Vol.2022-MUS-134 No.62
Vol.2022-SLP-142 No.62
2022/6/18

双曲空間への音色埋め込みを用いたガウス混合変分自己符号化器による楽音合成の検討

中島 風太[1,a)]　中村 友彦[1,b)]　高宗 典玄[1,c)]　深山 覚[2,d)]　猿渡 洋[1,e)]

概要：本稿では，音色の階層性を考慮した変分自己符号化器（variational autoencoder: VAE）による楽音合成法を提案する．VAE に基づく楽音合成では，楽器音をその特徴を捉えた低次元な潜在空間に射影しそこから再構成できるよう学習する．潜在空間で効率的な表現を得るためには，データの性質を反映した潜在空間を構築することが重要である．これに対し，本稿では物理的な機構に基づく楽器分類体系には階層性が存在することに着眼し，音色に関する潜在変数を階層性のあるデータを効率的に表現できる双曲空間上で定義した VAE を提案する．提案法は，音色と音高を別々の潜在空間で扱うことのできる従来法を拡張し，音色に関する潜在変数の事前分布として双曲空間上の正規分布（擬似双曲正規分布）を導入する．また，擬似双曲正規分布の導入を行っても，従来法と同様に確率的勾配降下法を用いて学習できることを示す．実験により，音色に関する潜在空間において，Euclid 空間を用いる場合に比べ双曲空間を用いることで，同一楽器類はより近く，異種楽器類はより遠くへと埋め込まれることが示唆された．

キーワード：楽音合成, 変分自己符号化器, 双曲空間, 音色埋め込み, 表現学習

1. 序論

楽音合成は，パラメトリックな表現を用いて計算機により楽器音を合成する技術である．深層ニューラルネットワーク（deep neural network: DNN）の導入により楽音合成の性能は大きく向上しており，変分自己符号化器（variational autoencoder: VAE）[1] に代表される深層生成モデルがよく用いられる [2].

VAE は，入力データを低次元な潜在空間上の確率分布のパラメータへと変換するエンコーダと，その確率分布からサンプルされた潜在変数からデータを生成するデコーダからなり，入力データを再構成できるよう学習される．学習により得られた潜在空間上での補間やサンプリングにより，異なるデータ間の補間や入力データに類似したデータの生成が可能となる．この性質を利用し，2 つの楽器音の補間や入力データと類似した楽器音の生成を行えるため，様々な VAE ベースの楽音合成手法が提案されてきた [3–6]．例えば，混合正規分布 VAE（Gaussian mixture VAE: GMVAE）[7] を用いた楽音合成手法では，音色と音高に対応する潜在空間をそれぞれ設けることで音色と音高を別々に変更し楽器音を合成できる [3]．また，音色毎，音高毎に異なる正規分布のパラメータを用いることで，楽器間の音色に関する補間も可能である．

VAE はデータを低次元な潜在空間で表現するため，いかにデータの性質を反映した潜在空間を構築するかが肝要である．音色は楽器の物理的な機構によって異なり，類似した機構をもつ楽器は類似した音色をもつ．物理的な機構に基づく楽器分類体系である Hornbostel–Sachs 分類 [8] では，楽器は階層的に分類される．例えば，気鳴楽器という大分類の中には，吹奏楽器，リード付き吹奏楽器，クラリネット類などの，この順に粒度の細かい下位分類が定義されている．この階層性を援用した埋め込み方法を用いることで，音響信号からの楽器識別性能が向上することが示されており [9]，階層性を潜在空間に反映することで音色をより効率的に表現できる可能性がある．

そこで，階層的な構造を潜在空間に導入するため双曲空間に着眼した．双曲空間は一定の負の曲率をもつ非 Euclid 空間であり，Euclid 空間とは異なり原点からの距離が指数的に増大する．階層的なデータが表現できる木構造も階層

[1] 東京大学
the University of Tokyo, Bunkyo, Tokyo 113–8656, Japan
[2] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305–8560, Japan
[a)] nakashima-futa@g.ecc.u-tokyo.ac.jp
[b)] tomohiko-nakamura@g.ecc.u-tokyo.ac.jp
[c)] norihiro_takamune@ipc.i.u-tokyo.ac.jp
[d)] s.fukayama@aist.go.jp
[e)] hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

が深くなるに連れ子ノードの数が指数的に増大する．そのため，原点からの距離が線形にしか増大しない Euclid 空間よりも，双曲空間の方が木構造型のデータを効率的に埋め込める [10]．双曲空間を導入することで，画像，自然言語，ソーシャルグラフなど階層性を持つデータに関するタスクの性能を向上させることが示されており [11]，VAE の潜在空間として双曲空間を用いる手法も提案されている [12,13]．例えば，[12] ではパラメータに関して勾配を計算できるようにしつつ双曲空間上で正規分布を定義する方法を提案し，それを用いて潜在空間として双曲空間を用いた VAE（双曲 VAE）を実現している．

本稿では，GMVAE に基づく楽音合成手法と双曲 VAE を組み合わせることで，音色に対応する潜在空間にのみ双曲空間を用いた VAE ベースの楽音合成手法を提案する．双曲 VAE はデータの全ての要素を一つの潜在空間で表現するため，そのままでは音色に対応する潜在空間のみを双曲空間にできない．これに対し，提案法では GMVAE をベースとすることで，音色と音高を別々の潜在空間で表現する方法を引き継ぎつつ，音色に対応する潜在空間にのみ双曲空間を導入する．また，双曲 VAE で提案された双曲空間上の正規分布を用いることで，提案法は GMVAE と同様に勾配降下法を用いて学習できる．楽音合成実験により，提案法と GMVAE に基づく従来手法 [3] を比較することで双曲空間を導入した効果を確認する．

2. 事前準備：VAE と双曲空間

2.1 VAE [1]

VAE [1] は DNN を用いた確率的生成モデルである．この手法では，対象となるデータの特徴を表す潜在変数 $\boldsymbol{z} \in \mathbb{R}^D$ からデータ $\boldsymbol{X}$ が確率的に生成される過程を考え，それを DNN を用いて表現する．最尤推定の枠組みに則り生成過程を表す DNN のパラメータをデータから決定する問題は，データに関する対数尤度 $\log p(\boldsymbol{X})$ の最大化問題として定式化できるものの，直接 $\log p(\boldsymbol{X})$ を計算することは難しい．そこで，データの生成過程 $p(\boldsymbol{X}|\boldsymbol{z})$ を表す DNN をデコーダとみなし，$\boldsymbol{z}$ に関する事後分布 $p(\boldsymbol{z}|\boldsymbol{X})$ を推定する DNN（エンコーダ）を導入する．ここで，デコーダのパラメータを θ，エンコーダのパラメータを ϕ とする．推定された事後分布 $q_\phi(\boldsymbol{z}|\boldsymbol{X})$ を導入することで，以下のような対数尤度の変分下限（evidence lower bound: ELBO）$\mathcal{L}(\theta, \phi; \boldsymbol{X})$ を導出できる．

$$\begin{aligned}\mathcal{L}(\theta, \phi; \boldsymbol{X}) \;=\; & \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{X})}\left[\log p_\theta(\boldsymbol{X}|\boldsymbol{z})\right] \\ & -\mathcal{D}_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{X})\|p(\boldsymbol{z}))\end{aligned} \tag{1}$$

