The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

segmented VRAE と遺伝的プログラミングに基づく音楽の創作

Music Composition based on the Genetic Programming with segmented VRAE

山本 周典 *1　　森 直樹 *1
Hironori Yamamoto　　Naoki Mori

*1大阪府立大学工学研究科
Graduate School of Engineering, Osaka Prefecture University

Automatic music composition is one of the most difficult and attractive challenges in the artificial intelligence (AI) field. In order to tackle this challenge, an approach using interactive evolutionary computation (IEC) is drawing attention because IEC takes human emotions into consideration. We have proposed an automatic music composition system based on IEC with a surrogate model called an evaluation model. In the previous study, the model is constructed with a Variational Recurrent Auto-Encoder (VRAE) to achieve quantitative evaluations. However, it is not easy for a simple VRAE to map tunes' features into a meaningful latent space regardless of their lengths.

This paper focuses on the way to map tunes with different length into a good latent space and the application for IEC. The evaluation model employs a hierarchical VRAE called segmented VRAE. The experiments are carried out to show the effectiveness of the proposed method.

1. はじめに

近年，計算機による楽曲の自動生成に関する研究が積極的になされており，新しい文化や産業の創造に繋がるとして多方面から強く期待されている．しかしながら，作品の評価は個人の嗜好や感性に大きく依存するため，これを定量的に評価することは非常に困難とされている．このような問題に対して，人間の評価系そのものを評価関数として最適化システムに導入した対話型進化型計算 (Interactive Evolutionary Computation: IEC)[Takagi 98] が注目をされている．

そこで我々はこれまで，適応度景観を学習し近似的に評価関数を推定する surrogate model[Jin 11] を応用した IEC による対話型楽曲自動生成システムを提案してきた．[山本 18] この従来システムでは，楽曲を表現するために音高および音価の概念を反映した木構造を用いた．また，Genetic Programming (GP) の拡張手法である Genetic Programming with Multi-Layered Population Structure (MLPS-GP)[Hasegawa 17] を楽曲の探索アルゴリズムとして用いた．surrogate model としては評価モデルと呼ばれるモデルを作成し，ユーザに代わって評価モデルに近似的に楽曲評価をさせることで IEC におけるユーザ負荷の問題を軽減した．この評価モデルにおける評価の過程では生成モデルの一種である Variational Recurrent Auto-Encoder (VRAE)[Fabius 14] を用いた．VRAE は Variational Auto-Encoder (VAE)[Kingma 13] の拡張手法で，VAE と同様に入力データが潜在的に内包している意味を反映しつつ写像した潜在空間を構築することが可能である．従来システムの楽曲評価では，潜在空間内における楽曲の潜在変数を評価の対象とすることで定量的な楽曲評価を実現した．しかしながら，単純な構造を持つ VRAE の場合，楽曲のような可変長データをその長さに関係なく適切な潜在変数に写像することは難しい．

本研究では以上の点を背景として，segmented VRAE と称す階層構造をもった VRAE に基づく対話型楽曲自動生成システムを提案する．segmented VRAE では楽曲を一定幅ごとに潜在変数に写像し，楽曲を複数の潜在変数によって扱う．この結果，楽曲の全体の長さに左右されることなく潜在変数を獲得することができる．得られた潜在変数をもとに適応度関数を作成することで，より入力楽曲の要素を適切に反映した楽曲生成を目指す．

連絡先: 山本 周典，大阪府立大学 工学研究科，大阪府堺市中区学園町 1-1，072-254-9273，h.yamamoto@ss.cs.osakafu-u.ac.jp

2. 提案モデル

本章では VRAE の概要および本システムで用いた segmented VRAE について説明する．

2.1 Variational Recurrent Auto-Encoder

VRAE[Fabius 14] は VAE[Kingma 13] と呼ばれる生成モデルの拡張手法で，エンコード部およびデコード部で Recurrent Neural Network (RNN) を使用している．このため楽曲のような時系列データに適している．VAE および VRAE の目的関数は，潜在変数の確率分布の形状に関わる項（latent loss）および潜在変数を事前分布として元データの復元に関わる項（reconstruction loss）で構成される．

2.2 segmented VRAE

次に，segmented VRAE について説明する．segmented VRAE は階層的な構造を持ち，入力データを幾つかの部分列に分けて扱う．

図 1 に segmented VRAE の構造を示す．segmented VRAE ではエンコード部において入力 x を n_x 個の部分データに分解し，各部分データに対応する潜在変数の集合 $\boldsymbol{Z} = \{\boldsymbol{z}_x^{(i)}\}_{i=1}^{n_x}$ を得る．また，デコード部では得られた潜在変数を上層の RNN に入力し，得られた出力を下層の RNN の初期状態および入力に結合する．各部分データの入力の初期値には “EOS” タグを入力する．本研究ではエンコード部では BiDirectional RNN を使用し，デコード部では Long Short-Term Memory (LSTM) [Schmidhuber 97] を使用した．

