは，複数の音を同時に出せないため，ある時間幅の連続的なメロディからコードとする分散和音（アルペジオなど）を抽出し認識することが必要である．しかし，従来モデルでは MIX 音源のみでの学習であるため，単一楽器やメロディのコード認識が困難であることが予想される．

そこで，事前実験として，文献[2]のモデルにボーカル音源（3.1 にて説明）でコード認識を行ったところ，MIX 音源の認識率に比べ，認識率が低い結果となった．このことから，従来モデルには，ボーカルのような単一メロディのコード認識が困難であると考えらえる．

本研究では，従来の MIX 音源によるモデル学習ではなく，ボーカル音源でモデル学習を行うことで，ボーカル音源に対するコード認識精度の向上を目的とする．これによって，分散和音に対応可能であるかの検討を行う．また音源によるコード認識の違いを明らかにすることで，幅広い音楽のコード認識が可能か検討する．

2　コード認識手法の概要

本手法の流れを Fig. 1 に示す.

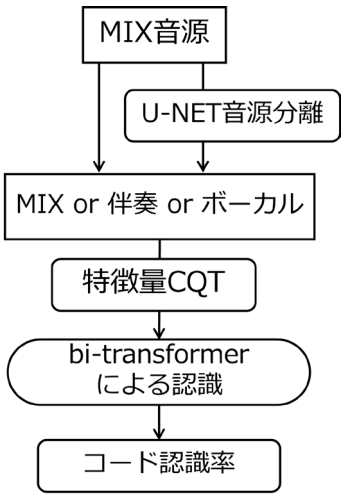

Fig. 1 システム概要

本研究では，MIX 音源，伴奏音源，ボーカル音源を使用する．ボーカル音源と伴奏音源を作成するにあたり U-NET による音源分離を用いた．その後，特徴量 CQT(Constant-Q transform)を音源から抽出し，得られた特徴量を入力とした bi-transformer によるコード認識を行う．その後，得られた予測コードと正解コードを比較し，コード認識率を算出する．以上がシステム一連の流れとなる．

---

* A study of chord recognition for singing voice，by KITABORI, Kazuyoshi and KOSAKA, Tetsuo (Yamagata Univ.)

3　システム構成

3.1 U-NET 音源分離

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，より小さくて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する．U-NET は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．文献[3]では，U-NET を音源分離に利用し，MIX 音源のスペクトログラムから目標とする単一音源（この研究ではボーカル音源か伴奏音源）のマスクを生成し，MIX 音源に被せることで音源分離を行う．この手法は従来の音源分離より高品質な音源分離を達成した．

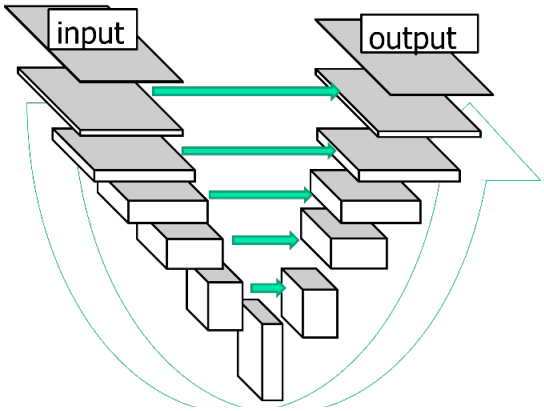

Fig. 3 U-NET アーキテクチャ

3.2 特徴量 CQT(constant-Q transform)

CQT[4]は，対数周波数を利用することで音高を捉えることのできる特徴量である．短時間フーリエ変換 STFT (short-time Fourier transform)では分析窓が固定なため，高周波数に対し情報が過多，低周波数に対して情報が不足するという問題があった．CQT は分析窓を周波数ごとに可変にすることで STFT の問題を解決している．これにより周波数分解能と時間分解能のバランスが取れた特徴量となっている．

3.3 Bi-transformer によるコード認識

従来，コード認識では DNN や CNN，RNN を用いていたが，transformer による前後関係を考慮した学習は，コード認識においても有効であった[2]．そこで，本研究でも Bi-transformer を用いてコード認識を行う．Bi-transformer 層の構造を Fig. 3 に示す．Masked Multi-head attention では, 入力をクエリ, キー,