ここで，$\mathbb{E}$ は期待値，$\mathcal{D}_{\mathrm{KL}}$ は Kullback–Leibler（KL）ダイバージェンスである．式 (1) の右辺第 1 項は，デコーダが生成したデータ $\hat{\boldsymbol{X}}$ が入力 $\boldsymbol{X}$ に近くなることを誘導する．通常，潜在変数の事前分布 $p(\boldsymbol{z})$ は標準正規分布が用いられ，式 (1) の右辺第 2 項により，エンコーダで推定される事後分布がこれに近づくよう誘導される．VAE では $\mathcal{L}(\theta, \phi; \boldsymbol{X})$ を最大化する θ, ϕ をデータから確率的勾配降下法を用いて学習する．

2.2 双曲空間

本節では，双曲空間における演算や確率分布を概説する．詳細な導出は [12,13] を参照されたい．以下，座標が双曲空間上のものであることを，記号 $\widetilde{\cdot}$ で表す．

2.2.1 Lorentz モデル

双曲空間は，空間の曲がり度合いを表す曲率 K が一定かつ負である空間である．一方，VAE の潜在空間として用いられる Euclid 空間は $K = 0$ となる空間であり，$K > 0$ の場合は超球面に対応する．曲率 K の代わりに半径 $R = 1/\sqrt{|K|}$ を用いて空間を表すこともある．

本稿では，[12] と同様に双曲空間のモデルとして Lorentz モデルを用いる．Lorentz モデルを用いることで，2.2.2, 2.2.3 節で述べる双曲空間での基本演算が閉形式で書ける．2 つの $d+1$ 次元実ベクトル $\boldsymbol{a} = [a_1, \ldots, a_{d+1}]^\top, \boldsymbol{b} = [b_1, \ldots, b_{d+1}]^\top \in \mathbb{R}^{d+1}$ に対し，Lorentz 内積を

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_L = -a_1 b_1 + \sum_{i=2}^{d+1} a_i b_i \tag{2}$$

と定義する．d 次元で曲率 K の Lorentz モデル $\mathbb{H}^d_K \subset \mathbb{R}^{d+1}$ は，Lorentz 内積を用いて以下のように定義される．

$$\mathbb{H}^d_K = \left\{ \boldsymbol{a} \in \mathbb{R}^{d+1} | \langle \boldsymbol{a}, \boldsymbol{a} \rangle_L = 1/K,\ a_1 > 0 \right\} \tag{3}$$

ここで，a_1 が最小となる点 $\widetilde{\boldsymbol{0}} = [R, 0, \ldots, 0]^\top$ を原点と呼ぶ．また，双曲空間上の 2 点 $\widetilde{\boldsymbol{a}}, \widetilde{\boldsymbol{b}}$ に関して最短となる曲線（測地線）の距離は，以下のように計算できる．

$$\mathcal{D}_{\mathbb{H}^d_K}(\widetilde{\boldsymbol{a}}, \widetilde{\boldsymbol{b}}) = \frac{1}{\sqrt{-K}} \cosh^{-1}(-K \langle \widetilde{\boldsymbol{a}}, \widetilde{\boldsymbol{b}} \rangle_L) \tag{4}$$

2.2.2 指数写像と対数写像

双曲空間に関する多くの演算は，双曲空間上の各点に関する接空間を介して行われる．双曲空間上の任意の点 $\widetilde{\boldsymbol{a}} \in \mathbb{H}^d_K$ に対し，接空間 $T_{\widetilde{\boldsymbol{a}}}\mathbb{H}^d_K \subset \mathbb{R}^{d+1}$ は d 次元のベクトル空間として以下のように定義される．

$$T_{\widetilde{\boldsymbol{a}}}\mathbb{H}^d_K = \left\{ \boldsymbol{u} \in \mathbb{R}^{d+1} | \langle \boldsymbol{u}, \widetilde{\boldsymbol{a}} \rangle_L = 0 \right\} \tag{5}$$

接空間上の点を双曲空間へ射影する写像を指数写像，双曲空間上の点を接空間に射影する写像を対数写像と呼ぶ [11]．指数写像は双曲空間上の点毎に定まり，当該点に関する接空間上の点を双曲空間に射影する．Lorentz モデルにおける指数写像 $\exp^K_{\widetilde{\boldsymbol{a}}}(\cdot)$ は，以下のように $\boldsymbol{v} \in T_{\widetilde{\boldsymbol{a}}}\mathbb{H}^d_K$ を $\mathbb{H}^d_K$ に射影する．

$$\begin{aligned}\exp^K_{\widetilde{\boldsymbol{a}}}(\boldsymbol{v}) \;=\; & \cosh(\sqrt{-K}\|\boldsymbol{v}\|_L)\widetilde{\boldsymbol{a}} \\ & + \sinh(\sqrt{-K}\|\boldsymbol{v}\|_L)\frac{\boldsymbol{v}}{\sqrt{-K}\|\boldsymbol{v}\|_L}\end{aligned} \tag{6}$$

ここで，$\|\cdot\|_L$ は Lorentz ノルムであり，Lorentz 内積を用いて $\|\boldsymbol{a}\|_L = \sqrt{\langle \boldsymbol{a}, \boldsymbol{a} \rangle_L}$ と計算される．対数写像 $\log_{\widetilde{\boldsymbol{a}}}^K(\cdot)$ は指数写像の逆変換であり，以下のように双曲空間上の点 $\widetilde{\boldsymbol{b}} \in \mathbb{H}_K^d$ を $T_{\widetilde{\boldsymbol{a}}}\mathbb{H}_K^d$ へ射影する．

$$\begin{aligned} &\log_{\widetilde{\boldsymbol{a}}}^K(\widetilde{\boldsymbol{b}}) \\ &= \frac{\cosh(K\langle \widetilde{\boldsymbol{a}}, \widetilde{\boldsymbol{b}} \rangle_L)}{\sinh(\cosh^{-1}(K\langle \widetilde{\boldsymbol{a}}, \widetilde{\boldsymbol{b}} \rangle_L))}(\widetilde{\boldsymbol{b}} - K\langle \widetilde{\boldsymbol{a}}, \widetilde{\boldsymbol{b}} \rangle_L \widetilde{\boldsymbol{a}}) \end{aligned} \tag{7}$$

2.2.3 平行移動

Euclid 空間ではベクトル加法により表現できる平行移動も，双曲空間では接空間を介して定義される．双曲空間上の点 $\widetilde{\boldsymbol{a}}$ から点 $\widetilde{\boldsymbol{b}}$ への平行移動 $\mathrm{PT}_{\widetilde{\boldsymbol{a}} \to \widetilde{\boldsymbol{b}}}^K(\cdot)$ は，$\boldsymbol{v} \in T_{\widetilde{\boldsymbol{a}}}\mathbb{H}_K^d$ を測地線に沿って $\widetilde{\boldsymbol{b}}$ に関する接空間 $T_{\widetilde{\boldsymbol{b}}}\mathbb{H}_K^d$ に移す写像である．ただし，$\mathrm{PT}_{\widetilde{\boldsymbol{a}} \to \widetilde{\boldsymbol{b}}}^K(\cdot)$ は任意の $\boldsymbol{v}, \boldsymbol{v'} \in T_{\widetilde{\boldsymbol{a}}}\mathbb{H}_K^d$ に関して Lorentz 内積を以下のように保存する．

$$\langle \mathrm{PT}_{\widetilde{\boldsymbol{a}} \to \widetilde{\boldsymbol{b}}}^K(\boldsymbol{v}), \mathrm{PT}_{\widetilde{\boldsymbol{a}} \to \widetilde{\boldsymbol{b}}}^K(\boldsymbol{v'}) \rangle_L = \langle \boldsymbol{v}, \boldsymbol{v'} \rangle_L \tag{8}$$

