Vol.2019-MUS-124 No.5
2019/8/28

音楽音響信号に対するラベル・テクスチャ分離型変分自己符号化器を用いた半教師ありコード推定

呉 益明[1,a)]　Tristan Carsault[2,b)]　中村 栄太[1,c)]　吉井 和佳[1,d)]

概要：
本稿では，正解コードラベル付きの音楽音響信号（教師ありデータ）に加えて，ラベルが付与されていない音響信号（教師なしデータ）を同時に利用するための，深層ニューラルネットワーク (DNN) に基づくコード推定法について述べる．従来の DNN に基づく識別的アプローチは，大量の教師ありデータを用いることで優れた推定精度を達成できるが，コードラベルの付与には多大な労力が必要であり，精度向上には限界があった．一方，隠れマルコフモデルなどの確率モデルに基づく生成的アプローチは，原理的に半教師あり学習が可能であるものの，モデルの表現力の貧弱さから，推定精度の面で劣っていた．これらの問題を解決するため，本研究では，高い表現力を持つ DNN に基づく深層生成モデルと，償却型変分推論法に基づく半教師あり学習法を提案する．具体的には，まず，コードラベル系列と音響テクスチャ系列を潜在変数とし，音響的特徴量を観測変数とする生成モデルを定式化する．観測変数が与えられた際に，潜在変数の事後分布を推定するため，音響的特徴量からコードラベル系列を推定する識別モデルと，音響的特徴量とコードラベル系列から音響テクスチャ系列を抽出する推論モデルを導入する．与えられた音楽音響信号に対して，教師ラベルの有無に関わらず，変分自己符号化器の枠組みでこれら三つの深層モデルを同時最適化することができる．実験の結果，教師なしデータに対しても，コードラベル情報と音響テクスチャ情報が適切に分離された表現学習を行うことができること，半教師あり学習を行った識別モデルが，教師ありデータのみで学習した識別モデルよりも高い認識精度を実現できることを確認した．

1. はじめに

自動コード推定の研究は，音楽音響信号からコードラベル系列を推定する過程を数理モデルで再現することを目指している．音楽音響信号の構造は複雑で，離散的な記号形式で定義されるコード進行との繋がりも非自明的な部分が多く，数理モデルを明示的に記述することが難しい．これまでの研究では，主に統計的機械学習の手法を用いて，データ駆動型アプローチで問題の解決を試みられてきた．

従来の研究では，コード記号系列から音響特徴量系列への「演奏過程」を確率的生成モデルとして定式化するアプローチが主に用いられてきた [1–4]．コード推定を行う際は，観測された音響特徴量の生成確率（尤度）が最大になるようなコードラベル系列を逆算する逆問題を解く．これらの研究では生成モデルの学習を容易にするため，音響モデルを音響信号から低次元な音響特徴量（主にクロマベクトル [5]）を抽出して観測値として扱っており，その音響モデルの設計も主な研究課題である [2, 6]．生成的モデルは解釈性が高く，モデルを拡張したり制約を課すことで生成過程に音楽的なドメイン知識を反映させることができる．また，理論的に半教師あり・教師なし学習が可能であるという利点もある．一方，効率的にコード推定が実行可能なモデルはほぼ隠れマルコフモデル (HMM) に基づいたものに限られており，モデルの表現力も限定的であるため，認識精度は高めるのは困難だった．

一方，近年では，観測された音響特徴量からコードラベル系列の事後確率を直接推定する識別的アプローチに基づく手法が注目されており，特に深層ニューラルネットワーク (DNN) モデルに基づいた教師あり学習法が多く提案されている．十分な量の学習データ（音響信号と正しいコードラベル系列のペアデータ）を強力な表現力を持つ DNN に学習させることで，未知データに対して比較的高い精度で事後確率を推定する認識モデルを獲得することに成功している．しかし識別的モデルの性能はアノテーションが存在する音響信号の質・量に強く依存しており，アノテーショ

[1] 京都大学情報学研究科
Graduate School of Informatics, Kyoto University
[2] IRCAM, CNRS, Sorbonne Université
[a)] wu@sap.ist.i.kyoto-u.ac.jp
[b)] carsault@ircam.fr
[c)] enakamura@sap.ist.i.kyoto-u.ac.jp
[d)] yoshii@sap.ist.i.kyoto-u.ac.jp

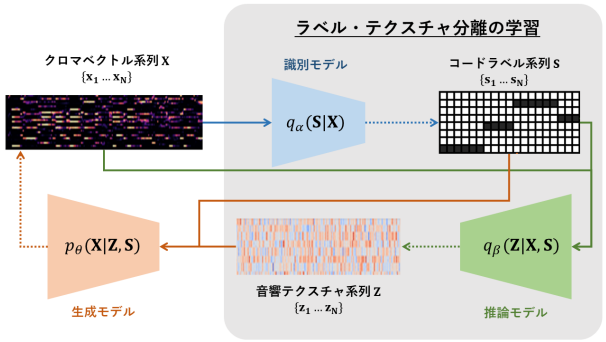

図 1 本論文が提案する，深層生成モデル・深層コード識別モデルと深層テクスチャ推論モデルを組み合わせた変分自己符号化器 (VAE) の概観図．点線は reparametrization trick によるサンプリングを表す．

