Vol.2021-MUS-130 No.31
Vol.2021-EC-59 No.31
2021/3/17

音色変動の変分モデルによる楽音生成

増田 尚建[1,a)]　齋藤 大輔[1]

概要：複雑な音色を生成できる手法として，深層生成モデルによる楽音の合成が近年注目されている．本研究では楽音における音色の時間的変動を考慮した深層生成モデルを提案する．変分オートエンコーダ（VAE）によって楽音の各フレームにおける瞬時的な音色をモデル化し，回帰型ニューラルネットによりVAE の事前・事後分布を制御することで音色の変動をモデル化する．提案手法により，楽器の種類および感性的な質感について制御した楽音合成が可能となった．

1. はじめに

新たな楽音合成の手段として，ニューラルネットワークを用いた音の生成モデルが注目されている．デジタル・オーディオ・ワークステーションソフトウェアが中心となった現在の音楽制作環境では，高品質なサンプラー音源の普及によりリアルな楽器の音を用いることがすでに可能になっている．一方で，ニューラルネットワークによる音の生成モデルはユニークな音色を今までにない操作方法で合成することを可能にし，音楽制作に新たな刺激をもたらすことが期待される．実際に音楽制作で用いることを目的とした音の生成モデルは多様な音色の生成，および音色の直感的な制御を可能にすることが好ましい．

生成モデルの中でもオートエンコーダ（autoencoder, AE）および変分オートエンコーダ（variational AE, VAE）は楽音の再合成，および獲得された潜在変数による音色の操作などの音楽における応用が考えられてきた [1]，[2]．楽音の AE モデルは音色について滑らかな潜在空間を学習でき，従来のクロスシンセシス技術と同様の音色モーフィングなど，様々なアプリケーションに応用可能である．AE ベースの楽音生成モデルの多くは，再構成が比較的容易な短いフレーム単位（0.1 秒程度）を対象とする [1]，[3]，[4]，[5]．フレーム単位の AE モデルは音のフレームについての潜在空間を獲得することで，瞬時的な音色の制御を可能にする．

しかし，フレーム単位で再構成を行う AE モデルの欠点として，音のフレーム間の関係をモデル化しないため，正しい構造を持った新しい楽音の生成が困難であるという点が挙げられる．新しい音を生成するには，音のフレームの潜在空間において軌跡を描く必要があるが，どのような軌跡を描けば正しい構造を持った音になるかは明らかでない．潜在空間におけるランダムウォークによる楽音の生成 [4] も試みられたが，この手法は多様な音色を出力するモデルには適用できない．一方で，音色の時間変化についての知見を得るため，スペクトログラムフレームの可視化を多次元尺度構成法によって行った研究では，各楽器はフレーム単位の音色空間においてそれぞれ異なる軌跡を描くことが示された [6]．本研究では，このような音色の軌跡を考慮した VAE の潜在空間における音色の変動のモデルにより，正しい構造を持った新しい楽音の合成を行う．

また，回帰型ニューラルネット（recurrent neural networks, RNN）や WaveNet [7] などの自己回帰モデルを AE モデルのエンコーダおよびデコーダに用いることで，音の時間的構造をモデル化し，各楽音について一つの全域的な表現を学習することができる [8]．しかし，このような AE モデルでは瞬時的な音色とその時間変動が一つの潜在変数にまとめられてしまうため，潜在変数の解釈が容易でない．

楽音の音色を適切に扱うためには瞬時的な音色とその変動の両方をモデル化するべきである．例えば，音の各フレームを解析することで音の「明るさ」などの瞬時的な音色の特徴を解釈可能な形で説明できる．一方で，一つの音の全体的な音色の知覚はフレーム間の変動によってのみ説明できる [9]．そこで本研究では，音色の変動を考慮した楽音の生成モデルを提案する．VAE により楽音の瞬時的な音色をモデル化し，VAE の事前分布と事後分布を制御する RNN によって音色の変動をモデル化する．この音色変動モデルに音の特徴を条件として与えることにより，音の合成の制御を可能にする．本稿では提案手法により正しい構造をもった楽音の事前分布が学習されることを示し，潜

