Restricted Boltzmann MachineとConditional RBMを用いたモデルによる音楽生成

A Music generation by Associative Memorization Model of The Music Features using Restricted Boltzmann Machine and Conditional RBM

○丹羽孔明 (北海道科学大) 成瀬継太郎 (会津大)
大江亮介 (北海道科学大) 木下正博 (北海道科学大)
三田村保 (北海道科学大) 川上敬 (北海道科学大)

Tadaaki NIWA, Hokkaido University of Science, r13301@hus.ac.jp
Keitaro NARUSE, University of Aizu
Ryousuke OOE, Hokkaido University of Science
Masahiro KINOSHITA, Hokkaido University of Science
Tamotsu MITAMURA, Hokkaido University of Science
Takashi KAWAKAMI, Hokkaido University of Science

In this paper, we develop an associative memorization architecture of the musical features from time sequential data of the music audio signals. This associative memorization architecture is constructed by using of Restricted Boltzmann Machine (RBM) and Conditional RBM (CRBM). Challenging purpose of our research is development of the new composition system that automatically creates a new music based on some existing music. In this composition system, characteristics or common musical features of existing music are learned and extracted, then a new different musical melody having such musical features is created. This paper illustrates the important former part of the whole system.

Key Words: Learning, Artificial Neural Network, Deep learning, Automatic composition system

1 はじめに

コンピュータサイエンスの方法論を用いた自動作曲システムの実現はいまだ魅力的な研究分野である。特に人間の感性や感動を反映した自動作曲はエキサイティングな研究テーマである。これらのシステムの多くでは対話型遺伝的アルゴリズムを用いて、人間が感じる音と言葉に対する印象から音楽が生成される。しかしながら、対話型遺伝的アルゴリズムによるアプローチは各解の評価と適応度の定量化を設計者やユーザに強制する。そこで我々は、音楽理論の専門知識を持たないアマチュアであっても使用することのできる完全自動作曲システムの実現を試みる。

本稿では、新しい楽曲の生成を行うために音楽音響信号の時系列データから得られる音楽的特徴の連想記憶アーキテクチャを開発する。この連想記憶アーキテクチャは Restricted Boltzmann Machine (RBM)[4] と Conditional RBM[5] によって構成される。本研究の挑戦的な目的は、いくつかの既存の楽曲に基づいて新しい楽曲の生成を行う知的な自動作曲システムの開発である。この作曲システムにおいては、既存の音楽に共通する音楽的な特徴や特性が学習、抽出され、その後、これらの音楽的特徴を持つ新しく様々な音楽のメロディーが生成される。このように、既存の音楽の音楽的特徴が連想記憶アーキテクチャによって学習される。本稿ではシステム全体の重要な前半部分を示している。

どのように人間の作曲家は楽譜や音楽作品を作るのか？ 一般に言えば、楽曲は音楽の制作手順において音楽的特徴の反復的な解析処理と再合成処理によって生成される。自動的に音楽を生成する知的なシステムを構築するためには、音楽生成アルゴリズムおよび音楽評価アルゴリズムを提供することが必要である。我々はまず、知的なシステムにおいて評価の対象となる特定の音楽を生成する楽曲生成アルゴリズムに焦点を当てる。

関連研究として音楽の自動作曲における代表的なシステムを紹介

介する。D.Cope らは音楽制作における人間の知的なプロセスを代替するシステムという位置付けで、楽曲を自動的に作曲するシステムを開発した。Experimental in Music Intelligence (EMI)[2] と呼ばれているこのシステムは与えられた楽曲をモチーフ抽出処理とルール解析処理によって、楽曲の断片であるモチーフとその楽曲を構成する規則に分解し、蓄積する。モチーフとは楽譜中のおおよそ 2 小節の音符のパターンであり、楽曲において意味を持つ最小のパターンとされている。この解析処理の後、EMI は蓄積したルールと乱数を用いてモチーフを再結合することによって作曲家の音楽様式を含む新しい楽曲を生成する。このような音楽の生成プロセスは、人工ニューラルネットワーク (Artificial Neural Network: ANN) を用いたアーキテクチャによるモデルでシミュレートすることが可能である。