ンが無い音響信号から学習ができない．また，コードラベル系列に関するドメイン知識の反映も容易ではない．

本研究では，両アプローチの限界を克服するため，深層生成モデルの半教師あり学習法を提案する．本モデルは，潜在変数であるコードラベル系列と音響テクスチャ系列から，観測変数であるクロマベクトル系列が生成される過程を表現する．ここで，音響テクスチャとは，クロマベクトルを表現するうえで必要となる，シンボリックなコードラベル以外の音響的な情報，例えば，コード区間に含まれる各音符の音量や経過音の存在などを示していると考えられる変数であり，便宜的に多変量標準ガウス分布に従うことを仮定する．また，コードラベル系列は，フレームレベルでの過剰な遷移を防ぐため，高い自己遷移確率をもつ一次マルコフモデルに従うことを仮定している．

さらに，クロマベクトル系列からコードラベル系列の事後確率を推定する深層識別モデルと，クロマベクトル系列とコードラベル系列から音響テクスチャ系列を推定する深層推論モデルを導入する．これらを深層生成モデルと統合することで，変分自己符号化器 (VAE) [7] を構成できる（図 1）．この統合モデルは，一部だけ教師データが存在する音楽音響データを用いて半教師あり学習することができ，コードラベルと音響テクスチャを適切に分離した推論・生成モデルを獲得することができる．

本研究の主な貢献は，強力な表現力と柔軟性を持つ深層生成モデルをはじめて自動コード推定モデルの研究に導入したことである．また，識別的・生成的モデルを統合した VAE を初めて自動コード推定モデルに応用し，半教師あり学習によるコード識別モデル精度が向上することを示す．技術面では，VAE を学習する際，コードラベル系列の事前分布をマルコフモデルで定義し，識別モデルの出力を正則化することで、コードラベル系列の時間的継続性を明示的に反映できることを示す．

2. 関連研究

本章では，音楽音響信号に対するコード推定と，生成的・識別的アプローチの統合の関連研究について述べる．

2.1 統計的コード推定

統計的コード推定の核心は，音響特徴量系列 $\mathbf{X}$ からコードラベル系列 $\mathbf{S}$ を採譜する過程を事後確率 $p(\mathbf{S}|\mathbf{X})$ で表現することである．モデル $p(\mathbf{S}|\mathbf{X})$ を定式化する方法は，生成的・識別的アプローチの二種に大別される．

生成的アプローチでは，隠れマルコフモデル (HMM) に基づく定式化が主流である [3, 8]．この手法では，潜在変数系列 $\mathbf{S}$ がマルコフモデルに従う $p(\mathbf{S})$ という仮定に基づき，観測系列 $\mathbf{X}$ の生成過程 $p(\mathbf{X}|\mathbf{S})$ を定式化する．$\mathbf{X}$ が与えられたときに $p(\mathbf{S}|\mathbf{X})$ を最大化する $\mathbf{S}$ は，ビタビアルゴリズムで効率的に求めることができる．Mauch ら [2] は，HMM の枠組みを拡張し，拍・調・コードおよび低音の状態を潜在変数とした動的ベイジアンネットワーク (DBN) を提案している．Ni ら [3] も，コードラベル系列・転置状態・調を潜在変数とした DBN を提案している．ほかには，周波数域が異なる複数のクロマ特徴系列をマルチストリーム HMM でモデリングする [6, 9]，コードの持続時間を明示的に考慮したセミマルコフモデルを用いる [4] などの拡張が提案されている．

一方識別的アプローチでは，Humphrey と Bello [10] が初めて畳み込みニューラルネットワーク（CNN）に基づくコード推定モデルを提案して以来，DNN に基づいた手法が多数提案されている．DNN モデルの特徴は，その深層構造によって入力データの高次な特徴が自律的に学習されることである．この特徴を活かし，自動コード推定のタスクにおいても，明示的な音響モデルによる処理を経ず、CQT スペクトログラムなどの低次な表現から高い精度でコードの事後確率を推定することに成功している [11]。そのほかに，タスクに適したな中間表現を DNN の学習によって獲得する手法 [12, 13] も提案されている．

正確なコード推定を行うためには，各時点（フレーム）における正確な事後確率の推定のほかに，コードラベル系列自体の時間特性を考慮する必要がある．HMM に基づく生成的アプローチでは，潜在変数であるコードラベル系列がマルコフ過程に沿うことを明示的に仮定しており，ビタビアルゴリズムで最尤なコードラベル系列を推定する過程でその時間特性（ラベル遷移確率）が反映される．一方識別的アプローチでは，フレーム単位の事後確率を「言語モデル」と統合し，後処理の段階でコードラベル系列の時間特性を反映させる推定手法が提案されている [14, 15]．また，リカレントニューラルネットワーク (RNN) による分類器を用いて，ニューラルネット自身に一定の時間特性を学

習させることも可能である [13,16].