3. 楽曲の自動生成システムの概要

本章では進化型計算を用いた楽曲自動生成システムについて示す．

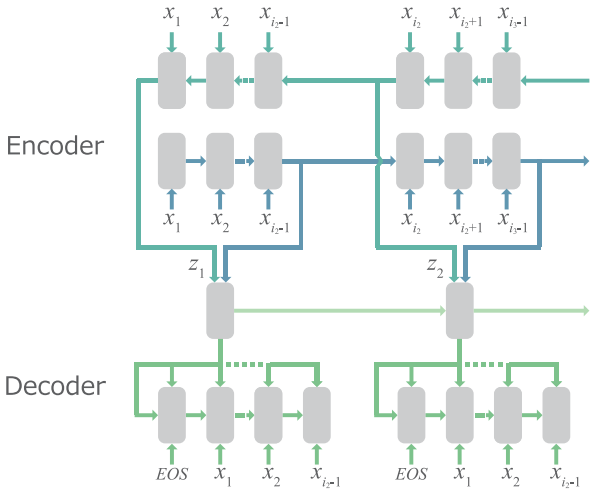

図 1: segmented VRAE の構造

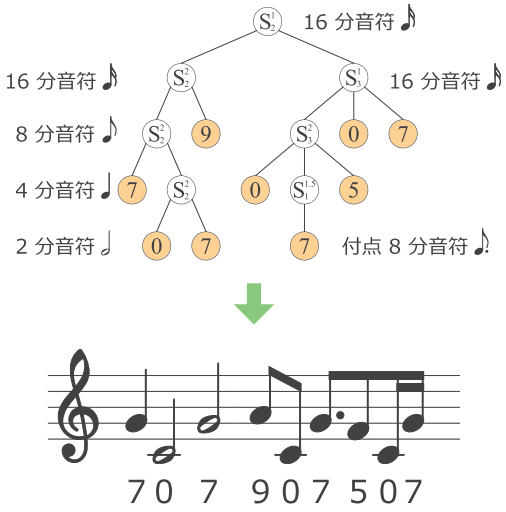

図 2: 木構造と楽譜の対応関係の例

3.1 提案システムの流れ

本システムはユーザの嗜好に沿った曲をユーザの評価に基づいて進化的に獲得することを目的としたシステムである．以下，評価モデルが楽曲の特徴量を学習するのに使用した曲を入力曲，進化型計算による探索中の個体を探索曲，探索終了後のエリート個体を進化曲と呼ぶ．

1. 入力曲に基づいて評価モデルが構築される．
2. 評価モデルによって探索曲を近似評価し，この評価に従い MLPS-GP が進化曲を生成する．
3. ユーザが進化曲を実評価する．
4. 評価モデルがユーザの実評価に基づいてユーザの嗜好を推定し，さらに IEC により評価モデルを更新し，初期個体群に進化曲を反映する．
5. ユーザが満足する曲が生成できた場合は終了する．そうでなければ 2. に戻る．

ただし，本研究では評価モデルによる近似評価に対して主眼を置くため，以下 2. までを対象として述べる．

3.2 入力曲の定義

本研究では，入力曲を一般入力曲および選好入力曲の 2 種類に大別し，評価モデルに学習させる．一般入力曲とは人が普段から耳にし聴き心地の良いと感じるような普遍的な曲のことで，これらを大量に学習することによって汎用的な楽譜情報の獲得が期待される．一方，選好入力曲とは，進化曲の特徴付けのために一般入力曲の中から抽出された曲のことで，ユーザの嗜好を獲得することが期待される．本システムでは，一般入力曲の学習によって獲得した普遍的な楽曲の特徴およびユーザに選択させた選好入力曲の特徴を組み合わせることで，ユーザの嗜好に合った進化曲の獲得を目的としている．