ANN のアーキテクチャを用いたモデルにおいて、データの記述形式はその学習性能と機能の形成に大きな影響を与える。このため、まず重要な問題となるのが音楽データの記述方法である。多くの関連研究では楽譜データや MIDI データ、あるいはそれらに基づく独自の記号形式など、音楽データの記号による表現方法を採用している。しかし、我々は人間の作曲家が単純な記号として楽曲を扱うことは無く、その記号に付随する音色、音の強さ、音高、音長、その他の聴覚感覚に由来する様々な情報を総合的に処理していると信じている。そのため、我々は生の音楽の音響信号を音楽データの記述に採用し、これをシステムへ入力として与え、またシステムからの出力として得る。このように音楽データの記述形式に音響信号を用いた音楽生成モデルはあまり研究されていない。

構築した連想記憶モデルは音楽音響信号の音楽的な特徴を記憶し、その音楽の連続したデータを再生する。この音楽の連想記憶のアーキテクチャは RBM と CRBM に基づいて構築される。構築したシステムについて、音楽の記憶と生成の実験を行い、音楽の音響信号データを記憶、生成した実験結果に基づいて、システムのパフォーマンスや有効性を検証する。

2 Restricted Boltzmann Machine

2.1 Restricted Boltzmann Machine

RBM は無向グラフィカルモデルによるエネルギーに基づいた生成メカニズムである。図 1 中の左の図に示すように、可視変数の層 $\mathbf{v} = [v_1, v_2, \dots, v_V]$ は無向な接続を用いて隠れ変数の層 $\mathbf{h} = [h_1, h_2, \dots, h_H]$ に接続される。 V, H は $\mathbf{v}, \mathbf{h}$ それぞれのユニット数である。RBM の一般的なモデルにおいて、可視ユニットの状態は $[0, 1]$ のバイナリ値によって定義され、また隠れユニットの状態も同様にバイナリ値によって表現される。この場合において、RBM のモデルは次のエネルギーを持つ。

$$E(\mathbf{v}, \mathbf{h}) = - \sum_i a_i v_i - \sum_j b_j h_j - \sum_{ij} v_i h_j w_{ij} \quad (1)$$

ここで $\mathbf{W} = [w_{11}, w_{12}, \dots, w_{VH}]$ は可視層と隠れ層間の無向接続の重み行列であり、 $\mathbf{a} = [a_1, a_2, \dots, a_V]$ と $\mathbf{b} = [b_1, b_2, \dots, b_H]$ はそれらの層のバイアスである。可視ユニット $\mathbf{v}$ と隠れユニット $\mathbf{h}$ の同時確率は、このエネルギー関数によって次のように与えられている。

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})) \quad (2)$$

ここで Z は $Z = \sum_{\mathbf{v}, \mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))$ で与えられる規格化定数である。

RBM のモデルパラメータ $\mathbf{a}, \mathbf{b}$ と $\mathbf{W}$ は対数尤度学習のための Contrastive Divergence (CD_k)[3] 学習アルゴリズムによって推定される。可視ユニットの発火状態と隠れユニットの発火状態は次の条件付き確率の式により求められる。

$$p(h_j = 1 | \mathbf{v}) = \text{sigmoid}(b_j + \sum_i v_i w_{ij}) \quad (3)$$

$$p(v_i = 1 | \mathbf{h}) = \text{sigmoid}(a_i + \sum_j h_j w_{ij}) \quad (4)$$

ここで $\text{sigmoid}(x)$ は $\text{sigmoid}(x) = 1/(1 + \exp(-x))$ によって与えられる標準シグモイド関数である。 CD_k アルゴリズムにおけるパラメータの更新規則は次のように与えられる。

$$\Delta W = \epsilon(\langle v_i, h_j \rangle_{data} - \langle v_i, h_j \rangle_{model}) \quad (5)$$

$$\Delta a = \epsilon(\langle v_i \rangle_{data} - \langle v_i \rangle_{model}) \quad (6)$$

$$\Delta b = \epsilon(\langle h_j \rangle_{data} - \langle h_j \rangle_{model}) \quad (7)$$

ここで $\langle \rangle_{data}$ は学習データ、 $\langle \rangle_{model}$ は式 (3) と (4) を用いたギブスサンプリングによって再構成された学習データを表す。