この写像も解析的に書け，以下のように計算できる．

$$\mathrm{PT}_{\widetilde{\boldsymbol{a}} \to \widetilde{\boldsymbol{b}}}^K(\boldsymbol{v}) = \boldsymbol{v} - \frac{K\langle \widetilde{\boldsymbol{b}}, \boldsymbol{v} \rangle_L}{1 + K\langle \widetilde{\boldsymbol{a}}, \widetilde{\boldsymbol{b}} \rangle_L}(\widetilde{\boldsymbol{a}} + \widetilde{\boldsymbol{b}}) \tag{9}$$

平行移動の逆写像は，移動の始点と終点を入れ替えた写像 $\mathrm{PT}_{\widetilde{\boldsymbol{b}} \to \widetilde{\boldsymbol{a}}}^K(\cdot)$ により定義できる．

以下では，原点 $\widetilde{\mathbf{0}}$ における接空間から $\widetilde{\boldsymbol{a}}$ に関する接空間 $T_{\widetilde{\boldsymbol{a}}}\mathbb{H}_K^d$ への平行移動と接空間 $T_{\widetilde{\boldsymbol{a}}}\mathbb{H}_K^d$ での指数写像の合成写像を $\mathrm{proj}_{\widetilde{\boldsymbol{a}}}^K(\cdot) = \exp_{\widetilde{\boldsymbol{a}}}^K \circ \mathrm{PT}_{\widetilde{\mathbf{0}} \to \widetilde{\boldsymbol{a}}}^K(\cdot)$ で表す．

2.2.4 擬似双曲正規分布 [12]

本節では，[12] で提案された双曲空間での正規分布（擬似双曲正規分布）について述べる．擬似双曲正規分布 $\mathcal{G}(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ は，Euclid 空間での正規分布と同じく平均 $\widetilde{\boldsymbol{\mu}} \in \mathbb{H}_K^d$ と分散共分散行列 $\boldsymbol{\Sigma}$ をパラメータにもつ．$\mathcal{G}(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma})$ に従う確率変数 $\widetilde{\boldsymbol{z}}$ は，Euclid 空間上で正規分布 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ に従う確率変数 $\boldsymbol{w} \in \mathbb{R}^d$ に対し，以下の変換を施すことで得られる．

$$\widetilde{\boldsymbol{z}} = \mathrm{proj}_{\widetilde{\boldsymbol{\mu}}}^K([0, \boldsymbol{w}^\top]^\top) \tag{10}$$

この分布の対数確率密度関数は，変数変換に関する連鎖律を用いて以下のように計算できる [12].

$$\begin{aligned} &\log p(\widetilde{\boldsymbol{z}}; \widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}) \\ &= \log p(\boldsymbol{w}; \mathbf{0}, \boldsymbol{\Sigma}) - \log \left| \det \frac{\mathrm{d}\widetilde{\boldsymbol{z}}}{\mathrm{d}\boldsymbol{w}} \right| \qquad (11) \\ &= \log p(\boldsymbol{w}; \mathbf{0}, \boldsymbol{\Sigma}) - (d-1) \log \left(\frac{R \sinh(\|\boldsymbol{u}\|_L / R)}{\|\boldsymbol{u}\|_L} \right) \qquad (12) \end{aligned}$$

ここで，$\boldsymbol{u} = \log_{\widetilde{\boldsymbol{\mu}}}^K(\widetilde{\boldsymbol{z}})$ とした．

擬似双曲正規分布の利点は，解析的に確率密度関数が表せ，サンプリングも容易であることである．この性質を利用することで，潜在空間を $\mathbb{H}_K^d$ とした VAE である双曲 VAE が構築できる [12]．また，$\widetilde{\boldsymbol{\mu}}, \boldsymbol{\Sigma}$ に関する勾配も計算できるため，通常の VAE と同じく勾配降下法により双曲 VAE も学習できる．

3. 関連研究

VAE に基づく楽音合成手法では，音楽を構成する重要な要素である音色，音高をいかに個別に操作できるような潜在変数を獲得できるかに焦点が当てられてきた [3–5]．Luo ら [3] は，音色と音高に対応する潜在空間を別々に用意し，入力データに付随した楽器と音高のラベルを用いてそれらを別々に操作できる VAE ベースの楽音合成を提案した．この手法では，潜在変数 $\boldsymbol{z} \in \mathbb{R}^D$ を，音色を表す $D^{(\mathrm{t})}$ 次元の潜在変数 $\boldsymbol{z}^{(\mathrm{t})}$ と音高を表す $D^{(\mathrm{p})}$ 次元の潜在変数 $\boldsymbol{z}^{(\mathrm{p})}$ に分解し，それらの独立性を仮定する．すなわち，$p(\boldsymbol{z}) = p(\boldsymbol{z}^{(\mathrm{t})})p(\boldsymbol{z}^{(\mathrm{p})})$ とする．ここで，$D = D^{(\mathrm{t})} + D^{(\mathrm{p})}$ である．事後分布も音色と音高に関して独立と仮定し，それぞれに関して別のエンコーダを用いる．一方，デコーダは VAE と同じく，音色と音高に関する潜在変数を結合した D 次元の潜在変数 $\boldsymbol{z}$ からのデータの生成過程を表現する．

さらに，Luo らの手法では GMVAE [7] を用いることで，楽器の種類と音高を反映した潜在変数の事前分布を設計している．GMVAE では，潜在変数 $\boldsymbol{z}$ は入力データに付随するラベル $\boldsymbol{y}$ に条件付けられる．Luo らの手法では，音高の潜在変数の事前分布を音高のラベル $y^{(\mathrm{p})}$ に依存させ $p(\boldsymbol{z}^{(\mathrm{p})}|y^{(\mathrm{p})})$ とすることで，音高毎に異なる事前分布を用いる．同様に，楽器のラベル $y^{(\mathrm{t})}$ に音色に関する事前分布を依存させ $p(\boldsymbol{z}^{(\mathrm{t})}|y^{(\mathrm{t})})$ とし，楽器毎に異なる事前分布を用いる．$p(\boldsymbol{z}^{(\mathrm{p})}|y^{(\mathrm{p})}), p(\boldsymbol{z}^{(\mathrm{t})}|y^{(\mathrm{t})})$ を正規分布とすることで $p(\boldsymbol{z})$ は混合正規分布となり，潜在空間で多峰性の確率分布を表現できるよう拡張している．この手法を足がかりに，ラベルが付与されていないデータにも利用できる拡張 [4] や，GMVAE の代わりに距離学習を用いて学習データにない楽器への汎化性を向上させる手法も提案されている [5].

