Vol.2021-MUS-131 No.56
Vol.2021-SLP-137 No.56
2021/6/19

変分自己符号化器を用いた距離学習による楽器音の音高・音色分離表現

田中 啓太郎[1,a)]　錦見 亮[2,b)]　坂東 宜昭[3,c)]　吉井 和佳[2,d)]　森島 繁生[4,e)]

概要：本稿では，任意の楽器音を音高と音色の潜在表現に分離するための表現学習手法について述べる．このような分離はこれまでにも，変分自己符号化器（variational autoencoder, VAE）を用いて，特に予め指定された楽器群を対象に試みられてきた．しかし，得られる潜在表現は人間の知覚に合致しておらず，空間を直感的に解釈することが困難であった．この問題を解決するため，本研究では VAE の各潜在空間に対して距離学習手法を導入し，類似した（していない）音高または音色同士が，潜在空間において近く（遠く）なるように埋め込みを行う．具体的には，同じ（異なる）音高または音色間の潜在距離が最小（最大）となるように，VAE の学習において対比的損失関数を追加する．さらに，実際の音高あるいは音色名ではなく，二つの音の音高あるいは音色が同一かどうかのみの情報を用いた，弱教師あり学習を行う．これにより，未知楽器に対する汎化性能の向上を実現する．実験では，提案手法によって未知楽器に対しても，音高と音色がクラスター化された，より良い構造の分離表現を獲得できることを確認した．

1. はじめに

分離表現学習とは，ある一つの特徴のみに影響するような独立した因子の組み合わせによって，複雑なデータを表現するものである．これにより潜在表現が解釈可能なものになるとともに，データ生成時には我々人間の直感に従って各因子を操作することができるようになる．分離表現学習における有名なアプローチは，深層潜在変数モデルを敵対的生成ネットワーク（generative adversarial network, GAN） [1] や変分自己符号化器（variational autoencoder, VAE） [2–5] の枠組みで訓練するというものである．代表的な対象である画像データ [6–8] の他にも，文章データ [9] や音声データ [10] も取り扱われている．

音楽情報処理分野においては，音の三要素である音量，音高，音色に音楽音響信号を分離することが，楽曲のスタイル変換 [11] や自動生成 [12]，分析 [13]，推薦 [14] の基礎をなすものとして重要視されている．このうち音量は計算可能であるため，音高と音色への分離が主な研究対象となる．例えば Mor ら [15] は自己符号化器を用いて，音楽音響信号中の音高は変えずに音色だけを操作する楽曲変換手法を提案した．彼らはこれを，各音色に対応する複数のデコーダを用いて実現した．Bitton ら [16] は β-VAE を拡張することで，単独のエンコーダ・デコーダ対による実現を可能にしている．しかしながら，いずれの手法においてもその潜在表現は VAE が楽器音を容易に生成するためのものであり，人間の知覚とはかけ離れたものであった．この問題を解決するため，Esling ら [17] は多次元スケーリング法を用いることで，人間の知覚に沿うような潜在音色空間を学習する VAE に基づく手法を提案した．

以上の先行研究は全て音色表現のみに焦点を当てており，音高表現は所与のものとして仮定されている．Hung ら [18] は楽曲のスタイル変換を目的として，初めて音高と音色双方の分離表現を試みた．最近では，Luo ら [19, 20] が混合ガウス VAE を用いた楽器音の音高・音色分離表現

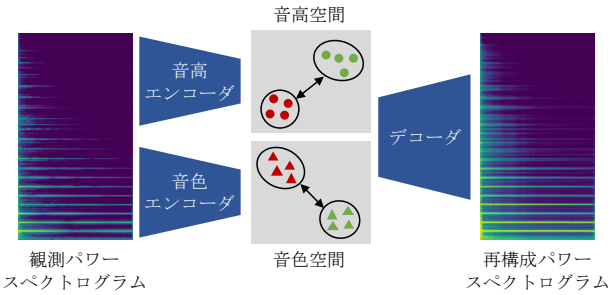

図 1　提案する対比的損失関数を用いた VAE の概念図

1　早稲田大学大学院先進理工学研究科
2　京都大学大学院情報学研究科
3　産業技術総合研究所人工知能研究センター
4　早稲田大学理工学術院総合研究所
a)　phys.keitaro1227@ruri.waseda.jp
b)　nishikimi@sap.ist.i.kyoto-u.ac.jp
c)　y.bando@aist.go.jp
d)　yoshii@i.kyoto-u.ac.jp
e)　shigeo@waseda.jp

手法を提案している．が，ここで使用されているモデルでは，各音高および音色（楽器）に対応するガウス分布を用意しているため，学習データに含まれない未知の音高や音色を取り扱うことはできない．