本稿においては、音楽生成モデルの構築に Gaussian Bernoulli RBM (GBRBM)[1] を用いる。GBRBM は可視ユニットの状態を実数値で定義し、また隠れユニットの状態をバイナリ値で表現したモデルで、エネルギー関数と可視ユニットの発火状態は次のように与えられる。

$$E(\mathbf{v}, \mathbf{h}) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{ij} \frac{v_i}{\sigma_i} h_j w_{ij} \quad (8)$$

$$p(h_j | \mathbf{v}) = N(h_j; b_j + \sum_i v_i w_{ij}, \sigma^2) \quad (9)$$

ここで $N(a; b; \sigma^2)$ はガウス関数であり、また σ は標準偏差である。

2.2 Conditional RBM

CRBM はエネルギーに基づいたグラフィカルモデルによる時系列な状態の生成モデルであり、RBM において時系列データを処理可能な拡張の一つである。図 1 の右の図は CRBM の各層の接続を表す。可視変数の層 $\mathbf{v}$ は無向な接続によって隠れ変数の層 $\mathbf{h}$ へ接続される。図中の層 $\mathbf{u} = [u_{t-1}, u_{t-2}, \dots, u_{t-N}]$ はある時刻 t における過去の可視ユニットの状態を表す履歴ベクトルであり、 N はその履歴数である。この履歴の層は有向な接続によって可視

Fig.1 Restricted Boltzmann Machine (RBM) and Conditional RBM (CRBM). left image illustrates RBM model, and right shows CRBM model.

Table 1 The number of units for each layer in model

1st layer	2nd layer	past stats	3rd layer
16,000	200	1,400	200

ユニットと隠れユニットの層に接続される。CRBM のモデルは次式で与えられるエネルギーを持つ。

$$E(\mathbf{v}, \mathbf{h} | \mathbf{u}) = \frac{1}{2} \sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_{ij} w_{ij} v_{i,t} h_{j,t} \sum_j \hat{b}_{j,t} h_{j,t} \quad (10)$$

ここで $\mathbf{W}$ は可視ユニットと隠れユニットの間の無向な接続の重み行列であり、 t は現在の時刻である。 $\hat{\mathbf{a}} = [\hat{a}_1, \hat{a}_2, \dots, \hat{a}_i]$ と $\hat{\mathbf{b}} = [\hat{b}_1, \hat{b}_2, \dots, \hat{b}_i]$ はそれらの層のバイアスであり、履歴ベクトル $\mathbf{u}$ を用いて次のように求められる。

$$\hat{a}_{i,t} = a_i + \sum_k A_{ki} u_k \quad (11)$$

$$\hat{b}_{j,t} = b_j + \sum_k B_{kj} u_k \quad (12)$$

ここで $\mathbf{A}$ は履歴ベクトル $\mathbf{u}$ から可視ユニット $\mathbf{v}$ への有向接続の重み行列であり、 $\mathbf{B}$ は履歴ベクトル $\mathbf{u}$ から隠れユニット $\mathbf{h}$ への有向接続の重み行列を表す。

CRBM のモデルパラメータは RBM と同様に CD_k アルゴリズムによって推定される。各層のユニットの発火状態は式 (3) と (4) を用いて求められる。ここで式 (3) 中の変数 a_i は $\hat{a}_{i,t}$ に、また式 (4) 中の変数 b_j は $\hat{b}_{j,t}$ にそれぞれ変更される。各パラメータの更新規則は式 (5), (6), (7) および次の式が用いられる。

$$\Delta A = \epsilon(\langle v_i, u_k \rangle_{data} - \langle v_i, u_k \rangle_{model}) \quad (13)$$

$$\Delta B = \epsilon(\langle h_j, u_k \rangle_{data} - \langle h_j, u_k \rangle_{model}) \quad (14)$$

3 音楽的特徴の連想記憶モデル

3.1 連想記憶モデルの構成

図 2 に示すように GBRBM と CRBM を積層し、音楽音響信号の時系列データを音楽情報の記述形式として採用した音楽的特徴の連想記憶モデルを構築した。連想記憶モデルの第一ブロックにおいて音響信号は特徴ベクトルにエンコードされ、また特徴ベクトルは音響信号にデコードされる。第二ブロックにおいては特徴ベクトルの時系列データが学習、記憶され、また新しい特徴ベクトルの時系列データが生成される。このモデルの各層のサイズは表 1 に示すユニット数のパラメータを用いて設定した。

3.2 連想記憶モデルの学習

CD_k 学習アルゴリズムを用いて、連想記憶モデル内の各ブロックに対するモデルパラメータを推定した。表 2 は、このモデルの訓練のための学習パラメータの設定である。訓練データには Vivaldi 作曲の「和声と創意への試み」より「春」「夏」「秋」「冬」各曲の第 1～第 2 楽章から冒頭 15 秒間を切り出し、これを繋げたものを用いた。この訓練用の音楽音響信号について、表 3 の設定を用いてリサンプリングとフレームへの分割を行った。フーリエ変換や主成分分析を用いたホワイトニングなど、信号処理を行う上で多く用いられる前処理は適用せず、本モデルは生の音声信号波形のデータを学習する。

Fig.2 Model for association of the music audio signals.