4. 提案手法

本節では，[3] で提案された楽音合成モデルをベースに，音色に対応する潜在空間を双曲空間に拡張した楽音合成手法を提案する．3 節で示したように，[3] で提案された手法は潜在変数をそれぞれ音色と音高に関連する次元に分解して表現する．提案法では，音色に関する潜在変数 $\boldsymbol{z}^{(\mathrm{t})}$ を双曲空間 $\mathbb{H}_K^{D^{(\mathrm{t})}}$ で定義し，事前分布として擬似双曲正規分布を導入する．これにより，音色に関して階層的な表現が学習できるように誘導する．

4.1 生成モデル

本節では，提案法の生成モデルに関して述べる．以下では，簡単のため [3] と同一の確率変数に関しては 3 節で用いた記法を利用する．

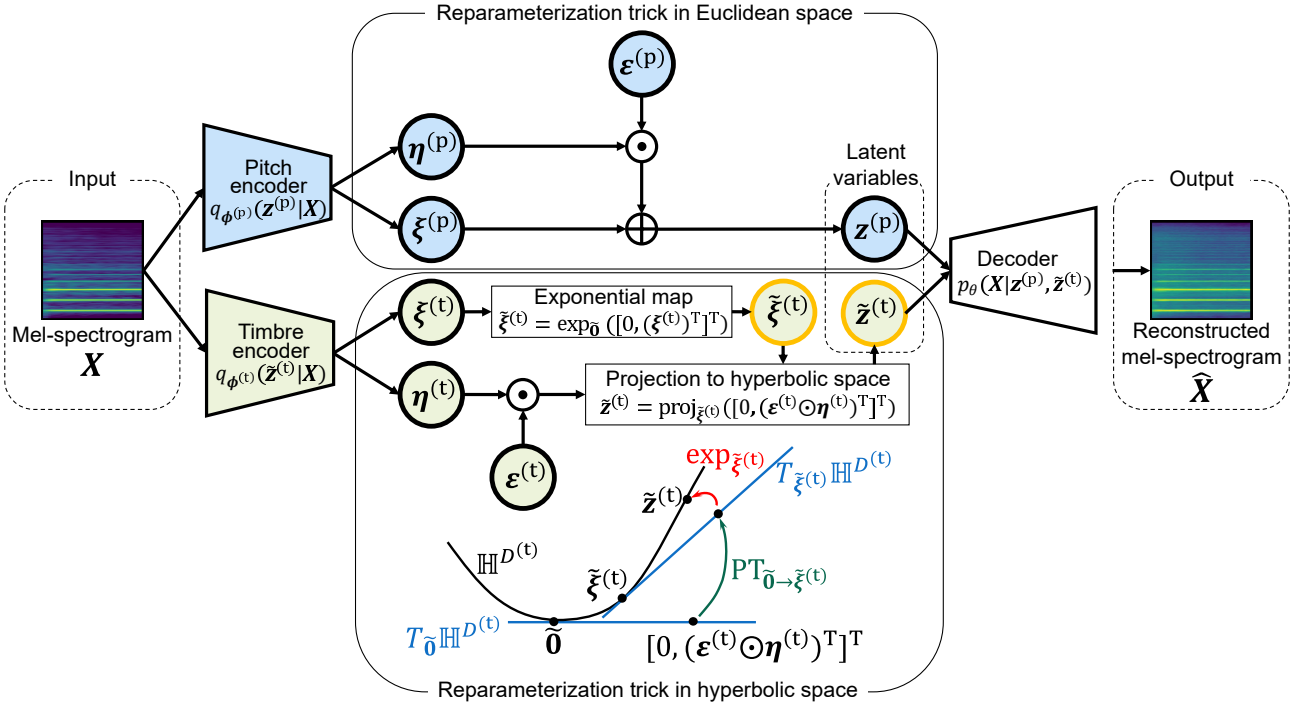

図 1 提案モデルの概略図．2 つのエンコーダが音色と音高を分離し，音色の潜在変数のみ双曲空間上で表現する．記号 $\odot$ はベクトルの要素毎の積を表す．曲率 K は省略して表記した．

提案モデルは，観測データに関する音高ラベル $y^{(\mathrm{p})} \in \mathcal{P}$ が与えられたもとでの，観測音響信号のメルスペクトログラム $\boldsymbol{X}$ の生成過程を記述する．ここで，$\mathcal{P}$ はとりうる音高の集合である．音色ラベル $y^{(\mathrm{t})} \in \mathcal{T}$ は一様な確率をもつカテゴリカル分布に従う．これはデータに含まれる楽器の偏りがないことを意味する．ここで，$\mathcal{T}$ はとりうる音色ラベルの集合である．音高に対応する潜在変数 $\boldsymbol{z}^{(\mathrm{p})}$ と音色に対応する潜在変数 $\tilde{\boldsymbol{z}}^{(\mathrm{t})}$ は，それぞれ Euclid 空間 $\mathbb{R}^{D^{(\mathrm{p})}}$，Lorentz モデル $\mathbb{H}_K^{D^{(\mathrm{t})}}$ 上で定義される．$\boldsymbol{z}^{(\mathrm{p})}$ は，$y^{(\mathrm{p})}$ によって異なる平均と分散共分散行列をもつ正規分布に従う．

$$\boldsymbol{z}^{(\mathrm{p})} \sim \mathcal{N}(\boldsymbol{\mu}^{(\mathrm{p})}_{y^{(\mathrm{p})}}, (\mathrm{diag}(\boldsymbol{\sigma}^{(\mathrm{p})}_{y^{(\mathrm{p})}}))^2) \tag{13}$$

ここで，$\mathrm{diag}(\cdot)$ は与えられたベクトルを対角要素とする対角行列を表し，$\boldsymbol{\mu}^{(\mathrm{p})}_{y^{(\mathrm{p})}} \in \mathbb{R}^{D^{(\mathrm{p})}}$ は平均，$\boldsymbol{\sigma}^{(\mathrm{p})}_{y^{(\mathrm{p})}}$ は各要素の標準偏差を並べた非負ベクトルを表す．$\tilde{\boldsymbol{z}}^{(\mathrm{t})}$ は，$y^{(\mathrm{t})}$ によって異なる平均と分散共分散行列を持つ擬似双曲正規分布に従う．

$$\tilde{\boldsymbol{z}}^{(\mathrm{t})} | y^{(\mathrm{t})} \sim \mathcal{G}(\tilde{\boldsymbol{\mu}}^{(\mathrm{t})}_{y^{(\mathrm{t})}}, (\mathrm{diag}(\boldsymbol{\sigma}^{(\mathrm{t})}_{y^{(\mathrm{t})}}))^2) \tag{14}$$

ここで，$\tilde{\boldsymbol{\mu}}^{(\mathrm{t})}_{y^{(\mathrm{t})}} \in \mathbb{H}_K^{D^{(\mathrm{t})}}$ は平均，$\boldsymbol{\sigma}^{(\mathrm{t})}_{y^{(\mathrm{t})}}$ は各要素の標準偏差を並べた非負ベクトルを表す．本稿では，[3] に倣い $\boldsymbol{\sigma}^{(\mathrm{t})}_{y^{(\mathrm{t})}}, \boldsymbol{\sigma}^{(\mathrm{p})}_{y^{(\mathrm{p})}}$ は固定値とした．デコーダは，これらの潜在変数を結合して得られる潜在変数 $[(\boldsymbol{z}^{(\mathrm{p})})^\top, (\tilde{\boldsymbol{z}}^{(\mathrm{t})})^\top]^\top$ から $\boldsymbol{X}$ の生成モデル $p(\boldsymbol{X}|\boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})})$ を表現する．これらをまとめると，$y^{(\mathrm{t})}, \boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})}, \boldsymbol{X}$ の同時分布は以下のように書ける．

$$\begin{aligned} &p(\boldsymbol{X}, \boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})}, y^{(\mathrm{t})}; y^{(\mathrm{p})}) \\ &= p(\boldsymbol{X}|\boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})}) p(\boldsymbol{z}^{(\mathrm{p})}; y^{(\mathrm{p})}) p(\tilde{\boldsymbol{z}}^{(\mathrm{t})}|y^{(\mathrm{t})}) p(y^{(\mathrm{t})}) \end{aligned} \tag{15}$$