3.3 個体表現

本研究では GP を適用するために曲を木構造によって表現する．この木構造によって曲の音高および音価が表現可能である．図 2 に，GP における曲の個体表現と楽譜の対応関係を示す．終端ノードを音高とし，非終端ノードを子ノードへの音高の倍率および子ノードへの分岐数とした．l を子ノードの音価への倍率，m を子ノードへの分岐数として，非終端ノードを S_m^l で表す．l，m に関しては，$l \in \{1.0, 1.5, 2.0\}$，$m \in \{1, 2, 3, 4\}$ と定めた．また本研究では，根ノードの音価を 16 分音符で固定し，終端ノードは 3 オクターブ分に対応する 36 種類を用いた．なお，$\mathrm{S}_m^{1.5}$ に関しては付点音価を表現することを目的としているため，ある終端ノードの音価を計算するに際して，その親ノードが $\mathrm{S}_m^{1.5}$ の場合においてのみ音価を 1.5 倍とし，それ以外の $\mathrm{S}_m^{1.5}$ に関しては音価を 1 倍とするように設定した．木構造表現から楽譜に変換する際は，終端ノードに関して深さ優先探索の先行順で辿ることで変換がなされる．

3.4 適応度

ここでは MLPS-GP で最適化する際に用いる適応度について示す．適応度を計算するにあたって，事前に VRAE に対して一般入力曲を学習させる．次に，選好入力曲 $m_t (t \in \mathcal{N})$ をユーザが設定し，これらを VRAE のエンコード部によって潜在空間に写像することで，潜在変数の集合 $\boldsymbol{Z}_{m_t} = \{\boldsymbol{z}_{m_t}^{(i)}\}_{i=1}^{n_{m_t}}$ を得る．また，探索曲 x についても同様に潜在変数の集合 $\boldsymbol{Z}_x = \{\boldsymbol{z}_x^{(i)}\}_{i=1}^{n_x}$ を算出し，これらをデコードすることで $\hat{x}$ を得る．

以上の変数を用いて，探索曲 x に対する直接的な評価指標である個体評価値 $f(x)$ を定義していく．個体評価値 $f(x)$ は以下の 2 項目によって定義される．

- ユーザ設定の選好入力曲 m_t および探索曲 x の類似度 $f_\mathrm{d}(m_t, x)$
- 探索曲 x に対する外れ値検知項 $f_\mathrm{o}(x, \hat{x})$

3.4.1 楽曲間距離 $f_\mathrm{d}(m_t, x)$

ユーザ設定の選好入力曲 m_t および探索曲 x の類似度は潜在変数間の Kullback-Leibler 情報量（KL 情報量）をもとに算出する．選好入力曲の潜在変数 $\boldsymbol{Z}_{m_t} = \{\boldsymbol{z}_{m_t}^{(i)}\}_{i=1}^{n_{m_t}}$ および探索曲の潜在変数 $\boldsymbol{Z}_x = \{\boldsymbol{z}_x^{(i)}\}_{i=1}^{n_x}$ に基づき，楽曲間距離 $f_\mathrm{d}(m_t, x)$ を以下で定める．

$$f_\mathrm{d}(m_t, x) = \frac{1}{n_{m_t} n_x} \sum_{v=1}^{n_{m_t}} \sum_{w=1}^{n_x} D_\mathrm{KL}(\boldsymbol{z}_{m_t}^{(v)}, \boldsymbol{z}_x^{(w)}) \qquad (1)$$

ただし，$D_{\mathrm{KL}}(\cdot,\cdot)$ は KL 情報量を表す．

3.4.2 外れ値検知項 $f_{\mathrm{o}}(x,\hat{x})$

外れ値検知は探索曲をデコードした際の reconstruction loss をもとに算出する．探索曲 x および VRAE のデコードによって得られた $\hat{x}$ を用いて，外れ値検知項 $f_{\mathrm{o}}(x,\hat{x})$ は以下で与えられる．

$$f_{\mathrm{o}}(x,\hat{x}) = |g(x,\hat{x}) - \beta| \tag{2}$$

ただし，g は VRAE 学習時に reconstruction loss を算出した関数とし，β は学習済み VRAE に対してテストデータを入力した際に得られる reconstruction loss とする．このようにテストデータの reconstruction loss に基づいて，教師無しで入力データと学習時のデータセットとの差異を測り，外れ値を検知する研究 [Erik 15, Bontemps 16] も報告されており，本研究でも有用であると考えられる．

3.4.3 個体評価値

（1）式および（2）式から，探索曲に対する個体評価値 $f(x)$ は以下で与えられる．

$$f(x) \quad = \quad \left(\frac{1}{T} \sum_{t}^{T} f_{\mathrm{d}}(m_t, x) + \alpha f_{\mathrm{o}}(x,\hat{x}) \right)^{-1} \tag{3}$$