[1] 東京大学　大学院工学系研究科
Graduate School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan
[a)] n_masuda@gavo.t.u-tokyo.ac.jp

在空間における音色の変動を可視化する．

関連研究として，通常の VAE によって楽音のフレームをモデル化し，外部の RNN モデルにより潜在変数の系列を生成することで，音を生成するフレームワークが，グラニュラーシンセシスの発展型として提案されている [10]．これに対して，本研究では RNN により VAE の事前分布および事後分布を制御することで音色の変動のモデルを推論時にも用いるため，通常の VAE に比べて再構成性能の向上が期待される．

2. 変分オートエンコーダ

VAE では変分推論を用いて生成モデルを訓練する．データ $\mathbf{x} \in \mathbf{X}$ がある分布 $p(\mathbf{x})$ に従って，ある生成的要因（潜在変数）$\mathbf{z}$ を持つとする．ここで，下式を最大化したい．

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} \tag{1}$$

これを直接計算することはできないが，変分推論により近似的に求めることができる．近似的な分布 $q(\mathbf{z}|\mathbf{x})$ を考える．事後分布の近似 $q(\mathbf{z}|\mathbf{x})$ と本来の事後分布 $p(\mathbf{z}|\mathbf{x})$ との KL ダイバージェンスは式 (2) の通りである．

$$\begin{aligned} &D_{KL}[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}|\mathbf{x})] \\ &= E_{\mathbf{z}\sim q}[\log q(\mathbf{z}|\mathbf{x}) - \log p(\mathbf{z}|\mathbf{x})] \end{aligned} \tag{2}$$

式 (2) を変形すると以下のように (3) 式が求まる．

$$\begin{aligned} &\log p(\mathbf{x}) - D_{KL}[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}\mathbf{x})] \\ &= E_{\mathbf{z}\sim q}[\log p(\mathbf{x}|\mathbf{z})] - D_{KL}[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})] \end{aligned} \tag{3}$$

VAE では $q(\mathbf{z}|\mathbf{x})$ の近似にニューラルネットワークを用いる．このネットワークの表現力が十分に高い場合，$D_{KL}[q(\mathbf{z}|\mathbf{x})|p(\mathbf{z}|\mathbf{x})]$ はゼロに近づくと考えられるため，$\log p(\mathbf{x})$ の最大化は一般的に変分下限と呼ばれる式 (3) の右辺の最大化によって行える．よって，VAE の目的関数は式 (4) のようになる．

$$L_{\theta,\phi} = E_{q_\phi(\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}[q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})] \tag{4}$$

ここで，θ, ϕ は学習時に最適化されるデコーダ・エンコーダネットワークのパラメータである．目的関数の第 1 項を最大化することはオートエンコーダの再構成誤差を最小化することに等しい．第 2 項は事後分布を事前分布に近づける正則化の効果を持つ．事後分布 $q_\phi(\mathbf{z})$ および事前分布 $p_\theta(\mathbf{z})$ にどのような分布を用いるかは自由であるが，KL ダイバージェンスの計算のしやすさを考えて，事前分布に標準正規分布 $\mathcal{N}(0, I)$ を用い，事後分布に対角成分以外が 0 となる分散共分散行列をもった正規分布を用いることが一般的である．エンコーダネットワークは事後分布の平均・分散をデータ $\mathbf{x}$ から求め，デコーダは潜在変数 $\mathbf{z}$ からデータの再構成を出力する．