2.2 生成的・識別的アプローチの統合

DNN に基づく生成的モデルと識別的モデルを統合する手法は，主に半教師あり・教師なし機械学習の文脈で検討されている．与えられたラベルと分離した潜在表現を学習する手法として，変分自己符号化器 (VAE) の拡張である条件付き変分自己符号化器 (CVAE) [17,18] の有効性が示されている．Makhzani ら [19] が提案した半教師あり学習法では，敵対的学習を用いて表現の分離を促進している．音声認識の研究では，音声をテキストに変換する音声認識モデルと，テキストから音声を合成する音声合成モデルを統合することで，少ないペアデータを用いてより高精度な音声認識モデルを獲得する手法が提案されている [20].

生成的・識別的モデルは，自動コード推定のタスクにおいても相補的な性質を持つと思われる．本研究では，その相補性を活かすために，二つのモデルを統一的に最適化する手法を提案し，識別的モデルの認識精度を保つと同時に，半教師あり学習が可能なコード推定モデル学習機構を実現する．この手法の特徴は，コード状態と分離した表現である「音響テクスチャ」を潜在変数として，深層生成モデルに導入したところで，これにより CVAE に類似した手法でモデルの最適化を行うことが可能になる．

3. 提案手法

本章では，提案手法である深層生成モデルに基づいた自動コード推定モデルの定式化・学習方法について述べる．

3.1 問題の定式化

音楽音響信号のフレームレベルの特徴量系列を $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ とする．$\mathbf{x}_n \in [0,1]^D$ は D 次元の非負ベクトルで，N は総フレーム数である．$\mathbf{x}_n$ は，各フレームの低域，中域，高域の音高クラス状態を三つのクロマベクトル ($D = 36$) で表し，あらかじめ DNN に基づく特徴抽出器 [13] で音響信号から獲得しておく．$\mathbf{X}$ のコードラベル系列を $\mathbf{S} = \{\mathbf{s}_1, \cdots, \mathbf{s}_N\}$ とする．$\mathbf{s}_n \in \{0,1\}^K$ はサイズ K のコード語彙のうちの一つを表す，K 次元の one-hot ベクトルである．本研究では，コード語彙は 6 種類のトライアドコード (*maj, min, dim, aug, sus2, sus4*) に加え，*no-chord* ラベルを含むため，$K = 73$ である．もう一つの潜在変数系列を $\mathbf{Z} = \{\mathbf{z}_1, \cdots, \mathbf{z}_N\}$ とする．$\mathbf{z}_n \in \mathbb{R}^L$ はフレーム n における特徴ベクトルの音響テクスチャを表す L 次元のベクトルである ($L = 8$).

本研究の目標は，学習データに対して識別モデル $p(\mathbf{S}|\mathbf{X})$ を学習し，未知の特徴量系列が与えられたときに，対応するコード系列を推定するために用いることである．通常，$p(\mathbf{S}|\mathbf{X})$ は，人手によるアノテーションで作成された $\mathbf{X}$ と $\mathbf{S}$ のペアデータを用いて教師あり学習される．一方，本研究では，ペアデータのほかに，教師データが存在しない音響信号も用いて，$p(\mathbf{S}|\mathbf{X})$ を半教師あり学習する．

3.2 モデルの定式化

潜在変数であるラベル系列 $\mathbf{S}$ およびテクスチャ系列 $\mathbf{Z}$ と，観測値であるクロマベクトル系列 $\mathbf{X}$ からなる確率的生成モデルを定式化する．

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

ここで，$p(\mathbf{X}|\mathbf{S}, \mathbf{Z})$ は $\mathbf{S}, \mathbf{Z}$ に条件付けられた $\mathbf{X}$ の尤度，$p(\mathbf{S})$ と $p(\mathbf{Z})$ はそれぞれ $\mathbf{S}$ と $\mathbf{Z}$ の事前分布である．標準的な HMM は $p(\mathbf{X}, \mathbf{S}) = p(\mathbf{X}|\mathbf{S}) p(\mathbf{S})$ で与えられるが，本研究では，$\mathbf{X}$ のをより精緻にモデル化するため，明示的に潜在変数 $\mathbf{Z}$ を導入した．具体的には，$p(\mathbf{X}|\mathbf{S}, \mathbf{Z})$ は深層生成モデルとして定式化される．

$$p_\theta(\mathbf{X}|\mathbf{S}, \mathbf{Z}) = \prod_{n=1}^{N} \prod_{d=1}^{D} \mathrm{Bernoulli}(x_{nd} | [\boldsymbol{\omega}_\theta(\mathbf{S}, \mathbf{Z})]_{nd}) \tag{2}$$

$\boldsymbol{\omega}_\theta(\mathbf{S}, \mathbf{Z})$ は，パラメータは θ を持ち，$\mathbf{S}$ と $\mathbf{Z}$ を入力にとる再帰型ニューラルネットワーク (RNN) の，次元数が ND の出力であり，$[\mathbf{A}]_{ij}$ は $\mathbf{A}$ の ij 番目の要素を指す．

事前分布 $p(\mathbf{S})$ と $p(\mathbf{Z})$ は，それぞれ潜在変数 $\mathbf{S}$ と $\mathbf{Z}$ の特性を反映するように定義する．まず，コードラベル系列 $\mathbf{S}$ は時間的な連続性を持つことから，$\mathbf{S}$ は一次マルコフモデルに沿うと仮定する．

$$\begin{aligned} p_\phi(\mathbf{S}) &= p(\mathbf{s}_1) \prod_{n=2}^{N} p(\mathbf{s}_n | \mathbf{s}_{n-1}) \\ &= \prod_{k=1}^{K} \phi_k^{s_{1k}} \prod_{n=2}^{N} \prod_{k'=1}^{K} \prod_{k=1}^{K} \phi_{k'k}^{s_{n-1,k'} s_{nk}} \end{aligned} \tag{3}$$