この同時分布を $\boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})}, y^{(\mathrm{t})}$ に関して周辺化することで，対数周辺分布 $\log p(\boldsymbol{X}; y^{(\mathrm{p})})$ が得られる．

4.2 学習

4.2.1 ELBO の導出

文献 [3] と同様に，潜在変数 $\boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})}$ の事後分布を推定するエンコーダをそれぞれ導入し，対数尤度 $\log p(\boldsymbol{X}; y^{(\mathrm{p})})$ の下限である ELBO を導出する．エンコーダは $\boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})}$ 毎に用意され，それぞれパラメータ $\phi^{(\mathrm{p})}, \phi^{(\mathrm{t})}$ をもつ DNN である．これらのエンコーダによって得られる変分事後分布を，それぞれ $q_{\phi^{(\mathrm{p})}}(\boldsymbol{z}^{(\mathrm{p})}|\boldsymbol{X}), q_{\phi^{(\mathrm{t})}}(\tilde{\boldsymbol{z}}^{(\mathrm{t})}|\boldsymbol{X})$ で表す．さらに，音色ラベルの事後分布 $p(y^{(\mathrm{t})}|\boldsymbol{X})$ を近似する変分事後分布を $q(y^{(\mathrm{t})}|\boldsymbol{X})$ とすると，変分下限 $\mathcal{L}(\theta, \phi^{(\mathrm{p})}, \phi^{(\mathrm{t})}; \boldsymbol{X}, y^{(\mathrm{p})})$ は

$$\begin{aligned} &\mathcal{L}(\theta, \phi^{(\mathrm{p})}, \phi^{(\mathrm{t})}; \boldsymbol{X}, y^{(\mathrm{p})}) \\ &= \mathbb{E}_{\substack{\tilde{\boldsymbol{z}}^{(\mathrm{t})} \sim q_{\phi^{(\mathrm{t})}}(\tilde{\boldsymbol{z}}^{(\mathrm{t})}|\boldsymbol{X}) \\ \boldsymbol{z}^{(\mathrm{p})} \sim q_{\phi^{(\mathrm{p})}}(\boldsymbol{z}^{(\mathrm{p})}|\boldsymbol{X})}} \left[\log p_\theta(\boldsymbol{X}|\boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})})\right] \\ &\quad - \mathcal{D}_{\mathrm{KL}}(q_{\phi^{(\mathrm{p})}}(\boldsymbol{z}^{(\mathrm{p})}|\boldsymbol{X}) \| p(\boldsymbol{z}^{(\mathrm{p})}|y^{(\mathrm{p})})) \\ &\quad - \mathbb{E}_{y^{(\mathrm{t})} \sim q(y^{(\mathrm{t})}|\boldsymbol{X})} \left[\mathcal{D}_{\mathrm{KL}}(q_{\phi^{(\mathrm{t})}}(\tilde{\boldsymbol{z}}^{(\mathrm{t})}|\boldsymbol{X}) \| p(\tilde{\boldsymbol{z}}^{(\mathrm{t})}|y^{(\mathrm{t})}))\right] \\ &\quad - \mathcal{D}_{\mathrm{KL}}(q_{\phi^{(\mathrm{t})}}(y^{(\mathrm{t})}|\boldsymbol{X}) \| p(y^{(\mathrm{t})})) \end{aligned} \tag{16}$$

と書ける．詳細な導出は [14] を参照されたい．

4.2.2 Reparameterization Trick

提案法の学習問題は，全学習データに対する ELBO の和を最大化する DNN のパラメータ $\theta, \phi^{(\mathrm{p})}, \phi^{(\mathrm{t})}$ と事前分布の平均 $\widetilde{\boldsymbol{\mu}}^{(\mathrm{t})}_{y^{(\mathrm{t})}}, \boldsymbol{\mu}^{(\mathrm{p})}_{y^{(\mathrm{p})}}$ を求める問題に帰着できる．式 (16) 中の $\boldsymbol{z}^{(\mathrm{p})}, \widetilde{\boldsymbol{z}}^{(\mathrm{t})}$ に関する期待値を解析的に求めることは難しいため，[1] で提案された reparameterization trick と呼ばれる近似計算法を用いる．以下では，$\boldsymbol{z}^{(\mathrm{p})}, \widetilde{\boldsymbol{z}}^{(\mathrm{t})}$ に関する reparameterization trick について述べる．

図 1 は提案法の概略図である．$\boldsymbol{z}^{(\mathrm{p})}$ は Euclid 空間にあるため，VAE や GMVAE と同様の reparameterization trick を適用できる．音高用エンコーダは，$\boldsymbol{X}$ から $\boldsymbol{z}^{(\mathrm{p})}$ の事後分布の平均 $\boldsymbol{\xi}^{(\mathrm{p})} \in \mathbb{R}^{D^{(\mathrm{p})}}$ と，対角分散の対数 $\boldsymbol{\zeta}^{(\mathrm{p})} \in \mathbb{R}^{D^{(\mathrm{p})}}$ を推定する．この後，$\boldsymbol{\zeta}^{(\mathrm{p})}$ の各要素に関して指数関数を適用したものを対角要素に並べた対角行列を $(\mathrm{diag}(\boldsymbol{\eta}^{(\mathrm{p})}))^2$ とする．これらを用いて，エンコーダにより推定された $\boldsymbol{z}^{(\mathrm{p})}$ の事後分布は $\mathcal{N}(\boldsymbol{\xi}^{(\mathrm{p})}, (\mathrm{diag}(\boldsymbol{\eta}^{(\mathrm{p})}))^2)$ と書ける．$\boldsymbol{z}^{(\mathrm{p})}$ を事後分布から直接サンプルしモンテカルロ近似すると，エンコーダのパラメータ $\phi^{(\mathrm{p})}$ に関する勾配が計算できない．そこで，標準正規分布 $\mathcal{N}(\boldsymbol{0}_{D^{(\mathrm{p})}}, \boldsymbol{I}_{D^{(\mathrm{p})}})$ からサンプルした $\boldsymbol{\varepsilon}^{(\mathrm{p})}$ を用いて，$\boldsymbol{\xi}^{(\mathrm{p})} + \boldsymbol{\varepsilon}^{(\mathrm{p})} \odot \boldsymbol{\eta}^{(\mathrm{p})}$ と計算することで，事後分布の平均と分散に従う $\boldsymbol{z}^{(\mathrm{p})}$ をサンプルする．ここで，$\boldsymbol{0}_{D^{(\mathrm{p})}}, \boldsymbol{I}_{D^{(\mathrm{p})}}$ はそれぞれ $D^{(\mathrm{p})}$ 次元の零ベクトルと $D^{(\mathrm{p})}$ 行 $D^{(\mathrm{p})}$ 列の単位行列を表し，$\odot$ は要素毎の積を表す．

$\widetilde{\boldsymbol{z}}^{(\mathrm{t})}$ は Lorentz モデル上にあるため，擬似双曲正規分布に対する reparameterization trick [12] を用いる．音色用エンコーダは，$\boldsymbol{\xi}^{(\mathrm{t})} \in \mathbb{R}^{D^{(\mathrm{t})}}$ と $\boldsymbol{\zeta}^{(\mathrm{t})} \in \mathbb{R}^{D^{(\mathrm{t})}}$ を推定する．$\boldsymbol{\xi}^{(\mathrm{t})}$ を双曲空間に射影して得られる

$$\widetilde{\boldsymbol{\xi}}^{(\mathrm{t})} = \exp^K_{\boldsymbol{0}}([0, (\boldsymbol{\xi}^{(\mathrm{t})})^\top]^\top) \in \mathbb{H}^{D^{(\mathrm{t})}}_K \tag{17}$$