3. 提案手法

提案手法のモデルの概要を図 1 に示す．まず，モデルは入力として音の波形 $\mathbf{x}$ を受け取り，フレーム系列 $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_T)$ に切り分ける．そして，モデルはデータから潜在変数の系列 $\mathbf{z} = (\mathbf{z}_1, ..., \mathbf{z}_T)$ を推論し，潜在変数からデータを再構成しようとする（$\hat{\mathbf{x}}$）．提案手法では VAE の事前分布と事後分布を RNN によって制御する variational recurrent neural network（VRNN）を用いる [11]．最後に，デコーダの出力を微分可能な信号処理モジュールで構成される differentiable synthesizer[5] に入力して楽音波形を合成する．

3.1 Variational Recurrent Neural Network

提案手法ではフレーム間をまたいだ音色の変動を VRNN によってモデル化する．同様の形式で VRNN を動画生成に用いた研究 [12] では，VAE によって動画の各フレームの空間的構造をモデル化して，RNN によって動画の潜在的な変動をモデル化している．

通常の VAE では事前分布を標準正規分布として固定するが，VRNN では式 (5) のように，時刻 t における事前分布の平均 $\boldsymbol{\mu}_{0,t}$ と標準偏差 $\boldsymbol{\sigma}_{0,t}$ を RNN の状態変数 $\mathbf{h}_{t-1}$ から計算する．

$$\begin{aligned} \mathbf{z}_t &\sim \mathcal{N}(\boldsymbol{\mu}_{0,t}, \mathrm{diag}(\boldsymbol{\sigma}_{0,t})) \\ [\boldsymbol{\mu}_{0,t}, \boldsymbol{\sigma}_{0,t}] &= \varphi^{\mathrm{prior}}(\boldsymbol{h}_{t-1}) \end{aligned} \tag{5}$$

ここで φ^{prior} はニューラルネットワークとして実装される．この変動する事前分布によって，VAE が楽音の時間的構造についての情報を用いることができるようになる．また，VAE の目的関数（式 (4)）の KL ダイバージェンス項によって，事前分布は事後分布を予想するように訓練される．つまり，この事前分布によって潜在変数の系列のサンプリングを行うことで，新しい楽音の生成が可能になる．

さらに，RNN を楽音の属性を表すベクトル $\mathbf{a}$ によって条件づけることで，様々な楽器の変動をモデル化し，楽音生成の制御が可能になる．RNN は式 (6) のように更新される．

$$\mathbf{h}_t = f(\mathbf{z}_t, \mathbf{a}, \mathbf{h}_{t-1}), \tag{6}$$

ここで，f は gated recurrent unit（GRU）として実装する．

また，事後分布のパラメータも式 (7) の通りに RNN の情報から求める．

$$\begin{aligned} \mathbf{z}_t \mid \mathbf{x}_t &\sim \mathcal{N}(\boldsymbol{\mu}_{z,t}, \mathrm{diag}(\boldsymbol{\sigma}_{z,t})) \\ [\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}_{z,t}] &= \varphi^{\mathrm{post}}(\varphi^{\mathrm{enc}}(\mathbf{x}_t), \mathbf{h}_{t-1}), \end{aligned} \tag{7}$$

ここで，φ^{enc} はエンコーダネットワークによって求められる．φ^{post} はエンコーダの出力を feature-wise linear modulation（FiLM）[13] 層に通すことで求める．FiLM 層のパ