ここで，ϕ_k はコード k の初期確率，$\phi_{k'k}$ はコード k' からコード k への遷移確率を指しており，自己遷移確率 ϕ_{kk} が高くなるよう設定する．VAE の枠組みで，潜在変数の事前分布にマルコフモデを用いたのは本研究が初めてである．一方，テクスチャ系列 $\mathbf{Z}$ は抽象化な表現であるので，一般的な VAE と同様に，標準ガウス分布に従うものとする．

$$p(\mathbf{Z}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n | \mathbf{0}_L, \mathbf{I}_L) \tag{4}$$

ここで，$\mathbf{0}_L$ は L 次元のゼロベクトル，$\mathbf{I}_L$ はサイズ L の単位行列である．

3.3 教師なし変分推論

$\mathbf{X}$ が与えられたうえでの事後分布 $p(\mathbf{S}, \mathbf{Z}|\mathbf{X})$ は解析的に求めることが困難なため，償却型変分推論法 (amortized variational inference) [21] を用いる．具体的には，DNN を用いてパラメタライズされた変分事後分布 $q(\mathbf{S}, \mathbf{Z}|\mathbf{X})$ を導入し，$p(\mathbf{S}, \mathbf{Z}|\mathbf{X})$ と $q(\mathbf{S}, \mathbf{Z}|\mathbf{X})$ の Kullback-Leibler 距離を最

小化する．これは，対数周辺尤度 $\log p(\mathbf{X})$ の変分下限を $q(\mathbf{S},\mathbf{Z}|\mathbf{X})$ に対して最大化することと等価である．本研究では，$q(\mathbf{S},\mathbf{Z}|\mathbf{X})$ に対して，以下の分解形を仮定する．

$$q(\mathbf{S},\mathbf{Z}|\mathbf{X}) = q(\mathbf{S}|\mathbf{X})q(\mathbf{Z}|\mathbf{X},\mathbf{S}) \tag{5}$$

ここで，$q(\mathbf{S}|\mathbf{X})$ と $q(\mathbf{Z}|\mathbf{X},\mathbf{S})$ は，それぞれ潜在変数 $\mathbf{S}$ と $\mathbf{Z}$ を推定する深層識別モデルと深層推論モデルであり，RNN を用いて定式化する．

$$q_\alpha(\mathbf{S}|\mathbf{X}) = \prod_{n=1}^{N} \mathrm{Categorical}(\mathbf{s}_n|[\boldsymbol{\pi}_\alpha(\mathbf{X})]_n) \tag{6}$$

$$q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{z}_n|[\boldsymbol{\mu}_\beta(\mathbf{X},\mathbf{S})]_n, [\boldsymbol{\sigma}^2_\beta(\mathbf{X},\mathbf{S})]_n) \tag{7}$$

ここで，$\boldsymbol{\pi}_\alpha(\mathbf{X})$ は $\mathbf{X}$ を入力とし，パラメータ α をもつ RNN の NK 次元の出力，$\boldsymbol{\mu}_\beta(\mathbf{X})$ と $\boldsymbol{\sigma}^2_\beta(\mathbf{X})$ は $\mathbf{X}$ と $\mathbf{S}$ を入力とし，パラメータ β をもつ RNN の NL 次元の出力である．

このとき，イェンセンの不等式を利用して，$\log p(\mathbf{X})$ の下限を以下のように導出できる．

$$\begin{aligned}
\log p(\mathbf{X}) &= \log \iint \frac{q(\mathbf{S},\mathbf{Z}|\mathbf{X})}{q(\mathbf{S},\mathbf{Z}|\mathbf{X})} \log p(\mathbf{X},\mathbf{S},\mathbf{Z}) d\mathbf{S} d\mathbf{Z} \\
&\geq \iint q(\mathbf{S},\mathbf{Z}|\mathbf{X}) \log \frac{p(\mathbf{X},\mathbf{S},\mathbf{Z})}{q(\mathbf{S},\mathbf{Z}|\mathbf{X})} d\mathbf{S} d\mathbf{Z} \\
&= \mathbb{E}_{q_\alpha(\mathbf{S}|\mathbf{X})q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})}[\log p_\theta(\mathbf{X}|\mathbf{S},\mathbf{Z})] \\
&\quad + \mathbb{E}_{q_\alpha(\mathbf{S}|\mathbf{X})q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})}[\log p(\mathbf{Z}) - \log q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})] \\
&\quad + \mathbb{E}_{q_\alpha(\mathbf{S}|\mathbf{X})}[\log p_\theta(\mathbf{S}) - \log q_\alpha(\mathbf{S}|\mathbf{X})]] \\
&\approx \frac{1}{I}\sum_{i=1}^{I} \big(\log p_\theta(\mathbf{X}|\mathbf{S}_i,\mathbf{Z}_i) - \mathrm{KL}(q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S}_i)\|p(\mathbf{Z}))\big) \\
&\quad + \mathrm{Entropy}[q_\alpha(\mathbf{S}|\mathbf{X})] + \mathbb{E}_{q_\alpha(\mathbf{S}|\mathbf{X})}[\log p_\theta(\mathbf{S})] \\
&\quad \triangleq \mathcal{L}_{\mathbf{X}}(\theta,\phi,\alpha,\beta)
\end{aligned} \tag{8}$$