を事後分布の平均とする．また，$\boldsymbol{\zeta}^{(\mathrm{t})}$ の各要素に対し指数関数を適用したものを対角要素に並べた対角行列 $(\mathrm{diag}(\boldsymbol{\eta}^{(\mathrm{t})}))^2$ を事後分布の分散共分散行列とする．これらを用いて，エンコーダを用いて推定された音色の事後分布は $\mathcal{G}(\widetilde{\boldsymbol{\xi}}^{(\mathrm{t})}, (\mathrm{diag}(\boldsymbol{\eta}^{(\mathrm{t})}))^2)$ と書ける．$\widetilde{\boldsymbol{z}}^{(\mathrm{t})}$ のサンプリングの手順は以下の通りである．まず Euclid 空間の標準正規分布 $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{D^{(\mathrm{t})}})$ に従うベクトル $\boldsymbol{\varepsilon}^{(\mathrm{t})} \in \mathbb{R}^{D^{(\mathrm{t})}}$ をとる．それに標準偏差を乗じた $\boldsymbol{v}^{(\mathrm{t})} = \boldsymbol{\varepsilon}^{(\mathrm{t})} \odot \boldsymbol{\eta}^{(\mathrm{t})} \in \mathbb{R}^{D^{(\mathrm{t})}}$ が，分布内での偏差を表す．これを双曲空間上の分布の平均 $\widetilde{\boldsymbol{\xi}}^{(\mathrm{t})} \in \mathbb{H}^{D^{(\mathrm{t})}}_K$ の周りに射影し，$\mathrm{proj}^K_{\widetilde{\boldsymbol{\xi}}^{(\mathrm{t})}}([0, (\boldsymbol{v}^{(\mathrm{t})})^\top]^\top) \in \mathbb{H}^{D^{(\mathrm{t})}}_K$ として $\widetilde{\boldsymbol{z}}^{(\mathrm{t})}$ がサンプルできる．

5. 実験的評価

5.1 実験条件

本節では，双曲空間の導入よる音色埋め込みの性能向上を評価するため行った実験について述べる．

5.1.1 データセット

実験では，先行研究 [3] と同じ Studio-On-Line [15] データセットを用いた．これには 12 種類の音色（イングリッシュホルン，フレンチホルン，テナートロンボーン，トランペット，ピアノ，バイオリン，チェロ，サキソフォン，ファゴット，クラリネット，フルート，オーボエ），82 種類の音高 (A0 ∼ F♯7) の単音 1885 個が含まれる．このうち 80%にあたる 1531 個を学習セット，10%にあたる 177 個を検証セット，10%にあたる 177 個をテストセットとして用いた．3 つのセットへの分割は，各セットに含まれる音色の比率が等しくなるように行った．

5.1.2 比較手法

比較のためのベースラインとして，音色と音高の両方の潜在空間を Euclid 空間としたモデルを使用した．これは従来手法 [3] に相当する．提案手法として，4 節で述べたように音色の潜在空間のみを双曲空間に拡張したモデルを用いた．双曲空間の中での比較も行うため，$R = 1, 2, 5, 10$ の 4 種類の半径を持つ Lorentz モデルを用いた．なお．これらの R の値は，曲率に換算するとそれぞれ $K = -1, -1/4, -1/25, -1/100$ に相当する． 音色表現への双曲空間の導入による効果を様々な条件で確認するため，音色の潜在次元数として $D^{(\mathrm{t})} = 2, 4, 8, 16$ の 4 種類を用いた．一方，音高の潜在空間は比較手法と同様であるため，常に $D^{(\mathrm{p})} = 16$ 次元の Euclid 空間とすることで，比較手法 [3] の条件に統一した．

5.1.3 ネットワーク構造

本研究で使用するモデルは，2 つのエンコーダと 1 つのデコーダを持つ．エンコーダの構造は 2 つとも共通であり，2 つの 1 次元畳み込み層と，その後に続く 1 つの全結合層からなる．全ての層の後には，レイヤー正規化と正規化線形ユニット (rectified linear unit: ReLU) による活性化が行われる．全結合層は潜在変数の事後分布を表すパラメータを出力し，その分布から reparameterization trick で潜在変数がサンプリングされる．デコーダは，2 つの全結合層と，それに続く 2 つの 1 次元逆畳み込み層からなる．最後の逆畳み込み層の後には双曲線正接関数 (tanh) による活性化が行われる．それ以外の 3 つの層の後には，レイヤー正規化と ReLU による活性化が行われる．1 次元畳み込み層では，入力スペクトログラム $\boldsymbol{X}$ の周波数ビン数をチャンネル数と解釈し，時間フレーム数を次元数にもつベクトルに対し畳み込みを行う．畳み込みのカーネルサイズは 3，カーネル数は 512 とする．また，全結合層のユニット数も 512 とする．

このネットワーク構造は，正規化の部分を除いて先行研究 [3] と同一である．先行研究でバッチ正規化を用いていた部分を本研究ではレイヤー正規化に変更し，バッチサイズへの依存性を回避している．

5.2 ハイパーパラメータ

本節の内容は，すべて先行研究 [3] と同一の設定である．音色・音高の条件付き事前分布 $p(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}|y^{(\mathrm{t})}), p(\boldsymbol{z}^{(\mathrm{p})}; y^{(\mathrm{p})})$ の

分散に用いる固定値は，すべての $y^{(\mathrm{t})} \in \mathcal{T}, y^{(\mathrm{p})} \in \mathcal{P}$ に対し，それぞれ $\boldsymbol{\sigma}^{(\mathrm{t})}_{y^{(\mathrm{t})}} = \mathbf{1}_{D^{(\mathrm{t})}}, \boldsymbol{\sigma}^{(\mathrm{p})}_{y^{(\mathrm{p})}} = e^{-2}\mathbf{1}_{D^{(\mathrm{p})}}$ とした．ここで，$\mathbf{1}_d$ は要素が全部 1 の d 次元ベクトルである．潜在変数の事前分布として用いる混合正規分布の混合数は，データセットに含まれるカテゴリの種類に合わせ，音色は 12 個，音高は 82 個とした．学習時のバッチサイズは 128 とした．モデルの初期化には Xavier の初期化を用い，学習率 10^{-4} の Adam でパラメータの最適化を行った．