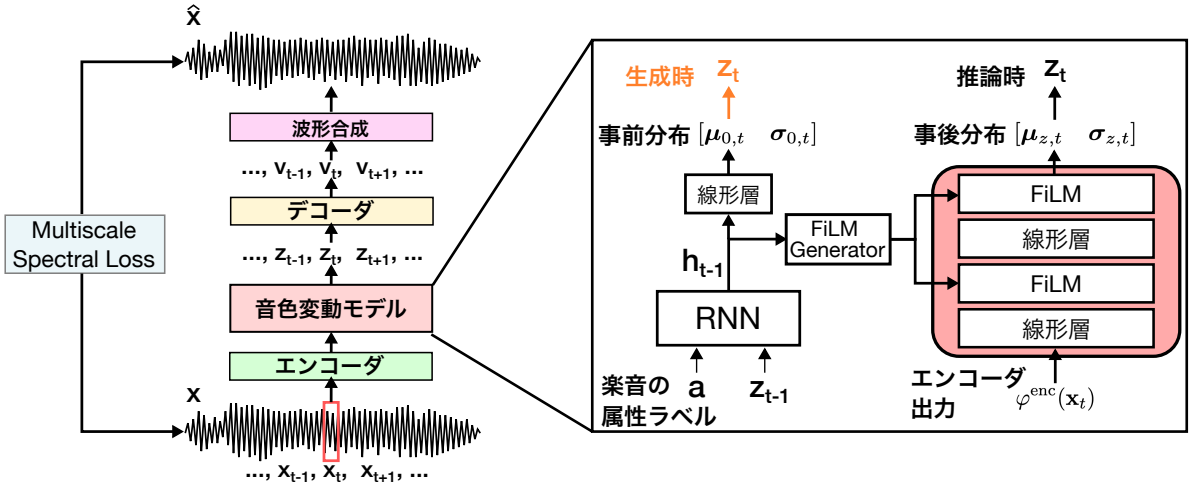

図 1 提案手法の概要図．フレーム単位で音を再構成する変分オートエンコーダが図の左側に，音色変動のモデルは図の右側に示されている．

ラメータを RNN の状態変数 $\mathbf{h}_{t-1}$ によって生成することで，RNN の情報を取り入れる．

よって，変分下限の KL 項は式 (8) の通りになる．

$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[\sum_{t=1}^{T}\left(-D_{\mathrm{KL}}\left[q\left(\mathbf{z}_t \mid \mathbf{x}_t, \mathbf{z}_{<t}\right) \| p\left(\mathbf{z}_t \mid \mathbf{z}_{<t}\right)\right]\right)\right] \quad (8)$$

モデルの訓練時には，デコーダネットワークは事後分布からサンプリングした $\mathbf{z}_t$ を入力として時点 t での音の合成パラメータ $\mathbf{v}_t$ を出力する．なお，本モデルと同様に Differentiable synthesizer を AE モデルで用いた DDSP[5] ではデコーダに楽音の基本周波数・ラウドネスの情報を与えることで再構成性能の向上を図った．また，最初に提案された VRNN[11] では RNN の状態変数 $\mathbf{h}_t$ をデコーダの入力とした．これに対し，提案手法のデコーダでは外部条件を用いずに，潜在変数 $\mathbf{z}_t$ の情報のみで音の再構成を行うように学習する．これにより，出力 $\hat{\mathbf{x}}_t$ が $\mathbf{z}_t$ のみによって決まるため，出力音の特徴について潜在空間の解析・可視化が可能になる．

3.2 Differentiable Synthesizer

提案手法では微分可能な信号処理モジュールで構成される differentiable synthesizer を用いて合成パラメータ $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_T)$ から音の波形 $\hat{\mathbf{x}}$ を出力する．本研究では DDSP[5] と同じ，倍音オシレータとフィルター付きノイズジェネレーターで構成されるシンセサイザーを用いた．このシンセサイザーは正弦波＋ノイズモデル [14] に基づいているが，正弦波の周波数が基本周波数の倍数に固定されている．これにより，一般的な楽音の構造に基づく強い帰納バイアスをモデルに課し，音色についての表現を獲得しやすくなる．また，倍音オシレータには楽音の基本周波数を条件として与えるため，潜在変数から基本周波数の影響が取り除かれる．スムーズな出力を得るため合成パラメータ $\mathbf{v} = (\mathbf{v}_1, ..., \mathbf{v}_T)$ に対し線形アップサンプリングを行う．