Table 2 Learning parameters for each blocks in model

rearning rate	momentum	weight decay	learning epochs
0.0001	0.9	0.001	6000

Table 3 Parameters for resampling and frame splitting

sampling rate	quantization bit rate	channels
16kHz	16bit	256kbps
window size	window function	overlap
16,00	rectangular window	0

4 実験結果

まず、訓練用データより冒頭 30 秒間を学習の完了した連想記憶モデルに与えて 180 秒間の音響信号を生成した。図 3 は訓練データとモデルより得られた音響信号のパワースペクトルである。生成された信号において開始から約 7 秒程度はノイズであったが、その後訓練データと同様の音を聞き取ることで音響信号が得られている。

次に、未学習の音響信号データを与えて 360 秒間の音響信号を生成した。このとき与えた信号データは、Vivaldi 作曲の「和声と創意への試み」より協奏曲第 9 番の第 1 楽章から冒頭の 20 秒間である。図 4 は訓練データとモデルより得られた音響信号の 25~205 秒間のパワースペクトルである。モデルより生成された信号は開始から約 90 秒までは強いノイズのみで、その後徐々に訓練データと同様の音が聞こえ始めた。およそ 145 秒から 205 秒までの間はノイズが減少し、訓練データと同じ楽曲がはっきりと聞き取れるようになり、205 秒を過ぎると再度、強いノイズのみとなった。

未学習のデータを与えた場合については学習した音響信号を生成するケースの他に、強いノイズの中に時折、学習した楽曲の断片が含まれる信号を生成するケースが確認され、音響信号の生成時に与える信号データの初期値に強く依存していることが考えられる。

これらの再構成された信号にはノイズが含まれている。しかしながら、再生された音響信号はノイズを含むにもかかわらず、人間のテスターにとって元の楽曲として認識することができる。つまり、この連想記憶アーキテクチャは、RBM と CRBM を用いて音響信号から音楽的特徴を抽出、記憶し、また記憶した信号を想起することが可能である。

4.1 おわりに

本稿では、新しい楽曲の生成を行うために、BM と CRBM のアーキテクチャを用いて音楽音響信号の時系列データを生成する連想記憶モデルを構築する。この連想記憶モデルについて、音楽音響信号を記憶することを確認するが、学習した楽曲の特徴を反映した新しい楽曲の生成は確認されていない。この原因として、連想記憶モデルに与える訓練データが楽曲の音響信号そのものという抽象化されていないデータのため、入力として与える楽曲データもしくは RBM による特徴ベクトル系列への変換後のデータにおいて、楽曲のデータを表現する際の抽象化の度合いによる影響が懸念され、またモデルの記憶する楽曲の断片において、そのパターンの多

Fig.3 Power spectrogram of the learning audio signals and generated audio signals. upper image is spectrogram of the audio signals for learning, and lower image is spectrogram of the audio signals by model generation.

Fig.4 Power spectrogram of the generated audio signals using unlearned data. upper image is spectrogram of the audio signals for learning, and lower image is spectrogram of the audio signals by model generation.

様性が十分ではないことが考えられる。

本研究の今後の課題については、上記の 2 点について RBM による Deep learning を用いて楽曲のより抽象的な表現を獲得する手法を検討するとともに、楽曲の断片のパターンをより多く記憶可能なネットワークパラメータを推定する。これらに加えて、連想モデルより生成された音楽音響信号のノイズを除去する方法についての検討も重要な課題の一つである。

参考文献

- [1] Cho, K. H., Ilin, A., Raiko, T., "Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines," Artificial Neural Networks and Machine Learning-ICANN 2011, pp.10-17, Springer, Berlin Heidelberg, 2011.
- [2] Cope, D., "Computers and Music Style," Computer Music and Digital Audio Series, 1991.
- [3] Hinton, G. E., "Training products of experts by minimizing contrastive divergence," Neural computation, vol.14-8, pp.1771-1800, 2002.
- [4] Hinton, G., "A Practical Guide to Training Restricted Boltzmann Machines," Tech. Rep. Department of Computer Science, University of Toronto, 2010.
- [5] Taylor, G. W., and Hinton, G. E., "Factored conditional restricted Boltzmann machines for modeling motion style," Proceedings of the 26th annual international conference on machine learning, pp.1025-1032, 2009.