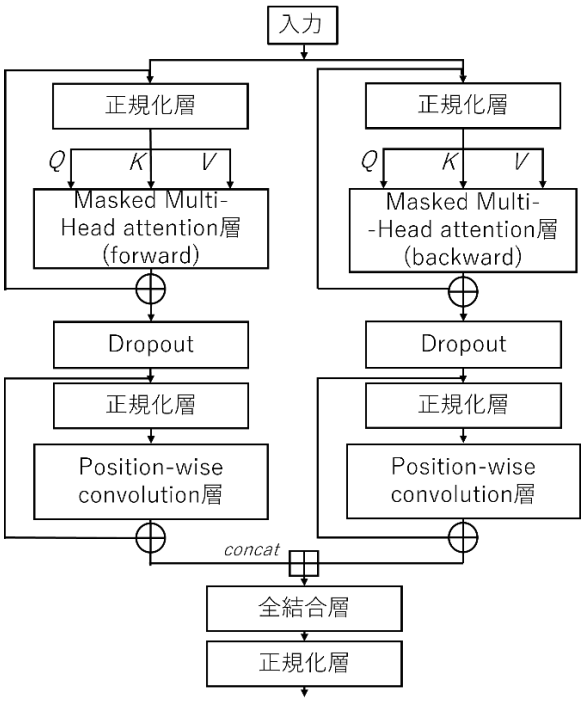

Fig. 2 Bi-transformer の構造

バリュー（q，k，v）に線形変換し，内積からフレーム毎の関係性を計算する．また，それらを複数ヘッドに分割し結合することで，よりフレーム同士の関係性を考慮することができる．また，マスクによって全体のフレームとの関係性ではなく，そのフレーム以前との関係性だけを計算している. これを forward と backward の双方向から行い，前後関係を計算する．position-wise convolution 層は，コード遷移を滑らかにするために用いられており, 1 次元の畳み込みブロックで構成されている．Forward と Backward の出力を結合し, 全結合，正規化を行ったのち，softmax 層へと出力される．softmax 層で 128 次元の尤度を 25 次元の尤度へ線形変換し，25 種類のいずれかのコードを決定する．その後，正解ラベルと比較され，その際の損失関数はクロスエントロピーを用いる．

4　コード認識実験

本手法で用いる音源データは，大きく分けて二種類あり，U-NET の音源分離モデルで使用する音源データとコード認識に使用する音源データである．U-NET で使用する音源は，MIX トラックに加え，ボーカル，ベース，ド

ラムなどの個別トラックが用意されている．コード認識に用いる音源データは，時間ごとのコードラベルが用意されている．

4.1 実験条件

4.1.1. U-NET の実験条件

U-NET で使用したデータは，DSD100[8]：100 曲，MUSDB18[9]：150 曲(うち 50 曲を検証データとした)，MedleyDB[7]：122 曲の合計 372 曲(23.04 時間)をサンプリングレート 44.100Hz，ステレオチャンネル WAV 形式で MIX 音源，ボーカルトラック，その他音源トラックの三種類を使用する．

音源の時間―周波数変換には librosa[5]の STFT を利用し，分析窓 1024，ホップサイズ 512，バッチサイズ 64，パッチ長 128 とした．得られた次元数 512，時間フレーム 128 のスペクトログラムをネットワークの入力とする．ネットワークパラメータは，カーネルサイズ 4，ストライド 2，パディング 1，チャンネル数は 1, 32, 64, 128， 256， 512 とした．エンコーダには活性化関数 leakyRelu，デコーダには活性化関数 Relu を使用し，最終層には sigmoid 関数を用いている．得られたマスクを混合音源のスペクトログラムに被せ，得られたスペクトログラムを逆短時間フーリエ変換 ISTFT（inverse short-time fourier transform）することで音源の復元を行った．

4.1.2. CQT の実験条件

CQT 抽出には librosa[5]を用いている．1 オクターブあたり 24 bin，ホップサイズ 2048，C1 から C7 の 6 オクターブについて，計 144 次元の CQT 特徴量を抽出した．実験では，音源を 10 秒ごとに入力とし，5 秒スライドさせ，CQT 特徴量を取得している．

4.1.3. Bi-transformer の実験条件

コード認識実験の楽曲データは，ビートルズ：180 曲，キャロルキング：7 曲，クイーンズ：19 曲，Zweieck：18 曲，RWC[6]：100 曲，BillBoard のヒットチャート：13 曲の計 337 曲を使用し，学習データ 317 曲，評価データ 20 曲とした．サンプリングレートは 22.050Hz，モノラルチャンネルで使用する．また，pyrubberband を利用し， 学習データに対して -5 度から +6 度までピッチ拡張を行った．コードの種類は，C から B までの 1 オクターブ 12 音それぞれにおけるメジャー/マイナーコードの 24 種類に無和音 N を加えた計 25 種類とした．bi-transformer の学習パラメータは，文献[2]を参考にし，レイヤー繰り返し数：8，self-attension heads：4， Q,K,V の次元数 と 隠れ層ノード数：128，Position-wise convolutional block の繰り返し数 2，カーネルサイズ：3，ストライド：1，パディング：1，各 Dropout を 20%とした．