3.3 Multi-scale Spectral Loss

音のオートエンコーダモデルの再構成誤差として出力波形と元の波形の平均二乗誤差を用いると，音の知覚的には無視できる位相成分の影響が大きくなってしまう．そのため，スペクトログラムによる再構成誤差が提案されている [15]．本研究では式 (9) で表される Multi-scale spectral loss を用いる．

$$\mathcal{L}_i = |S_i(x) - S_i(\hat{x})|_1 + \alpha \left|\log S_i(x) - \log S_i(\hat{x})\right|_1 \quad (9)$$

S_i は i 番目の FFT サイズでのパワースペクトログラムを表す．実験では [64, 128, 256, 512, 1024, 2048] の FFT サイズを用い，各フレームを 75%オーバーラップさせた．対数誤差の重み α は 1.0 に設定した．

よって，式 (8) で表される事前分布と事後分布の KL ダイバージェンスを $\mathcal{L}_{\mathrm{KL}}$，KL ダイバージェンス項の重みを β とすると，モデル全体の目的関数は式 (10) のようになる．

$$\mathcal{L} = \beta \mathcal{L}_{\mathrm{KL}} + \sum_i \mathcal{L}_i(\mathbf{x}, \hat{\mathbf{x}}), \quad (10)$$

4. 実験

本節ではモデルの訓練に関する詳細，提案手法による楽音の再構成の結果，及び新しい楽音の生成の結果を示す．なお，提案手法による楽音の再構成・生成結果の例はウェブサイト[*1]にアップロードした．

4.1 データセット

モデルの学習データには NSynth Dataset[8] を用いた．このデータセットの各エントリーは 3 秒間ホールドした MIDI ノート 1 つ分を 4 秒，標本周波数 16kHz で録音した音データである．データセットの各エントリーは音源によって “acoustic”，“electronic”，“synthetic” の 3 カテゴ

[*1] https://hyakuchiki.github.io/timbredynamicswebpage/

表 1 スペクトログラムの再構成性能における提案手法とベースラインの比較

	RMS ($\times 10^5$)	LSD ($\times 10^2$)
VAE	8.32	5.68
VRNN	8.55	5.45
VRNN-inst	8.22	5.39
VRNN-attr	**8.14**	**5.34**

リに分けられているが，ここでは電気的な増幅を用いない楽音を含む “acoustic” カテゴリのみ用いた．また，音高が C2 より下，もしくは B5 より上の音は除外した．さらに，今回用いる分割した．また，differentiable synthesizer の制約上出力できない音を排除するため，属性に “reverb”，“percussive” タグがついているエントリーは除外した．最終的に残った楽音 22595 個のデータセットを訓練用・検証用・テスト用で 8:1:1 に分割した．また，differentiable synthesizer での波形合成で必要となる楽音の基本周波数は CREPE[16] によって事前に計算した．

また，半自動的なアノテーションによってデータセットの各エントリーには “instrument family”（楽器族），及び “note quality”（音の質感）で構成される属性タグがつけられている．今回用いるデータセットに含まれる楽器族は “brass”，“flute”，“guitar”，“keyboard”，“mallet”，“reed” の 6 つとした．各楽器族には様々な楽器・奏法で演奏した楽音が含まれている．

4.2 モデル詳細

本項では，モデルの詳細を示す．まず，モデルの入力波形からメルスペクトログラムを抽出する．FFT 長は 1024 サンプル，ホップサイズは 512 サンプルとし，各楽音から 124 個のフレームが得られる．次に，エンコーダでは周波数次元をフィルタリングする 1 次元の畳み込み層 4 層，その後に layer normalization 層を挟んだ線形層を用いる．1 次元畳み込み層によって，基本周波数に対して不変な音色の表現を学習することが狙いである．さらに，エンコーダの出力を FiLM [13] 層を挟んだ 3 層の線形層に入力することで，事後分布のパラメータ $\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}_{z,t}$ を計算する．潜在変数の次元は 16 とした．予備実験により，これより次元を減らすと再構成の性能が落ちることが確認された．そして，デコーダは layer normalization を挟んだ 3 層の線形層で構成される．