ただし，α は可調整パラメータである．

（3）式では，選好入力曲および探索曲の楽曲間距離 $f_{\mathrm{d}}(m_t,x)$ を最小化することで探索曲を選好入力曲に近づける．また，外れ値検知項 $f_{\mathrm{o}}(x,\hat{x})$ を最小化することで学習に用いたデータに近い楽曲の生成を実現する．

3.5 適応度

ユーザの嗜好を反映する上で，生成する曲の長さは重要な要素である．ゆえに，本システムでは，ユーザが必要とする曲の曲長を t_{d}，進化曲の曲長を t_x とし，MLPS-GP における適応度 $F(x)$ は（3）式 を用いて，次式のように与えられる．

$$F(x) = \frac{t_{\mathrm{d}}}{t_{\mathrm{d}} + |t_{\mathrm{d}} - t_x|^2} f(x) \tag{4}$$

4. 実験

ここでは以下に示す 2 種類の実験をする．

- segmented VRAE の精度解析
- 探索過程における適応度の解析

実験で用いたデータは Essen フォークソングデータベースコレクション [Ess] から 8472 曲を引用した．各楽曲は 3 オクターブの範囲に規格化し，音高および音価の組合せに対して，楽譜と同様それぞれ固有のタグを割り振った．また，6434 曲を訓練データ，1609 曲をテストデータとして用いた．

4.1 segmented VRAE の精度解析

ここではネットワーク構造の違いによる影響を確認するために，従来 VRAE および segmented VRAE の精度を比較する．ただし本実験における従来 VRAE は，エンコード部を BiDirectional RNN，デコード部を LSTM から構成され，層状の構造を持たないものである．

表 1 に従来 VRAE および segmented VRAE 共通のパラメータを示す．segmented VRAE の場合，入力データを複数の部分データに分割する過程が必要となる．今回は予備実験により “5” タグずつ区切ることで部分データを作成した．

表 1: VRAE および segmented VRAE 共通のパラメータ

パラメータ名	値
batch size	512
embed units	256
hidden state	512
latent units	128
optimizer	Adam
alpha (Adam)	0.001
beta1 (Adam)	0.05
beta2 (Adam)	0.001
dropout	0.8
loss fuction	softmax cross entropy

表 2 に従来 VRAE および segmented VRAE のテストデータに対する精度を示す．ただし，epoch に関してはそれぞれテストデータに対する精度に基づき決定した．また，accuracy については高いほうが，それ以外の項目については低いほうが高い精度を示す．まず，全体の “loss” より segmented VRAE の方が従来 VRAE よりも高い精度を示している．また，“accuracy” および “reconstruction loss” の項目より，従来 VRAE に対して segmented VRAE の方が潜在変数からのデータ復元に関して高い精度を示すことがわかった．一方で “latent loss” の項目から，潜在変数に対する形状の最適化は従来 VRAE の方が高い精度を示すことがわかった．この理由として，VAE における loss は latent loss および reconstruction loss の和によって構成されており，reconstruction loss の方が全体の loss に対する影響が大きかったことが原因として考えられる．この結果，segmented VRAE は latent loss をあまり考慮せず reconstruction loss を下げる方向に最適化が進んだと考えられる．

4.2 適応度の解析

ここでは評価モデルによる探索曲の進化について解析する．選好入力曲を “SCHOENE AUGEN SCHOENE STRAHLEN”[Ess] の 1 曲に設定し，（4）式の適応度に基づき音楽を自動生成した．表 3 に本実験で用いた MLPS-GP のパラメータおよび評価モデルのパラメータを示す．

表 4 に進化曲の適応度および表 5 にランダム曲の適応度を示す．ただし，楽曲間距離 $f_{\mathrm{d}}(m_t,x)$ および外れ値検知項 $f_{\mathrm{o}}(x,\hat{x})$ は小さい値ほどよい結果を示しており，一方で適応度 $F(x)$ は大きい値ほど望ましい．表 4，5 よりすべての項目について進化曲がより良い結果を示した．また，これらの項目に対して進化曲およびランダム曲の間で Welch の t 検定にかけたところ，1% 水準で有意差が見られた．図 3 に評価回数に対する進化曲の適応度の推移を示す．これにより，評価回数を重ねるごとに適応度の上昇が確認できる．以上の結果は MLPS-GP による進化曲に対する最適化が設定した目的関数に基づき成功していることを示している．

5. まとめと今後の課題

本論文では segmented VRAE と称す階層構造をもった VRAE に基づく進化的楽曲自動生成システムを提案した．まず，segmented VRAE を導入し，これを活用した進化的楽曲生成システムを示した．実験では segmented VRAE の精度および生成過程における楽曲の最適化について解析した．今後の課題としては，深層学習に基づく手法の探索アルゴリズム，

