6T-01

バンド編曲に向けたギター音源からベース音源を生成するCNNモデル

香西 智雄† 北原 鉄朗†
† 日本大学文理学部情報科学科

1. はじめに

ギターは，軽音楽において中心的な楽器の1つである．そのため，多くのアマチュアギタリストが存在し，ギターを弾きながら作曲を楽しむ者も少なくない．しかし，作曲した楽曲をバンドで演奏するには，ベースやドラムスなど各楽器パートの演奏内容を決める作業（編曲）が必要になる．編曲を行うには，各楽器の特性を知る必要があったり，編曲した結果を人に伝えるには楽譜に書くか DTM (desktop music) を用いる必要があるため，簡単にできるものではない．

本研究が目指すのは，ユーザが作曲した楽曲の伴奏がギター1本で与えられたときに，他の楽器パートの演奏内容を自動で決めて、バンドで演奏できるようにするシステムの実現である．他の楽器パートとしては，ベース，ドラムス，キーボードなどが考えられるが，本稿ではベースのみを扱う．

従来の自動編曲は，ピアノを対象とするもの[1)2)]が多かった．ギターを対象とする研究[3)]も存在するが，ギターの音からドラム，ベースなどの音を付与するバンド編曲を行う研究をしているものはなかった．

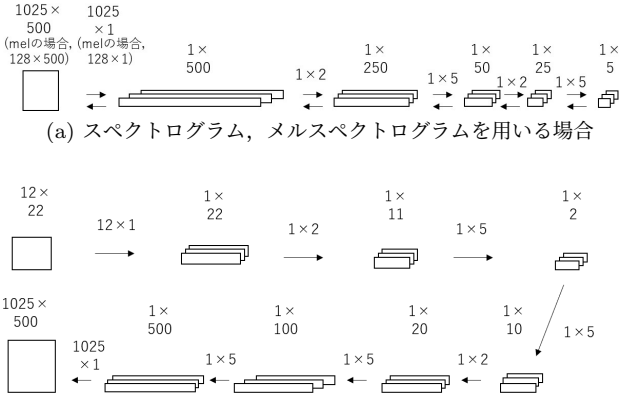
そこで，本稿はギター音源入力から，ベース音源を自動生成するモデルを提案する．ここで，オーディオ音源からオーディオ音源を返すようにしたのは，世の中には midi ギターが存在するが，弾いた音を midi に変換する精度が不十分であることが多く，対象とするユーザーが，ギターに関する知識しか持たないことを想定しているため，入力を midi などで作成する必要もなく，出力を楽譜に起こしても，読めない可能性があるため，この入出力の形にした．

2. 提案手法

本手法では，畳み込みニューラルネットワーク（CNN）を用いてギター音源からベース音源を生成する．この手法では，ギター音源とベース音源のペアデータが学習用に与えられることを前提とする．ギター音源を 0.5 秒ごとに区切り，そのスペクトログラムを CNN の畳み込み層に入力し，逆畳み込み層に与えることでベース音源のスペクトログラムを得る．

2.1 入力音源のスペクトログラムの計算

ギター音源（学習時はベース音源も）に対して，短時間フーリエ変換を用いてスペクトログラムを計算する．まず，サンプリング周波数を 22050Hz にダウンサンプリングする．次に，短時間フーリエ変換を行う．窓関数は hann 関数，窓幅は 2048，ホップサイズはサンプリング周波数の 1/1000 とした．その後，各値の絶対値を取ることで振幅スペクトログラムを得る．



(a) スペクトログラム，メルスペクトログラムを用いる場合

(b) クロマグラムを用いる場合

図1 CNN モデルのアーキテクチャー．長方形の上の数値はデータの形状，矢印の上の数値はフィルタの形状を表す．右向きの矢印は畳み込み層，左向きの矢印は逆畳み込み層である．