変分モデルの学習では，事後分布が事前分布に等しくなってしまう Posterior Collapse 現象がしばしば起きる．この問題は通常の VAE でも見られるが，変動する事前分布を用いる提案手法ではより顕著になる．本研究では誤差の KL 項の重み β をゼロから徐々に上げることで，この問題を回避する．β が小さすぎると，モデルの事前分布が事後分布の予測を行うようにならないが，β が大きすぎると事後分布に音色についての情報がエンコードされなくなってしまう．予備実験の結果，100 エポックをかけて β を 0 から 1 に上げた場合再構成の性能が最も高かったため，本実験ではこの通り β を設定する．ただし，属性タグを用いないモデルに関しては Posterior Collapse が顕著だったため，β を 10 エポック目から上げた．

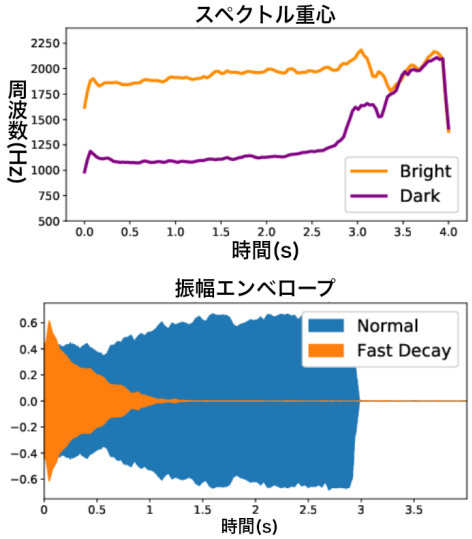

図 2 VRNN-attr で音の明るさについて指定した場合のスペクトル重心の軌跡を上図，音の減衰の早さについて指定した場合の振幅エンベロープに対する影響を下図に示す．

4.3 楽音の再構成

まず，提案手法の事前・事後分布により楽音の構造が学習されるかどうかを検証するため，楽音の再構成への影響を測った．属性タグなしのモデル（以下，VRNN），楽器族ラベルを入力したモデル（以下，VRNN-inst），楽器族と音の質感ラベルの両方を入力したモデル（以下，VRNN-attr）の 3 つと，通常の VAE によるベースライン（以下，VAE）による再構成誤差を表 1 に示す．元の音とモデルの出力の間の Log-spectral distortion（LSD）と平均二乗誤差（MSE）を計算した．提案手法の 3 つのモデルはベースラインよりも良い再構成性能を見せたほか，音色変動モデルに条件を与えることで再構成の質が改善された．これにより，音色変動についてのモデルが楽音の再構成に有用であることが確認できた．

なお，筆者らが出力音を主観的に評価した結果，各モデルは “brass” などの励起状態が長く続く楽音に関しては再構成に成功するが，“keyboard” などのノイズが冒頭に多く含まれる楽音の再構成の品質が低いことがわかった．この問題はベースラインモデルでも顕著なことから，differentiable synthesizer のノイズジェネレーターのパラメータの学習が困難であるということが考えられる．

図 3 VRNN-inst により生成された楽音の分類結果．混同行列の縦軸はモデルに与えた条件のラベルであり，横軸は予測されたラベル．各クラスについて 512 個の楽音を生成した．

4.4 楽音の生成

VRNN-attr により質感ラベルについて指定して楽音の生成を行った結果を図 2 に示す．スペクトル重心は音の「明るさ」に関連する特徴量であり，“bright” のタグをモデルに入力した場合のスペクトル重心は “dark” を入力した場合より高いことが期待される．生成された楽音から特徴量を抽出した結果，VRNN-attr による楽音の質感の制御はスペクトルに関連する質感と音の時間的特徴に関連する質感の両方について有効であったと言える．また，前述のウェブサイトにもアップロードされているモデルの出力から，1 つの楽器族に対して多様な楽音が生成できていることが確認できた．