$\{\mathbf{S}_i,\mathbf{Z}_i\}_{i=1}^I$ は分布 $q_\alpha(\mathbf{S}|\mathbf{X})q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})$ からの I 個のサンプルである．本研究では，標準的な VAE と同様に $I=1$ とするので，以降，添字 i は省略する．式 (8) の第一項から第三項は解析的に求めることができる．$\mathbb{E}_{q_\alpha(\mathbf{S}|\mathbf{X})}[\log p_\theta(\mathbf{S})]$ は，$\gamma(\mathbf{s}_t) = \mathbb{E}_{q_\alpha(\mathbf{S}|\mathbf{X})}[\log p(\mathbf{s}_{1:n})]$ とすると，HMM の前向きアルゴリズムに似た方法で再帰的に計算することができる．

$$\gamma(\mathbf{s}_1) = \log p(\mathbf{s}_1) \tag{9}$$

$$\gamma(\mathbf{s}_n) = \sum_{\mathbf{s}_{n-1}} q_\alpha(\mathbf{s}_{n-1}|\mathbf{X})\big(\gamma(\mathbf{s}_{n-1}) + \log p(\mathbf{s}_n|\mathbf{s}_{n-1})\big) \tag{10}$$

$$\mathbb{E}_{q_\alpha(\mathbf{S}|\mathbf{X})}[\log p_\theta(\mathbf{S})] = \sum_{\mathbf{s}_N} q_\alpha(\mathbf{s}_N|\mathbf{X})\gamma(\mathbf{s}_N) \tag{11}$$

期待値を計算するため，reparametrization trick を用いて確率変数 $\mathbf{S}$ と $\mathbf{Z}$ のサンプルを以下のように求める．

$$\boldsymbol{\epsilon}^{\mathbf{s}}_n \sim \mathrm{Gumbel}(\mathbf{0}_K, \mathbf{1}_K) \tag{12}$$

$$\mathbf{s}_n = \mathrm{softmax}(\log[\boldsymbol{\pi}_\alpha(\mathbf{X})]_n + \boldsymbol{\epsilon}^{\mathbf{s}}_n)/\tau) \tag{13}$$

$$\boldsymbol{\epsilon}^{\mathbf{z}}_n \sim \mathcal{N}(\mathbf{0}_L, \mathbf{I}_L) \tag{14}$$

$$\mathbf{z}_n = [\boldsymbol{\mu}_\beta(\mathbf{X},\mathbf{S})]_n + \boldsymbol{\epsilon}^{\mathbf{z}}_n \odot [\boldsymbol{\sigma}_\beta(\mathbf{X},\mathbf{S})]_n \tag{15}$$

ここで，$\mathbf{1}_K$ は K 次元の 1 値ベクトル，$\odot$ は要素ごとの乗算をあらわし，$\tau > 0$ は $\mathbf{s}_n$ の特性を調整するパラメータである（ここでは $\tau = 0.1$ とする）．

教師なし学習の目標は，モデルパラメータ $\theta, \phi, \alpha, \beta$ に関する関数 $\mathcal{L}_{\mathbf{X}}(\theta,\phi,\alpha,\beta)$ を最大化することである．モデルパラメータは Adam アルゴリズムを使い，勾配降下法で最適化される．

3.4 教師あり・半教師あり変分推論

$\mathbf{X}$ と $\mathbf{S}$ が既知の場合は，$q_\alpha(\mathbf{S}|\mathbf{X})$ の教師あり学習が可能である．同時に，$q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})$ と $p_\theta(\mathbf{X}|\mathbf{S},\mathbf{Z})$ の学習も行える．この場合では，以下のように $\log p(\mathbf{X}|\mathbf{S})$ の変分下限を求め，最大化する．

$$\begin{aligned}
\log p(\mathbf{X}|\mathbf{S}) &= \log \int \frac{q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})}{q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})} p(\mathbf{X},\mathbf{Z}|\mathbf{S}) d\mathbf{Z} \\
&\geq \int q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S}) \log \frac{p(\mathbf{X},\mathbf{Z}|\mathbf{S})}{q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})} d\mathbf{Z} \\
&\approx \frac{1}{I}\sum_{i=1}^{I} \log p_\theta(\mathbf{X}|\mathbf{S},\mathbf{Z}_i) - \mathrm{KL}(q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})\|p(\mathbf{Z})) \\
&\triangleq \mathcal{L}'_{\mathbf{X},\mathbf{S}}(\theta,\phi,\beta)
\end{aligned} \tag{16}$$