あらゆる音高と音色を対象とした任意の楽器音を取り扱う必要性から，我々はサンプル間の類似度を潜在空間における距離として表現する距離学習 [21–25] に着目する．その基本的なアプローチは，深層ニューラルネットワーク（deep neural network, DNN）を用いて，類似したサンプル同士は潜在空間において近く，逆に類似していないサンプル同士は潜在空間において遠くなるように学習を行うというものである．このアプローチの最大の長所は，DNN の学習にあたって具体的なカテゴリラベルは使用せず，二つのサンプル間のカテゴリの一致・不一致の情報のみを使用するため，未知のサンプル（音高や音色など）を取り扱うことが可能であるという点にある．このような未知のサンプルを見据えた学習は，zero-shot learning [26, 27] とも呼ばれ，近年注目を集めている．

本稿では，任意の楽器音を音高と音色の潜在表現に分離するための VAE に基づく表現学習手法を提案する（**図 1**）．我々の VAE は，観測スペクトログラムから潜在音高および音色表現を推論するための独立した二つのエンコーダと，これらの潜在空間からスペクトログラムを生成するための一つのデコーダで構成される．潜在音高および音色表現をクラスター構造化によって解釈可能にするため，同じ音高あるいは音色同士は近く，異なる音高あるいは音色同士は遠く埋め込まれるように，対比的損失関数を導入する．また，VAE は弱教師あり学習の枠組みで学習を行い，音高あるいは音色が同一かどうかの情報のみを使用する．これにより，VAE は具体的な音高ラベルと音色ラベルに依存しなくなるため，未知の音高と音色を持つ楽器音に対しても分離表現の獲得が可能になる．

本研究の主な貢献は，任意の楽器音に対する音高・音色分離表現を目的とした対比的損失関数に基づく VAE の学習にある．音高および音色表現は，具体的なラベルの事前定義なしに獲得することができる．実験によって，距離学習手法が知覚的な音高および音色の類似度を反映した潜在表現の獲得に効果的であることを示す．潜在表現は VAE に基づいて獲得されているため，音高および音色の情報の他にも，距離学習だけでは獲得できないような豊富な情報（トレモロやヴィブラート）を潜在的に含むと考えられる．

2. 提案手法

提案手法は，音高と音色の分離のための VAE による弱教師あり距離学習に基づく．単独楽器音の観測パワースペクトログラム $\mathbf{X} = \mathbf{x}_{1:T} \in \mathbb{R}_+^{F\times T}$ を入力として，二つの潜在音高および音色表現 $\mathbf{Z}^{\mathrm{p}} = \mathbf{z}^{\mathrm{p}}_{1:T} \in \mathbb{R}^{H\times T}$ および $\mathbf{Z}^{\mathrm{t}} = \mathbf{z}^{\mathrm{t}}_{1:T} \in \mathbb{R}^{H\times T}$ ($\mathbf{Z} = \{\mathbf{Z}^{\mathrm{p}}, \mathbf{Z}^{\mathrm{t}}\}$) を介した後，再構成パワースペクトログラム $\mathbf{Y} = \mathbf{y}_{1:T} \in \mathbb{R}_+^{F\times T}$ を出力する VAE を訓練することが目的である．ただし，T は時間フレーム長を，F は周波数ビン数を，H は各潜在空間の次元を表す．また，$\mathbf{x}_t \in \mathbb{R}_+^F$ と $\mathbf{y}_t \in \mathbb{R}_+^F$ は時刻 t における観測および再構成パワースペクトルを表し，$\mathbf{z}^{\mathrm{p}}_t \in \mathbb{R}^H$ と $\mathbf{z}^{\mathrm{t}}_t \in \mathbb{R}^H$ は時刻 t における音高および音色空間の潜在変数を表す．

2.1 生成モデル

観測スペクトログラム $\mathbf{X}$ に対する確率モデルを，潜在表現 $\mathbf{Z} = \{\mathbf{Z}^{\mathrm{p}}, \mathbf{Z}^{\mathrm{t}}\}$ を用いて式 (1) のように定式化する．

$$p_\theta(\mathbf{X}, \mathbf{Z}) = p_\theta(\mathbf{X}|\mathbf{Z})p(\mathbf{Z}) \tag{1}$$