楽器族による条件付き楽音生成の有用性を測るべく，VRNN-inst モデルの出力楽音を分類した結果を図 3 に示す．入力をメルスペクトログラムとする畳み込み層 4 層のネットワークによって楽器族ラベルを予測した．分類器は提案手法のモデルの訓練に用いたものと同じ NSynth データセットで訓練し，検証用データセットに対して正解率 99.4%であった．楽音生成時，differentiable synthesizer の基本周波数は比較的楽器族ラベルが均等に分散している C5=523.3Hz に設定した．

結果から，条件付き楽音生成は “reed”，“keyboard” などの一部の楽器族を除いてある程度効果的であったことが確認できた．“keyboard” の条件を与えた楽音が生成できないことは，4.3 項で述べたようにノイズジェネレーターが “keyboard” に分類される音のアタックに存在するノイズの再構成に失敗することが原因であると考えられる．また，“reed” の条件を与えると，比較的音色が似ている “brass” のような楽音が生成されてしまった．

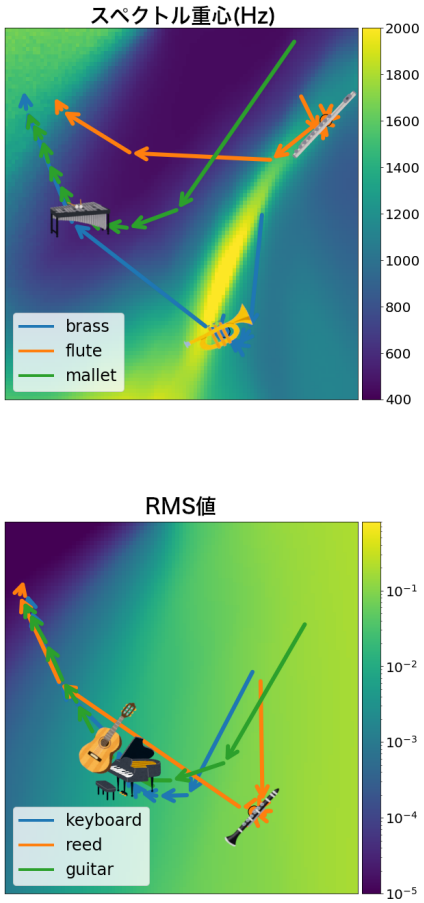

図 4 潜在空間における事前分布の軌跡と音響特徴量の分布．背景は RMS 値（上図），スペクトル重心（下図）のヒートマップである．楽器のアイコンは条件として与えた楽器族クラスを表し，軌跡の中点に位置する．

4.5 潜在空間の性質

提案手法によって獲得された音色変動の性質を可視化するため，VRNN-inst モデルから事前分布から自己回帰的にサンプリングした潜在変数の軌跡を図 4 に示す．各軌跡は異なる楽器族ラベルを指定して生成されている．16 次元の潜在空間を可視化するため，主成分分析により潜在変数の次元を削減した．さらに軌跡を見やすくするため，ガウシアンフィルタによる平滑化を行い，ダウンサンプリングした．また，軌跡を出力音と結びつけて解釈するため，スペクトル特徴量の分布を背景に表示した．楽音の生成モデルでは瞬時的な音色 [3]，及び楽器エンベディング [17] について同様の可視化を行った例がある．

事前分布の軌跡の終点は必ず左上の領域に位置した．RMS 値のヒートマップと合わせて考えると，この領域は各楽音の最後にある無音に該当することが推測される．ヒートマップ上で “reed” などの励起が一定時間続く楽器族は高エネルギーの領域にとどまるが，“guitar” や “keyboard” などの励起が瞬時的な楽器はより早く低エネルギーの領域

に遷移した．また，スペクトル重心のヒートマップからは，“trumpet” や “flute” は励起中に明るい音色を保っているが，打楽器である “mallet” はすぐに倍音成分を失っていることが読み取れる．これらの性質は実際の楽器の性質にも一致しており，提案手法の音色変動のモデリングにおける一定の有用性が示された．