2.2 入力音源に対する特徴抽出

入力音源のスペクトログラムを次項で述べる CNN に入力する他，スペクトログラムから特徴抽出したものを CNN に入力する方法も試行する．抽出する特徴として，メルスペクトログラムおよびクロマグラムを用いる．ただし，クロマグラムに関してはホップサイズを 512 とした．

2.3 CNN によるベース音源のスペクトログラムの生成

前項までの方法で得たギター音源のスペクトログラム（またはメルスペクトログラム、クロマグラム）を畳み込み層で圧縮を行い，その後，逆畳み込み層を適用することでベース音源のスペクトログラムに変換する．モデルの概要を図1に示す．入力されるデータは形状が 1024 × 500（周波数軸：1024 要素，時間軸：500 要素）であり（メルスペクトログラムのの場合は 128 × 500，クロマグラムの場合は 12 × 22），そこから図1の畳み込み層によって 1 × 5（クロマグラムの場合は 1 × 2）に圧縮したのち，逆畳み込み層によって 1025 × 500 のスペクトログラムに変換する．各層におけるフィルタのチャネル数は 1024 とし，ストライドは 1，パディングはなし，活性化関数は ReLU とした．

2.4 位相復元による音響信号の生成

出力されたベース音源のスペクトログラムに対して逆フーリエ変換および位相復元を行うことで，ベース音源の音響信号を得る．位相復元には Griffin-Lim アルゴリズム法を用いる．反復回数は 32，窓幅は 2048，ホップサイズはサンプリング周波数の 1/1000 とした．その後，HPSS（調波打楽器音分離）を用いて，打楽器音の分離をすることで，突発的な雑音を除去する．margin に関しては最小値を 1.0，最大値を 1.5 とする．抽出された調波楽器音を量子化ビット数を 16，サンプリング周波数を 44100Hz として wav 音源に変換する．

A CNN model that generates bass sounds from guitar sounds towards band arrangement *by Tomoo Kouzai and Tetsuro Kitahara(Nihon University)*

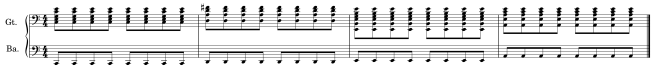


図 2 作成したギター，ベースのスコアの例

3. 実　　験

上で述べた手法に基づいて適切なベース音源が生成できるかどうかを次の実験により検証した．

3.1 データセット

Cakewalk by BandLab を用いてギターおよびベースパートの MIDI シーケンスを入力し，ソフトウェア音源を用いて wav 形式に変換した．BPM は 120，小節数は 4 小節（8 秒）とした．ギター音源には sforzand，ベース音源には Cakewalk by BandLab に付属する SI-Bass Guitar を用い，どちらにもエフェクターは適用しなかった．コード進行は 1 小節あたり 1 コードとし，メジャースケールに基づいて決定した．ベースは各コードのルート音とした．ギターおよびベースパートのリズムは八分音符とした．この基準に基づいてギター音源およびベース音源のペアを 16 個作成した．また，16 個のうち 4 個のコード進行に関してヴォイシングを変えたギター音源を 4 個作成した．一例を図 2 に示す．このうち，10 個を学習用に，10 個をテスト用に割り当てた．

3.2 実 験 条 件

入力音源の音響的特徴が学習データからどの程度異なっていても適切にベース音源の生成されるか検証するため，次の 3 つの条件を設定した．

条件 1　学習データとテストデータとでコード進行が異なるか，ヴォイシングが異なる．用いるソフトウェア音源や音響的条件が同じ。

条件 2　条件 1 に対してテストデータにローパスフィルタ（設定：1 オクターブ上がるごとに −3dB する）をかけた．

条件 3　学習データは前節の方法で作成したものであるが，テストデータは第 1 著者が本物のギターで演奏したもの．演奏の録音には M-Audio 社の M-Track を用いた．

