6V-07

モーフィングを用いたドラムループ素材の生成

川原瑞樹 † 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

音楽制作などにおいて, 多様な音色や効果音を創り出すことは非常に重要である. 特に, 数小節程度の音素材を組み合わせて音楽制作を行うループシーケンサでは, 搭載する音素材のバリエーションが, 制作可能な楽曲のバリエーションを決めることになる. しかしながら, 音素材に限りがあるため, 自分の求める音素材が常に見つかるとは限らない.

一方, 近年, 機械学習を用いた音楽生成のする研究が発展している[1),2)]. 特に, 2 つの音楽コンテンツの中間的なものを作り出そうとする「モーフィング」は, 多く研究がなされている[3),4)].

我々も以前, このモーフィングの考え方を用いて, ドラムループ素材を対象に, 2 つの音素材の中間的なものを生成する方法を検討した[5)]. この手法では, 変分オートエンコーダ（VAE）と畳み込みニューラルネットワーク（CNN）を組み合わせたモデルを用いている. 70 個程度の音素材を用いてこのモデルを実現したが, 聴取実験による主観評価をしていなかった. 本稿では, 学習する音素材を増やした上で, 聴取実験による主観評価を実施する.

2. 提案手法

本研究の提案手法は, 音源の特徴を表す低次元の潜在空間上でモーフィングを実現するために, CNN および VAE に基づくモデル（以下, CNN-VAE と呼ぶ）を使用する.

学習時では, まず, 音源をフーリエ変換によってスペクトログラムに変換する. その後, 畳み込み層を使用し, 音源を低次元の潜在空間内にマッピングする. そして, 逆畳み込み層を用いて元の音源のスペクトログラムに復元する. この復元後のスペクトログラムが元のスペクトログラムに近づくように, CNN-VAE モデルが学習される.

生成時では, 潜在空間にマッピングされた落とし込まれた音源のうち 2 つを選び, 潜在空間上でその 2 つの内分点を任意の内分比で取る. そのようにして得られた潜在空間上の点からスペクトログラムを生成し, 音響信号に変換する. 内分比を与えることで, 異なる音源の中間的なモーフィング結果が得られる.

2.1 使用データ

本研究では, 音楽ループ素材を Wave 形式で多数収録されたデータセットとして,『Sound PooL vol. 2』☆を使用した. このデータセットから, ジャンルが「Techno & Trans」で楽器パートが「Drums」のものを 224 個抽出して CNN-VAE モデルの学習に使用した.

Generation of drum loops using morphing *by Mizuki Kawahara, Tomoo Kouzai and Tetsuro Kitahara (Nihon University)*

☆ https://www.ah-soft.com/soundpool/

2.2 スペクトログラムの生成

はじめに, 入力音響信号を短時間フーリエ変換（STFT）によって周波数成分に分解する. このとき, ハミング窓を用いて窓幅を 2048, ホップサイズを窓関数サイズの 1/4 とした. サンプリング周波数を 22050Hz, 音響信号の長さを約 3.43 秒（BPM 140 で 8 拍分）であることを前提としているため, 得られるスペクトログラムは 1025 行 148 列の行列となる.

2.3 CNN-VAE モデルの構築

本システムは, 音源のモーフィングを実現するために, CNN と VAE を組み合わせた CNN-VAE モデルを構築する. まずスペクトログラムを VAE のエンコーダ部分である畳み込み層で圧縮し, 潜在空間にマッピングする. 入力データの形状が 1025 × 148（周波数軸が 1025 要素, 時間軸が 148 要素）である. 3 つの畳み込み層によって 1 × 1 に圧縮する. その後, VAE のデコーダ部分である逆畳み込み層を用いて, スペクトログラムを復元する. この過程で活性化関数として ReLU 関数を使用し, バッチサイズを 64, エポック数を 3000 として学習を行う. 損失関数には二乗平均誤差を用いた.

2.4 モーフィングの実現

本システムは, 学習済みの音源から 2 つ（s_i, s_j とする）を選び, それらの潜在空間上の座標を $\boldsymbol{z}_i$, $\boldsymbol{z}j$ とする. その後, $\alpha : 1-\alpha$ に内分する点 $\boldsymbol{z}\mathrm{new} = (1-\alpha)\boldsymbol{z}_i + \alpha\boldsymbol{z}_j$ を求めることで, s_i と s_j の中間的なスペクトログラムを生成する.

最終的には, 逆フーリエ変換および位相復元を行うことで, 生成されたスペクトログラムを音響信号に変換する. これにより, 入力した α の値に応じて, 異なる音源のモーフィング結果が返ってくる（ただし, 後述の実験では常に 0.5）.

3. 実験

3.1 実験方法

次の 2 つの手順で実験を行う. なお, どちらの実験もクラウドソーシングサービスを使って Web 上で行う. 実験では, 提案手法によって生成された音素材の品質と有用性を評価する. 以下に実験項目とその詳細を示す.

評価 1：ランダムなペアでの一対比較

提案手法によって生成された音源（以下, 生成音源）と Sound PooL 内の音源（以下, 市販音源）を対で提示し, 曲作りに有用と思うのはどちらか, 機械学習モデルによって生成されたと思うのはどちらかを質問する. 生成音源と市販音源が同程度に選ばれたら, 生成音源は市販音源と互角のクオリティがあると解釈できる.

使用データ

2. の手法で生成した音源 100 個（元音源はランダムに選ぶ）と Sound PooL に収録されている音源 100 個（ランダムに選ぶ）を使用する.