ただし，$p_\theta(\mathbf{X}|\mathbf{Z})$ は $\mathbf{Z}$ の $\mathbf{X}$ に対する尤度関数，$p(\mathbf{Z})$ は $\mathbf{Z}$ の事前分布，θ はモデルパラメータである．さらに，深層生成モデルである $p_\theta(\mathbf{X}|\mathbf{Z})$ を，式 (2) のように定式化する．

$$p_\theta(\mathbf{X}|\mathbf{Z}) = \prod_{f=1}^{F}\prod_{t=1}^{T} \mathrm{Exponential}(x_{ft}|[\boldsymbol{\kappa}_\theta(\mathbf{Z})]_{ft}) \tag{2}$$

$$= \prod_{f=1}^{F}\prod_{t=1}^{T} \frac{1}{[\boldsymbol{\kappa}_\theta(\mathbf{Z})]_{ft}} \exp(-x_{ft}/[\boldsymbol{\kappa}_\theta(\mathbf{Z})]_{ft}) \tag{3}$$

ただし，$\boldsymbol{\kappa}_\theta(\mathbf{Z})$ は $\mathbf{Z}$ を入力とし θ によりパラメタライズされた DNN の FT 次元の出力である．また，$[\mathbf{A}]_{ij}$ は $\mathbf{A}$ の ij 番目の要素を表す．$p(\mathbf{Z})$ には式 (4) のような標準ガウス分布を設定する．

$$p(\mathbf{Z}) = p(\mathbf{Z}^{\mathrm{p}})p(\mathbf{Z}^{\mathrm{t}}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^{\mathrm{p}}_t|\mathbf{0}_H, \mathbf{I}_H)\mathcal{N}(\mathbf{z}^{\mathrm{t}}_t|\mathbf{0}_H, \mathbf{I}_H) \tag{4}$$

ただし，$\mathbf{0}_H$ は H 次元の零ベクトルを，$\mathbf{I}_H$ は $H \times H$ の単位行列を表す．

2.2 VAE に基づく学習

与えられた観測スペクトログラム $\mathbf{X}$ に対して，尤度最大化の観点から潜在表現 $\mathbf{Z}$ とモデルパラメータ θ を推論する．DNN に基づく我々の生成モデル定式化において事後分布 $p_\theta(\mathbf{Z}|\mathbf{X})$ を直接計算することは困難であるため，VAE を用いて近似計算を行う．具体的には，ϕ によってパラメタライズされる変分分布 $q_\phi(\mathbf{Z}|\mathbf{X}) = q_\phi(\mathbf{Z}^{\mathrm{p}}|\mathbf{X})q_\phi(\mathbf{Z}^{\mathrm{t}}|\mathbf{X})$ を導入し，$q_\phi(\mathbf{Z}|\mathbf{X})$ の $p_\theta(\mathbf{Z}|\mathbf{X})$ に対するカルバック・ライブラー（Kullback-Leibler, KL）情報量の最小化によって最適化を行う．本稿では，$q_\phi(\mathbf{Z}|\mathbf{X})$ は ϕ によって式 (5) のようにパラメタライズされる DNN を用いて構築する．

$$q_{\phi^*}(\mathbf{Z}^*|\mathbf{X}) = \prod_{t=1}^{T} \mathcal{N}(\mathbf{z}^*_t|[\boldsymbol{\mu}^*_{\phi^*}(\mathbf{X})]_t, [\boldsymbol{\sigma}^{*2}_{\phi^*}(\mathbf{X})]_t) \tag{5}$$

ただし，$*$ は “p” または “t” を表し，$\boldsymbol{\mu}^*_{\phi^*}(\mathbf{X})$ と $\boldsymbol{\sigma}^{*2}_{\phi^*}(\mathbf{X})$ はパラメータ ϕ^* に依存した FT 次元の DNN の出力である．深層生成モデルと同様に，DNN の出力は確率分布のパラメータを表す．

モデルパラメータ θ に関して $\log p_\theta(\mathbf{X})$ を直接最大化する代わりに，$q_\phi(\mathbf{Z}|\mathbf{X})$ の導入を通して変分下限 $\mathcal{L}^{\text{vae}}$ を式 (6) により最大化する．

$$\mathcal{L}^{\text{vae}}=\mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X})}[\log p_\theta(\mathbf{X}|\mathbf{Z})]-\mathcal{D}_{\text{KL}}(q_\phi(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z})) \quad (6)$$