本来であれば，ギターやベースパートのリズムなどにも変化を付けるべきであるが，今後の課題とした．

生成されたベース音源の評価は，フレームごとに正解音源との音高の差を求め，差が 50cent 以内のときに正解とみなしたときの，正解率を用いて行う．

3.3 実 験 結 果

実験結果を表 1 に示す．表において STFT，mel，chroma はそれぞれ通常のスペクトログラム，メルスペクトログラム，クロマグラムを用いた場合の結果を示す．

3.3.1 実験条件 1

平均正解率が最も高いモデルは chroma，最も低いモデルは mel だった．mel は 10 個中 6 個で 3 モデル間で正解率が最下位であった．EABC♯m_voicing は chroma の正解率が極めて低い．マイナーコードを 3 つ以上含むデータは，STFT の正解率が chroma よりも高い結果となった．

表 1　モデルごとの正解率

条件	テストデータ	STFT	mel	chroma
条件 1	A♯CDmEm_voicing	0.37	0.32	0.66
	EABC♯m_voicing	0.19	0.14	0.03
	CDEmAm_voicing	0.39	0.49	0.53
	GABmD_voicing	0.55	0.54	0.62
	GCDEm	0.58	0.56	0.62
	CDmEmDm	0.42	0.21	0.17
	DmEmAmEm	0.57	0.40	0.32
	EmAmFG	0.59	0.39	0.81
	AmFGC	0.70	0.54	0.79
	FAmGDm	0.58	0.35	0.63
	平均	0.49	0.39	0.52
条件 2	A♯CDmEm_voicing	0.29	0.26	0.58
	EABC♯m_voicing	0.37	0.08	0.00
	CDEmAm_voicing	0.17	0.30	0.51
	GABmD_voicing	0.28	0.49	0.67
	GCDEm	0.24	0.31	0.52
	CDmEmDm	0.15	0.11	0.23
	DmEmAmEm	0.23	0.17	0.22
	EmAmFG	0.35	0.26	0.68
	AmFGC	0.17	0.35	0.76
	FAmGDm	0.41	0.42	0.58
	平均	0.27	0.28	0.48
条件 3	CDEmAm_Audio	0.20	0.09	0.35

テストデータの名称はコード進行を表す．学習に用いたコード進行と同じだがヴォイシングを変更したものは「_voicing」を付与した．

3.3.2 実験条件 2

実験条件 1 と同様，平均的に最も正解率の高かったモデルは chroma だった．一方，最も正解率が低かったモデルは STFT だった．STFT が条件 1 に比べて 0.2 以上正解率が下がったのに対して chroma ではそれほど下がらなかったのは，クロマグラムを計算する際に各オクターブの振幅を足すために，ローパスフィルタによって生じる周波数帯域の変化に頑健だったからと考えられる．一方，EABC♯m_voicing の結果を見ると，chroma の正解率が条件 1 と同様にほぼ 0 なのに対し，STFT の正解率は条件 1 よりも高かった．

3.3.3 実験条件 3

いずれも条件 1・2 より正解率は低いが，最も正解率の高いモデルは chroma，最も正解率が低いモデルは mel だった．

4. お わ り に

本稿では，CNN を用いてギター音源からベース音源を生成する手法を提案した．どの条件でもクロマグラムを用いたモデルの精度が最も良かった．ただ，どのモデルも十分な正解率を実現したとは言えない．今後は，複数の特徴量を併用したり学習データを増やすなどして精度を高めていきたい．

謝辞　本研究は，科研費 22H03711，21H03572 の支援を受けた．

参 考 文 献

1) 福田翼，池宮由楽，糸山克寿，吉井和佳: ユーザの技術に合わせた自動編曲機能をもつピアノ演奏練習システム，情報処理学会第 77 回全国大会，pp.2-402–403，2015.
2) S. Onuma and M. Hamanaka: Piano Arrangement System Based on Composers' Arrangement Processes, ICMC, 2010.
3) 丸山剛志，三浦雅展，柳田益造 : 与えられたメロディーとコード進行に基づくギター用編曲システムの構築，第 3 回情報科学技術フォーラム，pp.399–400, 2004.