$\{\mathbf{Z}_i\}_{i=1}^I$ は分布 $q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})$ からのサンプルである．コード推定モデル学習の核心である識別モデルの項 $q_\alpha(\mathbf{S}|\mathbf{X})$ が式 (16) に含まれていないため，CVAE の手法にならい，目標関数は式 (16) の変分下限とコード推定モデルの識別モデル項を足し合わせたものとする。

$$\mathcal{L}_{\mathbf{X},\mathbf{S}}(\theta,\phi,\alpha,\beta) = \mathcal{L}'_{\mathbf{X},\mathbf{S}}(\theta,\phi,\beta) + \log q_\alpha(\mathbf{S}|\mathbf{X}). \tag{17}$$

半教師あり学習，つまり一部の音楽音響信号にしか正解コードラベル系列が与えられていない場合，目標関数は式 (8) と式 (17) を足し合わせたものとする．

$$\mathcal{L}(\theta,\phi,\alpha,\beta) \triangleq \sum_{\mathbf{X}\text{ only}} \mathcal{L}_{\mathbf{X}}(\theta,\phi,\alpha,\beta) + \sum_{\mathbf{X}\text{ with }\mathbf{S}} \mathcal{L}_{\mathbf{X},\mathbf{S}}(\theta,\phi,\alpha,\beta) \tag{18}$$

実際に学習を行う際は，まず式 (18) の第二項を教師あり学習で収束するまで最適化してから，式 (18) 全体を教師あり・教師なしデータを用いて最適化する．

4. 評価実験

本章では，提案する半教師あり学習法の有効性を検証するために行ったコード推定実験の結果を報告する．

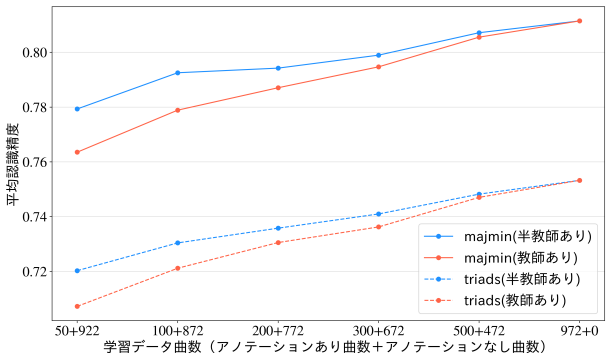

図 2 異なる教師ありデータ比率での平均コード認識精度グラフ．

4.1 実験条件

本研究では，正解コードラベルのアノテーションを持つ四つの公開データセット（計 1215 曲）を用いた．

- 220 曲（主にビートルズの作品）を含む Isophonics データセット [22]
- 100 曲の日本語・英語歌詞のポピュラー音楽を含む RWC-MDB-P データセット [23]
- 187 曲の英語歌詞のポピュラー音楽を含む uspop2002 データセット [24] [*1]
- 708 曲の音楽を含む McGill Billboard データセット [25]

これらのデータに対して 5 分割交差検証を行うことで，コード推定精度を評価した．各分割の学習データ部分（計 972 曲）のうち，ランダムに選らばれた 50・100・200・300・500 曲を教師ありデータ，残りを教師なしデータとし，提案した半教師あり学習法を実行した．コード推定結果は各楽曲ごとにテキスト形式に書き出してから，*mir_eval* ライブラリ [26] でアノテーションを基準に重み付き精度を計算した．認識結果の成否基準は，メジャー・マイナートライアドのみ（計 25 種）を考慮する *majmin* 基準と，提案モデルが対応するすべてのトライアド (計 73 種) を考慮する *triads* 基準を用いた．いずれも *mir_eval* ライブラリに実装されている．評価セット全体の精度は，各楽曲の長さで重み付き平均した認識精度で示す．

各深層モデルは多層双方向 LSTM ネットワークで実装した．層数は 3 とし，各隠れ層は 256 個の LSTM セルを，順方向・逆方向にそれぞれ持つ．各時間ステップ n では，$\mathbf{X}_n$ が $q_\alpha(\mathbf{S}|\mathbf{X})$ に入力され，$\mathbf{z}_n$ と $\mathbf{s}_n$ の連結ベクトル，$\mathbf{X}_n$ と $\mathbf{s}_n$ の連結ベクトルがそれぞれ $p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{S})$ と $q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})$ に入力される．式 (3) の遷移確率は，ここでは $\phi_{kk} = \frac{17}{89}$ とし，ϕ は学習過程では固定値として扱った．

4.2 実験結果

図 2 に，教師あり・なしデータの比率を変化させて学習したモデルのコード推定精度を示す．教師ありデータの量が多いほどコード推定精度は高くなった．また，データの比率によらず，半教師あり学習を行ったモデルは教師あり学習のみを行ったモデルよりも高い認識精度を達成している．教師なしデータの比率が大きいほど精度の差が顕著で，比率が小さくなるにつれ差が縮まった．この結果から，提案した半教師あり学習法の有効性を示された．

[*1] RWC-MDB-P および uspop2002 データセットのアノテーションは，Cho によって配布されたものを用いている．https://github.com/tmc323/ChordAnnotations から入手できる．