ただし，$\phi=\{\phi^{\text{p}},\phi^{\text{t}}\}$ であり，また，$q_\phi(\mathbf{Z}|\mathbf{X})=p_\theta(\mathbf{Z}|\mathbf{X})$ の時に限って $\mathcal{L}^{\text{vae}}$ は最大化される．なお，$p_\theta(\mathbf{Z}|\mathbf{X})$ の計算困難性のためこの条件を直接満たすことはできないことに注意されたい．式 (6) 中の $\log p_\theta(\mathbf{X})$ と $\mathcal{L}^{\text{vae}}$ との差は $q_\phi(\mathbf{Z}|\mathbf{X})$ の $p_\theta(\mathbf{Z}|\mathbf{X})$ に対する KL 情報量に一致するため，KL 情報量の最小化と $\mathcal{L}^{\text{vae}}$ の最大化は等価である．$\mathcal{L}^{\text{vae}}$ のうち期待値に関する項は，リパラメタライゼーショントリック [28] を用いて計算する．また，KL 情報量は解析的に計算可能であるため，ネットワークは勾配法を用いて最適化することができる．

VAE に基づく訓練によって，観測スペクトログラム $\mathbf{X}$ を統計的に独立した二つの潜在表現 $\mathbf{Z}^{\text{p}}$ と $\mathbf{Z}^{\text{t}}$ に分離することはできる．しかしながら，ここで獲得される表現は必ずしも有用でなく，潜在表現に知覚的な類似度が反映されていない．我々はこの問題を，潜在空間に距離学習を導入し，知覚の反映された分離を促進することで解決する．

2.3 分離のための対距離学習

エンコーダによって，観測スペクトル $\mathbf{x}_t$ は潜在変数 $\mathbf{z}_t^{\text{p}}$ および $\mathbf{z}_t^{\text{t}}$ に変換される．理想的には，同じ音高または音色（すなわち楽器）同士の潜在表現は互いに近く，異なる音高または音色同士の潜在表現は互いに遠く，各潜在空間において位置するべきである．

このような潜在表現を得るために，我々は対ごとの距離を活用する．楽器音の N 個のスペクトログラム $\{\mathbf{X}_n\}_{n=1}^N$ によるミニバッチ学習において，無作為に二つのスペクトログラム $\mathbf{X}_i$ および $\mathbf{X}_j$ $(i\neq j)$ を取り出すことを考える．ただし，N はバッチサイズを表す偶数であり，$i,j\in\{1,\ldots,N\}$ は学習サンプルを示している．各スペクトログラム対に対する潜在変数 $\mathbf{Z}_i^*$ および $\mathbf{Z}_j^*$ は，エンコーダによって独立に得られる．このもとで，距離学習を対比的損失関数 $\mathcal{L}_c^{\text{p}}$ および $\mathcal{L}_c^{\text{t}}$ を用いて行う．音高に対する損失関数 $\mathcal{L}_c^{\text{p}}$ は式 (7) のように計算される．

$$\mathcal{L}_c^{\text{p}}=\begin{cases}\sum_{i,j\in N}(\mathcal{D}_{ii}^{\text{p}}+\mathcal{D}_{jj}^{\text{p}}+\mathcal{D}_{ij}^{\text{p}}) \\ \qquad\qquad (i\text{ と }j\text{ が同じ音高の場合}) \\ \sum_{i,j\in N}(\mathcal{D}_{ii}^{\text{p}}+\mathcal{D}_{jj}^{\text{p}}-\mathcal{D}_{ij}^{\text{p}}) \quad (\text{それ以外})\end{cases} \quad (7)$$

ただし，$\mathcal{D}_{ii}^*,\mathcal{D}_{jj}^*,\mathcal{D}_{ij}^*$ は二つの潜在変数 $\mathbf{Z}_i^{\text{p}}$ および $\mathbf{Z}_j^{\text{p}}$ 間の距離の和であり，式 (8)–(10) のように定義される．

$$\mathcal{D}_{ii}^{\text{p}}=\sum_{t_1=1}^{T-1}\sum_{t_2=t_1+1}^{T}||\mathbf{z}_{it_1}^{\text{p}}-\mathbf{z}_{it_2}^{\text{p}}|| \quad (8)$$

$$\mathcal{D}_{jj}^{\text{p}}=\sum_{t_1=1}^{T-1}\sum_{t_2=t_1+1}^{T}||\mathbf{z}_{jt_1}^{\text{p}}-\mathbf{z}_{jt_2}^{\text{p}}|| \quad (9)$$

$$\mathcal{D}_{ij}^{\text{p}}=\sum_{t_1=1}^{T}\sum_{t_2=1}^{T}||\mathbf{z}_{it_1}^{\text{p}}-\mathbf{z}_{jt_2}^{\text{p}}|| \quad (10)$$

ただし，$||\cdot||$ はベクトルのユークリッド距離を表す．音色に対する損失関数 $\mathcal{L}_c^{\text{t}}$ も同様に計算される．これら対比的損失関数の値は，同じ音高または音色の潜在変数同士が互いに遠い，あるいは異なる音高または音色の潜在変数同士が互いに近い場合に大きな値を取る．したがって，提案する生成モデルの潜在空間を知覚的に解釈可能なものへと近づけることができる．