5. おわりに

本研究では楽音の生成を可能にする音色変動の変分モデルを示した．VAE により瞬時的な音色をモデル化し，音色変動をモデル化する RNN により VAE の事前分布と事後分布を制御した．また，音色変動のモデルに楽音の属性ラベルを条件として与えることで，再構成性能を改善し，楽音の条件づき生成を行えることを示した．提案手法では音色の情報を瞬時的な音色と音色変動の 2 つに分けてモデリングした．解釈性の観点から考えても，この 2 つが分離されたような表現は楽音全体の音色を表す全域的な表現より有用であると考えられる．音色変動に関するさらなる研究により，表現力の高い楽音の生成モデルが実現すると考えられる．

参考文献

[1] Sarroff, A. M. and Casey, M.: Musical audio synthesis using autoencoding neural nets, *Proceedings of the 40th International Computer Music Conference*, pp. 1411–1417 (2014).
[2] Roche, F., Hueber, T., Limier, S. and Girin, L.: Autoencoders for music sound modeling: a comparison of linear, shallow, deep, recurrent and variational models, *Proceedings of the Sound and Music Computing Conference* (2019).
[3] Esling, P., Chemla-Romeu-Santos, A. and Bitton, A.: Generative Timbre Spaces: Regularizing Variational Auto-Encoders with Perceptual Metrics, *Proceedings of the International Conference on Digital Audio Effects*, (online), available from ⟨https://github.com/acids-ircam/⟩ (2018).
[4] Subramani, K., Rao, P. and D'Hooge, A.: Vapar Synth - A Variational Parametric Model for Audio Synthesis, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (2020).
[5] Engel, J., Hantrakul, L., Gu, C. and Roberts, A.: DDSP:Differentiable Digital Signal Processing, *Proceedings of the International Conference on Learning Representations* (2020).
[6] Hourdin, C., Charbonneau, G. and Moussa, T.: A Multidimensional Scaling Analysis of Musical Instruments' Time-Varying Spectra, *Computer Music Journal*, Vol. 21, No. 2, pp. 40–55 (1997).
[7] van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, *ISCA Speech Synthesis Workshop* (2016).
[8] Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K. and Norouzi, M.: Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders, *Proceedings of the International Conference on Machine Learning*, pp. 1068–1077 (2017).
[9] McAdams, S. and Giordano, B. L.: The Perception of Musical Timbre, *The Oxford Handbook of Music Psychology* (2008).
[10] Bitton, A., Esling, P. and Harada, T.: Neural Granular Sound Synthesis, *ArXiv Preprint, ArXiv ID: 2008.01393* (2020).
[11] Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. and Bengio, Y.: A Recurrent Latent Variable Model for Sequential Data, *Advances in Neural Information Processing Systems*, pp. 2980–2988 (2015).
[12] He, J., Lehrmann, A., Marino, J., Mori, G. and Sigal, L.: Probabilistic Video Generation Using Holistic Attribute Control, *Proceedings of the European Conference on Computer Vision* (2018).
[13] Perez, E., Strub, F., De Vries, H., Dumoulin, V. and Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer, *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3942–3951 (2018).
[14] Serra, X.: Musical Sound Modeling with Sinusoids plus Noise, *Musical Signal Processing* (Roads, C., Pope, S., Picialli, A. and Poli, G. D., eds.), pp. 91–122 (1997).
[15] Défossez, A., Zeghidour, N., Usunier, N., Bottou, L. and Bach, F.: SING: Symbol-to-instrument neural generator, *Advances in Neural Information Processing Systems*, Vol. 2018-Decem, pp. 9041–9051 (2018).
[16] Kim, J. W., Salamon, J., Li, P. and Bello, J. P.: CREPE: A Convolutional Representation for Pitch Estimation, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (2018).
[17] Kim, J. W., Bittner, R., Kumar, A. and Bello, J. P.: Neural Music Synthesis for Flexible Timbre Control, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 176–180 (2019).