表 2: 従来 VRAE および segmented VRAE の精度

種類	epoch	accuracy	latent loss	reconstruction loss	loss
従来 VRAE	100	0.2866	0.02724	2.542	2.569
segmented VRAE	300	0.4183	0.1664	2.081	2.248

表 3: MLPS-GP および評価モデルのパラメータ

パラメータ名	値
適応度評価回数 N_{e}	30000
初期化における減衰率 r_{dump}	0.8
初期の深さ制限 D_{init}	3
α（個体評価値）	1
β（外れ値検知項）	2.081
曲長 t_{d} [sec]	15

表 4: 進化曲の適応度

評価指標	楽曲間距離 $f_{\mathrm{d}}(m_t, x)$	外れ値検知項 $f_{\mathrm{o}}(x, \hat{x})$	適応度 $F(x) \times 10^{-2}$
平均	16.02	2.622	5.196
標準偏差	1.59	1.210	0.611

探索オペレータの構築およびユーザとの対話評価の導入があげられる．

謝辞

本研究は一部，日本学術振興会科学研究補助金基盤研究（C）（課題番号 26330282）および一般財団法人カワイサウンド技術・音楽振興財団の補助を得て行われたものである．

参考文献

[Bontemps 16] Bontemps, L., McDermott, J., Le-Khac, N.-A., et al.: Collective Anomaly Detection Based on Long Short-Term Memory Recurrent Neural Networks, in *International Conference on Future Data and Security Engineering*, pp. 141–152, Springer (2016)

[Erik 15] Marchi, E., Vesperini, F., Eyben, F., Squartini, S., Bj ö rn Schuller : A Novel Approach for Automatic Acoustic Novelty Detection Using a Denoising Autoencoder with Bidirectional LSTM Neural Networks, in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference On*, pp. 1996–2000, IEEE (2015)

[Ess] *The Essen Folksong Collection*http://www.esac-data.org/

[Fabius 14] Fabius, O. and van Amersfoort, J. R.: Variational Recurrent Auto-Encoders, *arXiv preprint arXiv:1412.6581* (2014)

[Hasegawa 17] Hasegawa, T., Mori, N., and Matsumoto, K.: Genetic Programming with Multi-Layered Population Structure, *Proceedings of the 2017 Annual Conference on Genetic and Evolutionary Computation* (2017)

[Jin 11] Jin, Y.: Surrogate-Assisted Evolutionary Computation: Recent Advances and Future Challenges, *Swarm and Evolutionary Computation*, Vol. 1, No. 2, pp. 61–70 (2011)

[Kingma 13] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes., *CoRR*, Vol. abs/1312.6114, (2013)

[Schmidhuber 97] Schmidhuber, J. and Hochreiter, S.: Long Short-Term Memory, *Neural computation*, Vol. 9, No. 8, pp. 1735–1780 (1997)

[Takagi 98] Takagi, H.: Interactive Evolutionary Computation: Cooperation of Computational Intelligence and Human Kansei, in *Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems*, pp. 41–50 (1998)

[山本 18] 山本 周典, 長谷川 拓, 森 直樹, 松本 啓之亮：深層学習によるユーザ評価モデルを導入した遺伝的プログラミングによる音楽自動生成手法の提案, 2018 年度人工知能学会全国大会 (2018)

表 5: ランダム曲の適応度

評価指標	楽曲間距離 $f_{\mathrm{d}}(m_t, x)$	外れ値検知項 $f_{\mathrm{o}}(x, \hat{x})$	適応度 $F(x) \times 10^{-2}$
平均	51.96	12.71	1.489
標準偏差	9.56	2.77	0.267

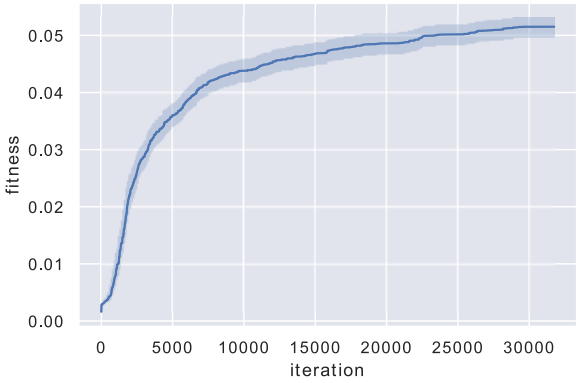

図 3: 評価回数に対する進化曲の適応度の推移．ただし，横軸は評価回数，縦軸は進化曲の適応度を示す．