我々は提案ネットワークを弱教師あり学習によって訓練する．すなわち，VAE に与えられる観測スペクトログラム対の音高と音色が同一かどうかの情報のみが必要であり，実際の音高名と音色名（楽器名）であるラベルは必要でない．実際の学習は，式 (6) の VAE の損失関数と式 (7) の対比的損失関数を組み合わせた，式 (11) で表される総合損失関数 $\mathcal{L}^{\text{total}}$ によって行う．

$$\mathcal{L}^{\text{total}}=-\mathcal{L}^{\text{vae}}+\alpha\mathcal{L}_c^{\text{p}}+\beta\mathcal{L}_c^{\text{t}} \quad (11)$$

ただし，α と β は二つの対比的損失関数の重みを調整するためのハイパーパラメータである．

3. 評価実験

本章では，分離のための提案手法の性能を実験によって，定性的および定量的に評価する．

3.1 使用データ

評価にあたっては，RWC 研究用音楽データベース [29] に収録されている楽器音のうち，尺八，ソプラノ，アルトを除いたものを使用した．データベース中の各ファイルは楽器名のアノテーションが付いており，当該楽器の演奏可能な範囲で全ての音高が録音されている．我々は，各ファイルの音響信号をミュート検出によって自動的に各音高に区切った後，各音冒頭の無音部分をオンセット検出によって削除した．そのうち，C3 から B5 までの音高を持つ音を選び，これらを評価実験に使用した．以上のようにして得られた音（40914 音，50 楽器）を，訓練セット（29957 音，40 楽器）と評価・検証セット（10957 音，10 楽器）とに分けた．交差検証および学習終了時期の決定のため，評価・検証セットを各 5 楽器から構成される二つのサブセットへとさらに分割した．以上三つのセットは音高を共有しているが，楽器は共有していない．

全ての音は 22050Hz でリサンプリングし，各音の最初の 2 秒のみを使用した．スペクトログラム作成にあたっては，窓幅 1024 サンプル，シフト幅 256 サンプルのハン窓による短時間フーリエ変換（short-time Fourier transform, STFT）を適用した（$F=513$，$T=173$）．各スペクトログラムは，平均パワーが 1 となるように正規化を行った．

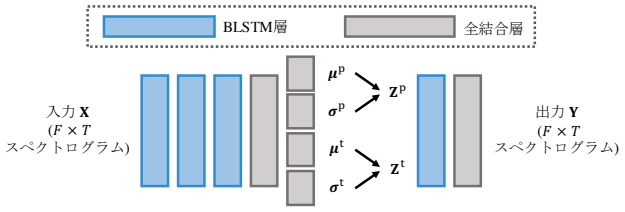

図 2 提案手法の VAE の構成

表 1 各空間における密集度と発散度

手法	音高表現		音色表現	
	密集度	発散度	密集度	発散度
通常の VAE	3.334	2.279	3.640	1.541
提案手法の VAE	**2.891**	**3.551**	**3.420**	**2.654**

3.2 モデル構成

我々の VAE に基づく手法では，音の時間的特徴を捉えるため，エンコーダとデコーダに双方向長短期記憶（bidirectional long short-term memory, BLSTM）を用いた（**図 2**）．エンコーダは，三層のノード数 2×300 の BLSTM と，全結合層から構成される．エンコーダの各 BLSTM については，ドロップアウト率を 0.3 に設定した．共有された全結合層は，600 次元を 256 次元に落としている．BLSTM とこの全結合層の出力は，全て leaky ReLU 層を介している．続く四つの全結合層は，それぞれ独立に 256 次元を $H = 16$ 次元まで落としており，潜在変数の平均と分散を表している．デコーダは一層のノード数 2×300 の BLSTM と，それに続く全結合層から構成される．BLSTM の出力は，エンコーダと同様に leaky ReLU 層を介している．スペクトログラムの各時間周波数ビンは非負値を取るため，全結合層の出力に対してのみ softplus 関数を適用した．バッチサイズ N は 16 であり，学習は Adam [30] を学習率 0.001 で用いて行った．潜在空間における密集度と発散度（詳細は 3.3 を参照のこと）は，対比的損失関数の重み（α と β）の調整によって操作可能であった．我々は実験的に，提案手法が重み 0.5 以上ではスペクトログラムの再構成に失敗することを確認し，ともに 0.2 とした．