5.3 評価指標

本節では，実験結果を定量評価するために用いた 2 つの指標，尤度と楽器間距離比について述べる．

5.3.1 尤度

VAE に関する文献 [12,13] に倣い，尤度を評価指標の 1 つとして用いた．データ $\boldsymbol{X}$ と音高の正解ラベル $y^{(\mathrm{p})}$ に関するモデルの尤度 $p(\boldsymbol{X}; y^{(\mathrm{p})})$ を，以下の式変形を用いて算出した．

$$
\begin{aligned}
&p(\boldsymbol{X}; y^{(\mathrm{p})}) \\
&= \int_{\boldsymbol{z}^{(\mathrm{p})}} \int_{\widetilde{\boldsymbol{z}}^{(\mathrm{t})}} p_\theta(\boldsymbol{X}|\boldsymbol{z}^{(\mathrm{p})}, \widetilde{\boldsymbol{z}}^{(\mathrm{t})}) \\
&\qquad \times p(\boldsymbol{z}^{(\mathrm{p})}; y^{(\mathrm{p})}) p(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}) \mathrm{d}\widetilde{\boldsymbol{z}}^{(\mathrm{t})} \mathrm{d}\boldsymbol{z}^{(\mathrm{p})} \qquad (18) \\
&= \int_{\boldsymbol{z}^{(\mathrm{p})}} \int_{\widetilde{\boldsymbol{z}}^{(\mathrm{t})}} p_\theta(\boldsymbol{X}|\boldsymbol{z}^{(\mathrm{p})}, \widetilde{\boldsymbol{z}}^{(\mathrm{t})}) \\
&\qquad \times \frac{p(\boldsymbol{z}^{(\mathrm{p})}; y^{(\mathrm{p})})}{q_{\phi^{(\mathrm{p})}}(\boldsymbol{z}^{(\mathrm{p})}|\boldsymbol{X})} \frac{p(\widetilde{\boldsymbol{z}}^{(\mathrm{t})})}{q_{\phi^{(\mathrm{t})}}(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}|\boldsymbol{X})} \\
&\qquad \times q_{\phi^{(\mathrm{p})}}(\boldsymbol{z}^{(\mathrm{p})}|\boldsymbol{X}) q_{\phi^{(\mathrm{t})}}(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}|\boldsymbol{X}) \mathrm{d}\widetilde{\boldsymbol{z}}^{(\mathrm{t})} \mathrm{d}\boldsymbol{z}^{(\mathrm{p})} \qquad (19) \\
&= \mathbb{E}_{\substack{\widetilde{\boldsymbol{z}}^{(\mathrm{t})} \sim q_{\phi^{(\mathrm{t})}}(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}|\boldsymbol{X}) \\ \boldsymbol{z}^{(\mathrm{p})} \sim q_{\phi^{(\mathrm{p})}}(\boldsymbol{z}^{(\mathrm{p})}|\boldsymbol{X})}} \Bigg[p_\theta(\boldsymbol{X}|\boldsymbol{z}^{(\mathrm{p})}, \widetilde{\boldsymbol{z}}^{(\mathrm{t})}) \\
&\qquad \times \frac{p(\boldsymbol{z}^{(\mathrm{p})}; y^{(\mathrm{p})})}{q_{\phi^{(\mathrm{p})}}(\boldsymbol{z}^{(\mathrm{p})}|\boldsymbol{X})} \frac{p(\widetilde{\boldsymbol{z}}^{(\mathrm{t})})}{q_{\phi^{(\mathrm{t})}}(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}|\boldsymbol{X})} \Bigg] \qquad (20) \\
&\approx \frac{1}{M} \sum_{m=1}^{M} \Bigg(p_\theta(\boldsymbol{X}|\boldsymbol{z}^{(\mathrm{p})}_m, \widetilde{\boldsymbol{z}}^{(\mathrm{t})}_m) \\
&\qquad \times \frac{p(\boldsymbol{z}^{(\mathrm{p})}_m; y^{(\mathrm{p})})}{q_{\phi^{(\mathrm{p})}}(\boldsymbol{z}^{(\mathrm{p})}_m|\boldsymbol{X})} \frac{p(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}_m)}{q_{\phi^{(\mathrm{t})}}(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}_m|\boldsymbol{X})} \Bigg) \qquad (21)
\end{aligned}
$$

ここで，式 (21) では重点サンプリング [16] による近似を用いている．m は各サンプルに対応するインデクスであり，サンプリング点数 M は 500 とした．$\boldsymbol{z}^{(\mathrm{p})}_m$ は $q(\boldsymbol{z}^{(\mathrm{p})}_m|\boldsymbol{X})$ に，$\widetilde{\boldsymbol{z}}^{(\mathrm{t})}_m$ は $q(\widetilde{\boldsymbol{z}}^{(\mathrm{t})}_m|\boldsymbol{X})$ に従ってサンプリングする．

5.3.2 楽器間距離比

本稿では，類似した楽器が潜在空間上で近い位置にあるか否かを以下のように評価した．学習済みモデルにおける $\widetilde{\boldsymbol{\mu}}^{(\mathrm{t})}_j$ は，音色の潜在空間上に埋め込まれた楽器 $j \in \mathcal{T}$ の分布の平均である．これを用いて，異なる 2 つの楽器 $j, j' \in \mathcal{T}$ の潜在空間上での距離 $\mathcal{D}_{\mathbb{H}^{D^{(\mathrm{t})}}_K}(\widetilde{\boldsymbol{\mu}}^{(\mathrm{t})}_j, \widetilde{\boldsymbol{\mu}}^{(\mathrm{t})}_{j'})$ が式 (4) から定まる．これを楽器間距離と呼ぶ．

データセットに含まれる楽器を以下の 4 つの楽器類に分類し，同一楽器類と異種楽器類の楽器間距離を用いて楽器間距離比を定める．

- 木管楽器（6 種類：イングリッシュホルン，サキソフォン，ファゴット，クラリネット，フルート，オーボエ）
- 金管楽器（3 種類：フレンチホルン，テナートロンボーン，トランペット）
- 擦弦楽器（2 種類：バイオリン，チェロ）
- 打弦楽器（1 種類：ピアノ）

データセットに含まれる楽器が 12 種類であるため，異なる 2 つの楽器のペアは 66 通り存在し，そのうち 19 ペアは同一楽器類に属する楽器同士，残りの 47 ペアは異種楽器類に属する楽器同士のペアである．よって，同一楽器類の全ペアに関する楽器間距離の平均は

$$
\mathcal{D}^{(\mathrm{same})} = \frac{1}{19} \sum_{(j,j') \in \mathcal{C}^{(\mathrm{same})}} \mathcal{D}_{\mathbb{H}^{D^{(\mathrm{t})}}_K}(\tilde{\boldsymbol{\mu}}_j, \tilde{\boldsymbol{\mu}}_{j'}) \qquad (22)
$$

異種楽器類の全ペアに関する楽器間距離の平均は

$$
\mathcal{D}^{(\mathrm{different})} = \frac{1}{47} \sum_{(j,j') \in \mathcal{C}^{(\mathrm{different})}} \mathcal{D}_{\mathbb{H}^{D^{(\mathrm{t})}}_K}(\tilde{\boldsymbol{\mu}}_j, \tilde{\boldsymbol{\mu}}_{j'}) \qquad (23)
$$

と表せる．ここで，$\mathcal{C}^{(\mathrm{same})} \subset \mathcal{T} \times \mathcal{T}$ は同一楽器類に属する音色インデクスペアの集合，$\mathcal{C}^{(\mathrm{different})} \subset \mathcal{T} \times \mathcal{T}$ は異種楽器類に属する音色インデクスペアの集合である．音色埋め込みに楽器分類の階層性を活用する目的においては，j と j' が同一楽器類のときは距離が小さく，j と j' が異種楽器類のときは距離が大きくなることが望ましい．その両観点を一つの指標で評価するため，楽器間距離比

$$
S = \frac{\mathcal{D}^{(\mathrm{different})}}{\mathcal{D}^{(\mathrm{same})}} \qquad (24)
$$

を定義する．この値 S が大きいほど，楽器間の分類上の近さを反映した埋め込みといえる．

表 1 音色の潜在空間の幾何・半径・次元数を変更したときの観測データの対数尤度

Geometry of timbre latent space	R	$D^{(\mathrm{t})}$ = 4	8	16
Euclid	-	**-187**	-258	**-381**
Hyperbolic (Proposed)	1	-203	**-250**	-568
	2	-210	-328	-568
	5	-246	-386	-665
	10	-273	-446	-771