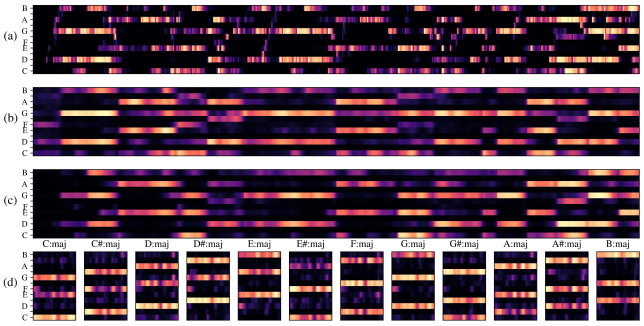

図 3 (a) クロマベクトル系列 $\mathbf{X}$．(b) 正解コードラベル系列 $\mathbf{S}$ と，$p(\mathbf{Z})$ からサンプルした $\mathbf{Z}$ を入力した場合の $p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{S})$ の期待値．(c) 正解コードラベル系列 $\mathbf{S}$ と，$q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})$ からサンプルした $\mathbf{Z}$ を入力した場合の $p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{S})$ の期待値．(d)$p(\mathbf{Z})$ からサンプルした $\mathbf{Z}$ を固定し，異なる $\mathbf{S}$ を入力した場合の $p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{S})$ の期待値．いずれの図も，$\boldsymbol{\omega}_\theta(\mathbf{Z},\mathbf{S})$ の真ん中の 12 次元のみを表示している．

我々はさらに，$\mathbf{S}$ と $\mathbf{Z}$ が分離された表現であるかを確認するために，定性的な実験を行った．**図 3** は，生成モデル $p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{S})$ に異なる $\mathbf{S}$ と $\mathbf{Z}$ が入力された時の期待値 $\boldsymbol{\omega}_\theta(\mathbf{Z},\mathbf{S})$ を可視化したものである．同じ $\mathbf{S}$ から生成された図 3(b) と図 3(c) は，いずれも $\mathbf{S}$ のコードの構成音を強調する傾向がみられた．しかし $q_\beta(\mathbf{Z}|\mathbf{X},\mathbf{S})$ からサンプリングした $\mathbf{Z}$ を入力とする図 3(b) は，事前確率 $p(\mathbf{Z})$ からサンプリングした $\mathbf{Z}$ を入力とする図 3(c) よりも，元の特徴量 $\mathbf{X}$(図 3(a)) に近い値が生成された．また，$\mathbf{Z}$ を固定し，$\mathbf{S}$ のみを変化させると，図 3(d) のようにコードの構成音に従って特徴量が生成された．これらの生成結果は，$\mathbf{S}$ と $\mathbf{Z}$ は分離された表現であり，それぞれ生成モデルにコード構成音とテクスチャ情報を与えていることを示唆している．

4.3 深層生成・識別モデル同時学習の有効性

通常，識別モデルの教師あり学習では，ペアデータに対する分類損失関数が利用されるが，生成モデルに基づく識別モデルの半教師あり学習では，生成モデルが識別モデルを改善するため評価関数の役割を担うことが期待されている（生成モデル自身も同時に更新される点に注意）．提案手法において，深層生成モデル $p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{S})$ が期待通りに機能しているか確認するために，我々は提案モデルを使い，以下の手順で学習を試みた．

(1) $\mathcal{L}_{\mathbf{X},\mathbf{S}}$ を最大化するようパラメータ β と θ を更新する．
(2) パラメータ β と θ を固定し，$\mathcal{L}_{\mathbf{X}}$ を最大化するよう識

別モデルのパラメータ α のみを更新する．

この手順により，疑似的に，識別モデル $q_\alpha(\mathbf{S}|\mathbf{X})$ を教師なし学習させたことになる．5 分割交差検証を行った結果，疑似的な教師なし学習で得られたコード推定モデルの精度は，*majmin* と *triad* 基準でそれぞれ 66.2%と 61.2%だった．これはアノテーションを用いた教師あり学習で得られたモデルの精度である 81.1%と 75.3%には及ばないものの，学習済みの生成モデルがある程度 $q_\alpha(\mathbf{S}|\mathbf{X})$ の評価関数として機能していることが示された．

5. おわりに

本研究では，DNN に基づくコード推定モデルの半教師あり学習法を提案した．音響特徴ベクトル系列 $\mathbf{X}$ からコードラベル系列 $\mathbf{S}$ とテクスチャ系列 $\mathbf{Z}$ を推定する深層推論モデル $q(\mathbf{S},\mathbf{Z}|\mathbf{X})$ と，$\mathbf{S}$ と $\mathbf{Z}$ から $\mathbf{X}$ を確率的に生成する深層生成モデル $p(\mathbf{X}|\mathbf{S},\mathbf{Z})$ を組み合わせた変分自己符号化器 (VAE) を構成し，正則化のため，$\mathbf{S}$ と $\mathbf{Z}$ に適切な事前分布を与えることで，これらのモデルを一挙に同時最適化できることを示した．

本手法は，主にポピュラー音楽で構成されたデータセットを用いて有効性を確認しているが，多様な音楽データを用いてモデルの汎化性能を改善し，コード推定精度をさらに向上することが課題である．また，現在のフレーム単位のコードラベル系列モデルの代わりに，拍単位・記号単位のコード言語モデルを事前分布に導入することで，音響特徴量の曖昧さを解消した音楽的に妥当なコード推定を実現することも重要な課題である．我々は提案した枠組みをさらに拡張し，調・拍・音符などの音楽要素を潜在変数として導入し，階層的な生成モデルを定式化することで，総合的な音楽採譜システムの実現を目指したい．