3.3 評価尺度

同じ音高あるいは音色に対する潜在変数同士の近さを表す密集度と，異なる音高あるいは音色に対する潜在変数同士の遠さを表す発散度を評価する．音高空間に対するこれらの評価尺度は，式 (12) および式 (13) のように計算する．

$$(\text{密集度}) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{9N_m}\sum_{n=1}^{N_m}\sum_{t=1}^{9}||\mathbf{z}^{\mathrm{p}}_{mnt} - \boldsymbol{\eta}^{\mathrm{p}}_{m}|| \quad (12)$$

$$(\text{発散度}) = \frac{2}{M(M-1)}\sum_{m_1=1}^{M-1}\sum_{m_2=m_1+1}^{M}||\boldsymbol{\eta}^{\mathrm{p}}_{m_1} - \boldsymbol{\eta}^{\mathrm{p}}_{m_2}|| \quad (13)$$

$$\boldsymbol{\eta}^{\mathrm{p}}_{m} = \frac{1}{9N_m}\sum_{n=1}^{N_m}\sum_{t=1}^{9}\mathbf{z}^{\mathrm{p}}_{mnt} \quad (14)$$

ただし，M は音高の数を，N_m は音高 m を持つ音の数を，$\boldsymbol{\eta}_m$ は音高 m に対する全ての潜在変数の平均を表す．音色に対しても同様に計算を行う．なお，各音の終盤における無音部分を除外するため，評価にあたっては各音の最初の 9 フレームのみを使用した．また，空間の大きさの影響を除外するため，各潜在空間は正規化した．

3.4 実験結果

表 1 に実験結果を示す．距離学習の導入によって，潜在音高および音色空間の双方において密集度は小さく，発散度は大きくなった．これより，提案手法の効力を持つことがわかる．**図 3** に，潜在音高および音色空間の t 分布型確率的近傍埋め込み法（t-distributed stochastic neighbor embedding, t-SNE） [31] による可視化結果を示す．提案手法が未知楽器に対して，音高および音色についてクラスター化されたより良い構造の分離表現を獲得できていることが見て取れる．**図 4** からは，提案手法は対比的損失関数を用いることで，ほとんどの音高および音色に対して密集度が改善（低下）していることがわかる．**図 5** の左上と右上との比較からは，対角成分付近の要素の値が小さく，対角成分から離れた要素の値は大きくなっていることがわかる．このことは，我々の提案手法が潜在音高空間において同じ音高同士は近く，異なる音高同士は遠くなるような潜在変数を構築できていることを示している．図 5 の下二つの図における比較からは，右図の非対角成分の値が左図の非対角成分の値よりも大きくなっていることがわかる．これらの結果は，提案手法が異なる音色同士を遠く埋め込むことに成功していることを示している．

4. おわりに

本稿では，任意の楽器音を音高と音色の潜在表現に分離するための VAE に基づく表現学習手法について述べた．各分離空間を操作するため，音の類似性に基づいた距離学習手法を用いた．さらに，距離学習を行うための弱教師あり学習手法を提案した．実験により，提案手法は通常の VAE と比較して，より優れた潜在音高および音色表現を獲得できることを確認した．今後の展望としては，潜在空間に時間的な情報をも取り込み，一つの楽器音を一つの音高表現と一つの音色表現へと埋め込むことが考えられる．また，スペクトログラム再構成の質を上げるため，より良い潜在空間を獲得できる他の距離学習手法を提案手法に融合することも，興味深い方向性であると考えられる．

謝辞 本研究の一部は，JST ACCEL No. JPMJAC1602 および PRESTO No. JPMJPR20C，JSPS 科研費 Nos. 16H01744，19H04137 および 20K21813 の支援を受けた．