実験参加者

Web 上のクラウドソーシングサービスを用いて集められた 201 名．参加者募集にあたって，年齢，性別，音楽的能力・経験は問わなかった．

実 施 手 順

実験参加者が指定された Web サイトにアクセスすると，生成音源と市販音源が 1 つずつランダムに並ぶ．参加者はこれらを聴き，「曲作りに使えそうなものはどちらか」「機械学習モデルによって生成されたと思われるのはどちらか」に答える．それぞれに対して理由も入力する．入力が終わって送信ボタンを押すと，次のペアに切り替わるようになっており，これを 10 回繰り返す．

なお，100 個のペアを 10 個のグループに分けられ，どのグループを聴くかはランダムに割り当てられる．

評価 2：元音源との類似度評価

生成音源に対して，モーフィングの元となる 2 つの音源のどちらに似ているかを問う実験である．回答が半数ずつに割れれば，生成音源はどちらにも似ていないことになり，中間的であると評価できる．

使用データ

生成音源は評価 1 と同じ 100 個の音源である．それに加え，それらを作る際に用いた元音源（述べ 200 個）を用いる．

実験参加者

評価 1 と同じ方法で集められた 145 名である．評価 1 とは別で集めたが一部重複している可能性はある．

実 験 手 順

実験用サイトにアクセスすると，一番上に生成音源（「音源 X」と呼ぶ）の再生ボタンが表示される．その下に 2 つの音源（「音源 A」「音源 B」と呼ぶ）の再生ボタンが表示される．実験参加者は，音源 X，音源 A，音源 B を聴き，X が A と B のどちらに似ているかを答える．選択肢は「A に近い」「どちらかといえば A に近い」「どちらかといえば B に近い」「B に近い」の 4 つである．この選択肢の 1 つを必ず選ぶこととする．理由も回答する．これを 10 回繰り返すと終了である．

3.2 結　　果

3.2.1 評価 1：ランダムなペアでの一対比較

「曲作りに使えそうなものはどちらか」「機械学習モデルによって生成されたと思われるのはどちらか」の質問に対して生成音源を答えた参加者の割合をペアごとに求め，その平均と標準偏差を求めた．

結果を表 1 に示す．2 つの質問のどちらも平均値が 0.5 に近いことから，提案手法によって生成された音素材を，既存の市販音源と同程度の品質があることが示唆される．ただし，標準偏差が両方の質問において比較的大きいことから，個人の好みや音楽に対する経験が評価に影響を与えている可能性がある．

3.3 評価 2：元音源との類似度評価

回答の選択肢である「A に近い」「どちらかといえば A に近い」「どちらかといえば B に近い」「B に近い」を便宜上感覚尺度とみなして，それぞれ整数 4，3，2，1 を付与した．

表 1　評価 1 の結果 (生成音源を選んだ割合)

	平均	標準偏差
曲作りに有用か	0.479	0.353
どちらが生成音源か	0.439	0.347

その後，音源ごとに参加者による回答の平均を求めた．そして，その値が 1 以上 2 未満，2 以上 3 未満，3 以上 4 未満である割合を求めた．2 以上 3 未満の割合が大きければ，参加者による判断が割れたことを示し，A と B のどちらにも似ていない（中間的である）と解釈できる．

結果を表 2 に示す．100 個の生成音源のうち 48%の結果が「2 以上 3 未満」であったことから，半数のものは，A と B のどちらにも似すぎずに中間的なものが生成されたと言える．ただし，本実験の結果は，A と B のどちらにも似ていないことを示すのみで，A と B の特徴をともに引き継いでいることを示しているわけではない．潜在空間内に A と B の中間に別の音素材 C が存在し，それに似たものが生成されている可能性も考えられる．今後，そのような状況の発生の有無も検証していく必要がある．

表 2　評価 2 の結果

	割合
1 以上 2 未満	28%
2 以上 3 未満	48%
3 以上 4 未満	24%

4. お わ り に

本研究では, 音楽制作におけるドラムループ素材の生成を目的に VAE と CNN を組み合わせたモーフィング手法を提案した．主観評価の結果，市販の音素材に対して遜色のない品質であること，モーフィングの元となる 2 つの音源のどちらかに酷似するケースは少ないことが示唆された．今後は, より多様なジャンルや楽器, スタイルに対応できるよう, モデルの改良や新しい学習手法を検討する必要がある．

謝辞　本研究は科研費 JP22H03711, JP21H03572 の支援を受けた．

参 考 文 献

1) Gaëtan Hadjeres, François Pachet, and Frank Nielsen: DeepBach: a Steerable Model for Bach Chorales Generation, *Proceedings of the 34th International Conference on Machine Learning* (ICML 2017), pp. 1362—1371 (2017).
2) Li-Chia Yang, Szu-Yu Chou, and Yi-Hsuan Yang: MidiNet: A Convolutional Generative Adversarial Network for Symbolic-Domain Music Generation, *Proceedings of International Society for Music Information Retrieval Conference* (ISMIR 2017), pp. 324–311 (2017).
3) 村田叡, 坂東宜昭, 糸山克寿, 吉井和佳, Variational Auto Encoder を用いたメロディとコードのモーフィング, 情報処理学会第 80 回全国大会 (2018).
4) 小林瑞季, 浜中雅俊, 新しい GTTM メロディモーフィング手法の提示：既存手法とマリンバ作品を経て, 情報処理学会研究報告, Vol. 2019-MUS-123 No. 40 (2019).
5) 川原瑞樹, 奥田真輝, 香西智雄, 北原鉄朗, CNN-VAE を用いたドラム音源のモーフィングの試み, 情報処理学会研究報告（2023）.