謝辞 本研究の一部は，JST No.JPMJAC1602 および JSPS 科研費 No. 16H01744, No.19H04137 による支援を受けた．

参考文献

[1] Sheh, A. and Ellis, D.: Chord segmentation and recognition using EM-trained hidden Markov models, *ISMIR*, pp. 185–191 (2003).
[2] Mauch, M. and Dixon, S.: Approximate Note Transcription For the Improved Identification of Difficult Chords, *ISMIR*, pp. 135–140 (2010).
[3] Ni, Y., McVicar, M., Santos-Rodriguez, R. and De Bie, T.: An End-to-End Machine Learning System for Harmonic Analysis of Music, *IEEE TASLP*, Vol. 20, No. 6, pp. 1771–1783 (2012).
[4] Chen, R., Shen, W., Srinivasamurthy, A. and Chordia, P.: Chord recognition using duration-explicit hidden Markov models, *ISMIR*, pp. 445–450 (2012).
[5] Fujishima, T.: Realtime Chord Recognition of Musical Sound: A System Using Common Lisp Music, *ICMC*, pp. 464–467 (1999).
[6] Khadkevich, M. and Omologo, M.: Time-frequency reassigned features for automatic chord recognition, *ICASSP*, pp. 181–184 (2011).
[7] Kingma, D. P. and Welling, M.: Auto-Encoding Variational Bayes, *ICLR*, pp. 1–14 (2014).
[8] McVicar, M., Santos-Rodriguez, R., Ni, Y. and Bie, T. D.: Automatic Chord Estimation from Audio: A Review of the State of the Art, *IEEE TASLP*, Vol. 22, No. 2, pp. 556–575 (online), DOI: 10.1109/TASLP.2013.2294580 (2014).
[9] Cho, T.: Improved techniques for automatic chord recognition from music audio signals, PhD Thesis, New York University (2014).
[10] Humphrey, E. J. and Bello, J. P.: Rethinking automatic chord recognition with convolutional neural networks, *ICMLA*, Vol. 2, pp. 357–362 (2012).
[11] Korzeniowski, F. and Widmer, G.: A fully convolutional deep auditory model for musical chord recognition, *MLSP*, pp. 13–16 (2016).
[12] Korzeniowski, F. and Widmer, G.: Feature Learning for Chord Recognition: The Deep Chroma Extractor, *ISMIR*, pp. 37–43 (2016).
[13] Wu, Y. and Li, W.: Automatic audio chord recognition with MIDI-trained deep feature and BLSTM-CRF Sequence Decoding Model, *IEEE TASLP*, Vol. 27, No. 2, pp. 355–366 (online), DOI: 10.1109/TASLP.2018.2879399 (2019).
[14] Sigtia, S., Boulanger-Lewandowski, N. and Dixon, S.: Audio Chord Recognition with a Hybrid Recurrent Neural Network, *ISMIR*, pp. 127–133 (2015).
[15] Korzeniowski, F. and Widmer, G.: Improved Chord Recognition by Combining Duration and Harmonic Language Models, *ISMIR*, pp. 10–17 (2018).
[16] Deng, J. and Kwok, Y.-K.: Large Vocabulary Automatic Chord Estimation with an Even Chance Training Scheme, *ISMIR*, pp. 531–536 (2017).
[17] Kingma, D. P., Mohamed, S., Rezende, D. J. and Welling, M.: Semi-supervised learning with deep generative models, *NIPS*, pp. 3581–3589 (2014).
[18] Maaløe, L., Sønderby, C. K., Sønderby, S. K. and Winther, O.: Auxiliary Deep Generative Models, *ICML*, pp. 1445–1453 (2016).
[19] Makhzani, A., Shlens, J., Jaitly, N. and Goodfellow, I.: Adversarial Autoencoders, *ICLR*, pp. 1–16 (2016).
[20] Hori, T., Astudillo, R. F., Hayashi, T., Zhang, Y., Watanabe, S. and Roux, J. L.: Cycle-consistency training for end-to-end speech recognition, *ICASSP*, pp. 6271–6275 (2019).
[21] Gershman, S. J. and D.Goodman, N.: Amortized inference in probabilistic reasoning, *Proceedings of the annual meeting of the cognitive science society*, Vol. 36, No. 36, pp. 517–522 (2014).
[22] Harte, C.: Towards automatic extraction of harmony information from music signals, PhD Thesis, Queen Mary University of London (2010).
[23] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC music database: Popular, classical, and jazz music databases, *ISMIR*, pp. 287–288 (2002).
[24] Berenzweig, A., Logan, B., Ellis, D. and Whitman, B.: A large-scale evaluation of acoustic and subjective music-similarity measures, *Computer Music Journal*, Vol. 28, No. 2, pp. 63–76 (2004).
[25] Burgoyne, J. A., Wild, J. and Fujinaga, I.: An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis, *ISMIR*, pp. 633–638 (2011).
[26] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D. and Ellis, D. P. W.: mir_eval: A transparent implementation of common MIR metrics, *ISMIR*, pp. 367–372 (2014).