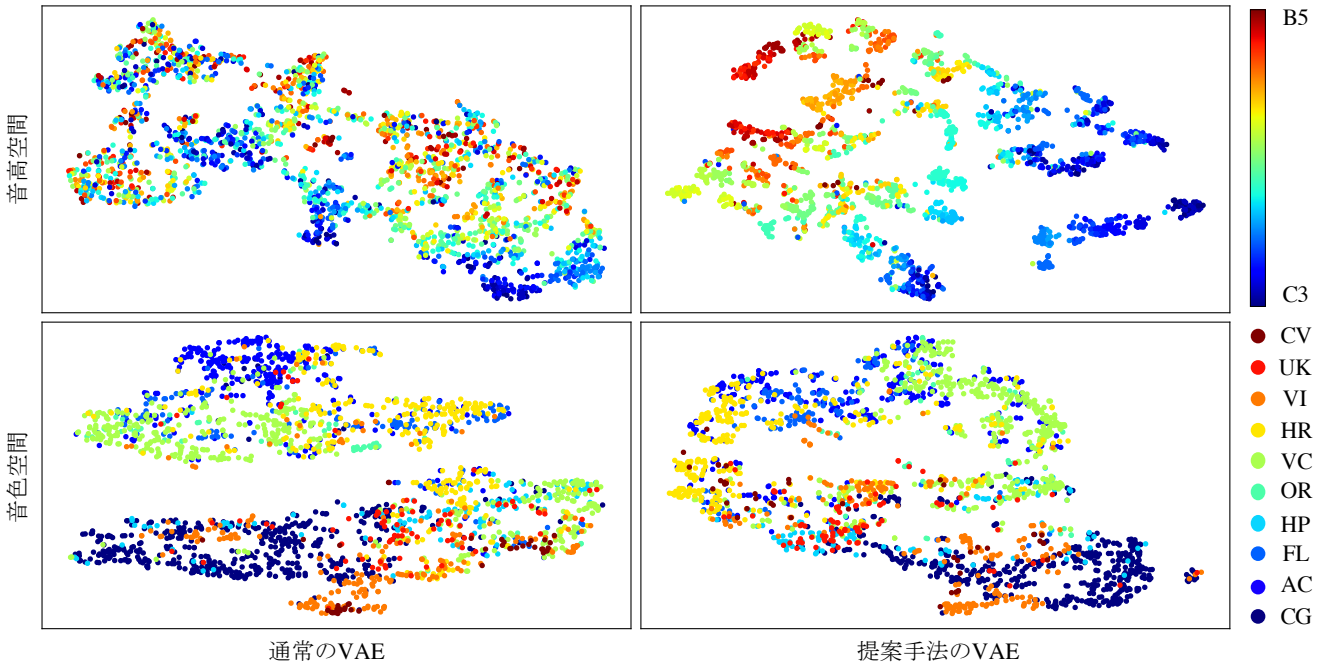

図 3 潜在音高および音色空間の可視化結果

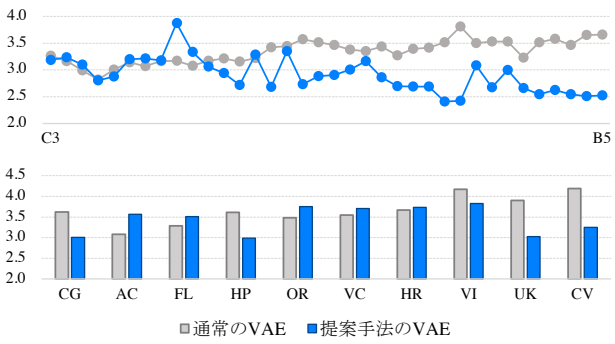

図 4 密集度で見た分離結果の分析（上：音高空間，下：音色空間）

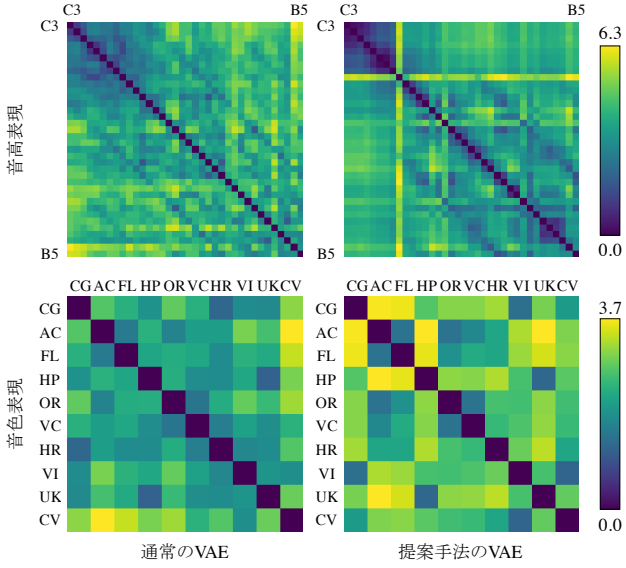

図 5 発散度で見た分離結果の分析

参考文献

[1] Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I. and Abbeel, P.: InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets, *Advances in Neural Information Processing Systems (NIPS)*, pp. 2172–2180 (2016).

[2] Ishfaq, H., Hoogi, A. and Rubin, D.: TVAE: Triplet-Based Variational Autoencoder using Metric Learning, *arXiv:1802.04403*, pp. 1–4 (2018).

[3] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S. and Lerchner, A.: Beta-VAE: Learning basic visual concepts with a constrained variational framework, *International Conference on Learning Representations (ICLR)* (2017).

[4] Kim, H. and Mnih, A.: Disentangling by factorising, *International Conference on Machine Learning (ICML)* (2018).