4.2 実験結果と考察

4.2.1. 各音源モデルの認識結果と考察

本実験では，従来法の MIX 音源モデルと提案法のボーカル音源モデルの音源毎のコード認識性能を比較し，ボーカル音源の認識性能を明らかにする．ボーカル音源モデルは，MIX 音源を U-NET で分離し，得たボーカル音源から CQT を抽出，bi-transformer へ入力して学習することで生成した．従来法の MIX モデル，提案法のボーカルモデル，U-NET で分離した伴奏音源で学習したモデルの三種に MIX 音源，伴奏音源，ボーカル音源の三種を評価した．table 1 に今回の実験結果を示す．

table 1．音源毎の認識結果

学習＼評価	MIX	伴奏	ボーカル
MIX	79.51	79.28	48.33
伴奏	79.05	80.31	36.21
ボーカル	77.21	77.02	**70.44**

table 1 より，ボーカル音源の評価において提案法は従来法に比べ大幅に改善された．また，ボーカル音源のみの学習でコード認識率が約 70%であることから，分散和音でのコード認識が可能であると考えられる．

MIX 音源モデルと伴奏音源モデルを比較すると，MIX 音源と伴奏音源の認識率がほぼ変わらないのに対し，ボーカル音源の認識に約 12%の差があることから，音源にボーカルが含まれることで，認識性能が向上すると考えられる．

提案法は，MIX 音源や伴奏音源の認識率が高くなっているが，分散和音だけで同時和音をここまで認識できるとは考えにくい．これは，ボーカル音源に伴奏が少量入り込んでいることが原因だと考えられる．正確な認識を行うには，ボーカル分離精度を高める必要がある．

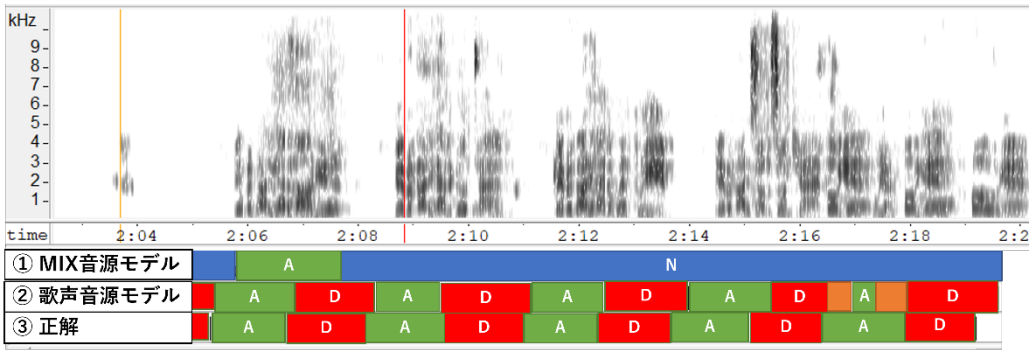

Fig. 4 ボーカル音源における 2 種のモデルの認識結果の比較

4.2.2. 予測ラベルの考察

Fig. 4 に本実験から得られた評価データの予測ラベルを示す．ラベルはそれぞれ，①MIX モデルを用いボーカル音源を評価したときの予測ラベル，②ボーカル音源モデルにボーカル音源を評価したときの予測ラベル，③正解ラベルとなっている．ボーカル音源で学習することで，MIX 音源で無音と予測されていた区間が正解ラベルと一致していることが確認できた．これは，MIX 音源モデルのような同時和音データでの学習であったため，ボーカル音源のような単一メロディを認識するのが難しかったからだと考えられる．

また，今回の実験ではボーカル音源を使用する学習においても MIX 音源のラベルを使用している．従って演奏音は存在するがボーカルは存在しない区間にもなんらかのコードラベルが付与されており，不正確な学習が行われている可能性がある．よって今後はボーカル音源に対しても正確なラベルを付与し再学習する必要がある．

5 まとめ

本研究では，学習データを MIX 音源からボーカル音源に変更することで，ボーカル音源に対するコード認識精度向上の検討を行った．結果としてボーカル音源で学習を行うことで，ボーカル音源に対して高いコード認識率を得られることが明らかになった．一方，U-NET によるボーカル分離が不十分であることや，ボーカル音源に対応した正解ラベルが必要となるとがわかった．今後，これらを改善するとともに，様々な音源のコード認識が可能なシステムの構築を目指したい．

参考文献

[1] 稲葉，他，“ディープニューラルネットワークを用いたコード認識の性能向上の検討”，第 2 回 東北地区音響学研究会 2019.

[2] Jonggwon Park, et al.,“A BI-DIRECTIONAL TRANSFORMER FOR MUSICAL CHORD RECOGNITION”, Proceedings of the 20th ISMIR Conference, pp. 4-8, 2019.

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

[4] Judith C. Brown,“Calculation of a constant Q spectral transform”, The Journal of the Acoustical Society of America vol.89 No.1 1991.

[5] “librosa” https://librosa.github.io/librosa/feature.html#

[6] “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

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

[8] 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.

[9] Rafii, et al., “ The MUSDB18 corpus for music separation ”, December, 2017. https://doi.org/10.5281/zenodo.1117372