5.4 結果と考察

各実験条件における対数尤度を**表 1** に示す．音色の潜在空間が 8 次元のとき，提案手法の条件の一つが従来手法を上回る尤度を達成したものの，他の次元数では Euclid 空間を用いた手法の方が高かった．

これらの尤度の差が聴感的にも顕著かを調べるため，著者の一人が生成音を聴取した．楽音を生成する際は，事前分布として学習された混合正規分布から，所望の音高 $y^{(\mathrm{p})}$・

表 2 音色の潜在空間の幾何・半径・次元数を変更したときの楽器間距離比

Geometry of timbre latent space	R	$D^{(\mathrm{t})}$ 4	8	16
Euclid	-	1.210	1.075	1.100
Hyperbolic (Proposed)	1	1.206	**1.271**	**1.186**
	2	1.161	1.084	1.176
	5	**1.312**	1.042	1.168
	10	1.181	1.051	1.162

音色 $y^{(\mathrm{t})}$ に対応する $\boldsymbol{\mu}^{(\mathrm{p})}_{y^{(\mathrm{p})}}, \tilde{\boldsymbol{\mu}}^{(\mathrm{t})}_{y^{(\mathrm{t})}}$ を $\boldsymbol{z}^{(\mathrm{p})}, \tilde{\boldsymbol{z}}^{(\mathrm{t})}$ として用いてメルスペクトログラムを生成した．それを振幅スペクトログラムに変換し，Griffin–Lim アルゴリズム [17] により位相を付与して逆短時間 Fourier 変換により波形に変換した．聴取した限り，全体的に音高が所望のものと異なるもの，音色が掠れているもの，生成音の音色がほぼ区別できないものが観察されたものの，8 次元で半径が 1 の潜在空間を用いた場合のモデルに関してはそのような問題は観察されなかった．この結果から，生成音の品質と埋め込みの尤度は必ずしも一致しないことが示唆される．そのため，尤度のみでは十分評価できない可能性があり，より正確に生成音の品質を比較するには主観評価を行う必要がある．

次に，音色に関する潜在空間を調べるため楽器間距離比を計算した（**表 2** 参照）．すべての次元数において，提案手法の条件のうち少なくとも一つが従来手法を上回る楽器間距離比を達成した．よって，適切な曲率の Lorentz モデルを用いることで，従来の VAE の確率モデルを保ちながら，楽器間の分類上の近さを反映した埋め込みを獲得できることが示唆される．特に，8 次元で半径 1 のときの楽器間距離比が同一次元の他の条件より高い値であることは，上述の聴感による評価と整合する．このことから，楽器分類の階層性を反映した埋め込みが楽音合成の品質にも寄与することが示唆される．

6. 結論と展望

本研究では，楽器の種類が階層的に分類できることに着目し，階層構造を埋め込むのに適した双曲空間を潜在空間に用いた VAE を提案した．提案法では，Luo ら [3] の楽音合成手法をベースとして，音色に関する潜在変数を双曲空間で定義し，その事前分布として擬似双曲正規分布を導入した．擬似双曲正規分布の性質により，双曲空間であっても事後分布や事前分布のパラメータに関する勾配が計算できるため，従来法と同様に確率的勾配降下法を用いて学習できることも示した．楽音合成実験により，音色に関する潜在空間において，Euclid 空間を用いる場合に比べ双曲空間を用いることで，同一楽器類はより近くに異種楽器類はより遠くへと埋め込まれることが示唆された．

現在使用しているデータセットには 12 種類の音色が含まれるが，より多様な音色を収録したデータセットを用いることで，音色の階層性の表現力を評価しやすくなると考えられる．また，評価指標に関してもより階層性を捉えた指標を調査する予定である．

謝辞 本研究は JSPS 科研費 19H01116 の助成を受けたものである．

参考文献

[1] D. P. Kingma and M. Welling: Auto-Encoding Variational Bayes, *Proc. International Conference on Learning Representations* (2014).
[2] M. Huzaifah and L. Wyse: Deep Generative Models for Musical Audio Synthesis, *Handbook of Artificial Intelligence for Music: Foundations, Advanced Approaches, and Developments for Creativity* (E. R. Miranda, ed.), Springer International Publishing, pp. 639–678 (2021).
[3] Y.-J. Luo, K. Agres and D. Herremans: Learning Disentangled Representations of Timbre and Pitch for Musical Instrument Sounds Using Gaussian Mixture Variational Autoencoders, *Proc. International Society for Music Information Retrieval Conference*, pp. 746–753 (2019).
[4] Y.-J. Luo, K. W. Cheuk, T. Nakano, M. Goto and D. Herremans: Unsupervised Disentanglement of Pitch and Timbre for Isolated Musical Instrument Sounds, *Proc. International Society for Music Information Retrieval Conference*, pp. 700–707 (2020).
[5] K. Tanaka, R. Nishikimi, Y. Bando, K. Yoshii and S. Morishima: Pitch-Timbre Disentanglement Of Musical Instrument Sounds Based On VAE-Based Metric Learning, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 111–115 (2021).
[6] A. Caillon and P. Esling: RAVE: A Variational Autoencoder for Fast and High-Quality Neural Audio Synthesis, *arXiv:2111.05011* (2021).
[7] D. P. Kingma, S. Mohamed, D. J. Rezende and M. Welling: Semi-supervised Learning with Deep Generative Models, *Proc. Advances in Neural Information Processing Systems*, Vol. 2, pp. 3581–3589 (2014).
[8] E. M. von Hornbostel and C. Sachs: Classification of Musical Instruments: Translated from the Original German by Anthony Baines and Klaus P. Wachsmann, *The Galpin Society Journal*, Vol. 14, pp. 3–29 (1961).
[9] H. F. F. Garcia, A. Aguilar, E. Manilow and B. Pardo: Leveraging Hierarchical Structure for Few-Shot Musical Instrument Recognition, *Proc. International Society for Music Information Retrieval Conference*, pp. 220–228 (2021).
[10] C. De Sa, A. Gu, C. Ré and F. Sala: Representation Tradeoffs for Hyperbolic Embeddings, *Proc. International Conference on Machine Learning*, Vol. 80, pp. 4460–4469 (2018).
[11] W. Peng, T. Varanka, A. Mostafa, H. Shi and G. Zhao: Hyperbolic Deep Neural Networks: A Survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
[12] Y. Nagano, S. Yamaguchi, Y. Fujita and M. Koyama: A Wrapped Normal Distribution on Hyperbolic Space for Gradient-Based Learning, *Proc. International Conference on Machine Learning*, pp. 4693–4702 (2019).
[13] O. Skopek, O.-E. Ganea and G. Bécigneul: Mixed-Curvature Variational Autoencoders, *Proc. International Conference on Learning Representations* (2020).

[14] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen and R. Pang: Hierarchical Generative Modeling for Controllable Speech Synthesis, *Proc. International Conference on Learning Representations* (2019).
[15] G. Ballet, R. Borghesi, P. Hoffmann and F. Lévy: Studio Online 3.0: An Internet "Killer Application" for Remote Access to IRCAM Sounds and Processing tools, *Journées d'Informatique Musicale* (1999).
[16] Y. Burda, R. B. Grosse and R. Salakhutdinov: Importance Weighted Autoencoders, *Proc. International Conference on Learning Representations* (2016).
[17] D. Griffin and J. Lim: Signal Estimation from Modified Short-time Fourier Transform, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 32, No. 2, pp. 236–243 (1984).