[5] Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J. and Meent, J.: Structured Disentangled Representations, *Proceedings of Machine Learning Research (PMLR)*, pp. 2525–2534 (2019).

[6] Liu, Z., Luo, P., Wang, X. and Tang, X.: Deep Learning Face Attributes in the Wild, *Proceedings of International Conference on Computer Vision (ICCV)* (2015).

[7] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, pp. 2278–2324 (1998).

[8] Aubry, M., Maturana, D., Efros, A., Russell, B. and Sivic, J.: Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models, *Conference on Computer Vision and Pattern Recognition (CVPR)* (2014).

[9] Ravfogel, S., Elazar, Y., Goldberger, J. and Goldberg, Y.: Unsupervised Distillation of Syntactic Information from Contextualized Word Representations, *arXiv:2010.05265* (2020).

[10] Hsu, W., Zhang, Y. and Glass, J.: Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data, *Advances in Neural Information Processing Systems (NIPS)* (2017).
[11] Dai, S., Zhang, Z. and Xia, G. G.: Music Style Transfer: A Position Paper, *International Workshop on Music Metacreation (MUME)* (2018).
[12] Briot, J., Hadjeres, G. and Pachet, F.: Deep Learning Techniques for Music Generation – A Survey, *Computational Synthesis and Creative Systems*, pp. 1–249 (2020).
[13] Yang, R., Wang, D., Wang, Z., Chen, T., Jiang, J. and Xia, G.: Deep Music Analogy Via Latent Representation Disentanglement, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 596–603 (2019).
[14] Paul, D. and Kundu, S.: A Survey of Music Recommendation Systems with a Proposed Music Recommendation System, *Emerging Technology in Modelling and Graphics*, pp. 279–285 (2020).
[15] Mor, N., Wolf, L., Polyak, A. and Taigman, Y.: A Universal Music Translation Network, *arXiv:1805.07848* (2018).
[16] Bitton, A., Esling, P. and Chemla-Romeu-Santos, A.: Modulated Variational auto-Encoders for many-to-many musical timbre transfer, *arXiv:1810.00222* (2018).
[17] Esling, P., Chemla–Romeu-Santos, A. and Bitton, A.: Bridging audio analysis, perception and synthesis with perceptually-regularized variational timbre spaces, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 175–181 (2018).
[18] Hung, Y., Chiang, I., Chen, Y. and Yang, Y.: Musical Composition Style Transfer via Disentangled Timbre Representations, *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4697–4703 (2019).
[19] Luo, Y., Agres, K. and Herremans, D.: Learning Disentangled Representations of Timbre and Pitch for Musical Instrument Sounds Using Gaussian Mixture Variational Autoencoders, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 746–753 (2019).
[20] Luo, Y., Cheuk, K. W., Nakano, T., Goto, M. and Herremans, D.: Unsupervised Disentanglement of Pitch and Timbre for Isolated Musical Instrument Sounds, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 700–707 (2020).
[21] Lu, R., Wu, K., Duan, Z. and Zhang, C.: DEEP RANKING: TRIPLET MATCHNET FOR MUSIC METRIC LEARNING, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 121–125 (2017).
[22] Royo-Letelier, J., Hennequin, R., Tran, V. and Moussallam, M.: Disambiguating Music Artists at Scale with Audio Metric Learning, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 622–629 (2018).
[23] Karsdorp, F., Kranenburg, P. and Manjavacas, E.: Learning Similarity Metrics for Melody Retrieval, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 478–485 (2019).
[24] McCallum, M. C.: Unsupervised Learning of Deep Features for Music Segmentation, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 346–350 (2019).
[25] Schindler, A. and Knees, P.: Multi-Task Music Representation Learning from Multi-Label Embeddings, *Content-Based Multimedia Indexing (CBMI)*, pp. 1–6 (2019).
[26] Larochelle, H., Erhan, D. and Bengio, Y.: Zero-data Learning of New Tasks, *National Conference on Artificial Intelligence*, pp. 646–651 (2008).
[27] Palatucci, M., Pomerleau, D., Hinton, G. and Mitchell, T. M.: Zero-Shot Learning with Semantic Output Codes, *Advances in Neural Information Processing Systems (NIPS)* (2009).
[28] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes, *International Conference on Learning Representations (ICLR)* (2014).
[29] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Music Genre Database and Musical Instrument Sound Database, *International Society for Music Information Retrieval Conference (ISMIR)*, pp. 229–230 (2003).
[30] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv:1412.6980* (2014).
[31] van der Maaten, L. and Hinton, G.: Visualizing data using t-SNE, *JOURNAL of Machine Learning Research*, pp. 2579–2605 (2008).