2S-08

階層隠れセミマルコフモデルと深層学習に基づく楽曲セクションの境界推定とラベル付け

柴田 剛　　錦見 亮　　中村 栄太　　吉井 和佳

京都大学 大学院情報学研究科

1. はじめに

音楽構造解析とは，ポピュラー音楽の音響信号を A メロ・サビなどのセクションに分割しラベル付けするタスクであり，楽曲再生時のナビゲーションや試聴用音源の生成に有用である．本タスクは，音楽音響信号をセクションに分割する「セグメンテーション」，各セクションを楽曲の繰り返し構造に基づいてクラス分類する「クラスタリング」，各クラスに「サビ」などの具体的なラベルを付与する「ラベリング」の 3 ステップから構成される．

従来研究では，セグメンテーションとクラスタリングのみを扱うものが多く，非負値行列分解を用いる方法 [1] やグラフ分解を用いる方法 [2] が提案されている．ラベリングを扱う研究では，Maddage ら [3] は典型的な楽曲構造や各セクションの役割に基づく手法を提案している．また Paulus ら [4] は，セクションの N グラムに基づく適合度関数を構成し，セグメンテーションとクラスタリング，ラベリングを同時推定する手法を提案している．

本稿では，階層隠れセミマルコフモデル (HHSMM) に基づくセグメンテーションとクラスタリングを行う手法 [5] の結果と深層学習によるラベリングの結果を統合する手法を提案する．一般に，統合システムの構築では，多段処理的な方法より同時的に行う方法の方が，誤差が累積しない点で有用である．HHSMM をラベリングまで扱えるよう拡張する方法として，音声認識で用いられる DNN-HMM と同様の方式が考えられるが，第一段階として本稿では多段処理による統合を考える．具体的には，再帰型ニューラルネットワーク (RNN) を用いて音響特徴量系列からセクションラベル系列を推定した結果と，HHSMM の結果とを統合する手法を構築する．

2. 提案手法

音楽音響信号からセクションの境界位置とラベル（N_Z 種類あるものとする）を推定する問題に取り組む．図 1 に示すように，HHSMM によるセグメンテーションとクラスタリング（2.1 節）と RNN によるフレーム単位のラベリング（2.2 節）の結果を統合して各クラスターに対しラベルを付与する手法を提案する（2.3 節）．

2.1 セグメンテーションとクラスタリング

各セクション内における音色の同質性，同じクラスのセクションにおけるコード進行の反復性，およびセクション長の規則性を同時に扱える HHSMM を用いる [5]．HHSMM は，セクション系列とコード系列に対応する二階層の潜在変数系列を持つ．入力の音響特徴量系列に対して，ギブスサンプリングによりパラメータを推定した後，最尤の潜在変数系列をビタビアルゴリズムで推定することでセグメンテーションとクラスタリングを行う．

Boundary Detection and Labeling of Musical Sections Based on a Hierarchical Hidden Semi-Markov Model and Deep Learning: *Go Shibata, Ryo Nishikimi, Eita Nakamura, and Kazuyoshi Yoshii (Kyoto Univ.)*

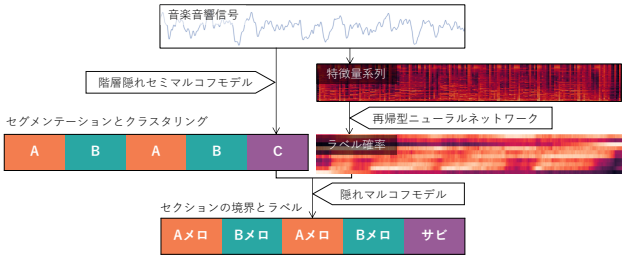

図 1: 提案する音楽構造解析手法

2.2 フレーム単位のラベリング

双方向長短期記憶 (BILSTM) 層，全結合層，softmax 出力層からなる RNN を用いる．この RNN は，入力音響特徴量系列 $\mathbf{x}_{1:T}$ に対して，セクションラベル系列 $z_{1:T}$ ($z_t \in \{1,\dots,N_Z\}$) の確率 $\{p(z_t|\mathbf{x}_t)\}_{t=1}^{T}$ を出力する（T は楽曲のフレーム数）．損失関数には focal loss を用いる．

2.3 統合

隠れマルコフモデル (HMM) を用いて，HHSMM および RNN に基づく解析結果を統合する．HHSMM で推定されたセクション数を N とすると，本節の HMM は，セクション単位のラベル系列 $z_{1:N}$ を潜在状態系列とし，フレーム単位の特徴量系列 $\mathbf{x}_{1:T}$ を生成する．

$$\begin{aligned} p(\mathbf{x}_{1:T}, z_{1:N}) &= p(z_{1:N})p(\mathbf{x}_{1:T}|z_{1:N}) \\ &= p(z_1)\prod_{n=1}^{N-1} p(z_{n+1}|z_n)\prod_{n=1}^{N}\prod_{t=1}^{T_n} p(\mathbf{x}_t|z_t=z_n) \end{aligned} \quad (1)$$

ここで，T_n は n 番目のセクションのフレーム数を表す（$\sum_n T_n = T$）．言語モデル $p(z_{1:N})$ として 1 次マルコフモデルを用いた場合，初期確率は $p(z_1)$，遷移確率は $p(z_{n+1}|z_n)$ $(n=1,\dots,N-2)$，終了確率は $p(z_N|z_{N-1})$ となる．一方，音響モデル $p(\mathbf{x}_{1:T}|z_{1:N})$ に関しては，RNN から得られた確率 $p(z_t|\mathbf{x}_t)$ に対し，ベイズの定理を適用することでフレーム単位の確率 $p(\mathbf{x}_t|z_t)$ を求める．

$$p(\mathbf{x}_t|z_t) = \frac{p(z_t|\mathbf{x}_t)p(\mathbf{x}_t)}{p(z_t)} \propto \frac{p(z_t|\mathbf{x}_t)}{p(z_t)} \quad (2)$$

ここで，確率 $p(z_t)$ はラベルの unigram 確率である．

いま，特徴量系列 $\mathbf{x}_{1:T}$ に対して，最大事後確率をとるセクション単位のラベル系列 $z_{1:N}$ を推定したい．

$$\begin{aligned} p(z_{1:N}|\mathbf{x}_{1:T}) &\propto p(\mathbf{x}_{1:T}, z_{1:N}) \\ &= p(z_1)\prod_{n=1}^{N-1} p(z_{n+1}|z_n)\prod_{n=1}^{N}\left(\prod_{t=1}^{T_n} p(\mathbf{x}_t|z_t=z_n)\right)^{\beta/T_n} \end{aligned} \quad (3)$$

ここで，β は言語モデルと音響モデルのバランスを調整するパラメータである．ラベル系列の全探索は計算量的に現実的ではないため，HHSMM で推定された各クラスターに含まれる全フレームについて，各ラベルごとに RNN で得られた確率の総積を取ることで，そのクラス

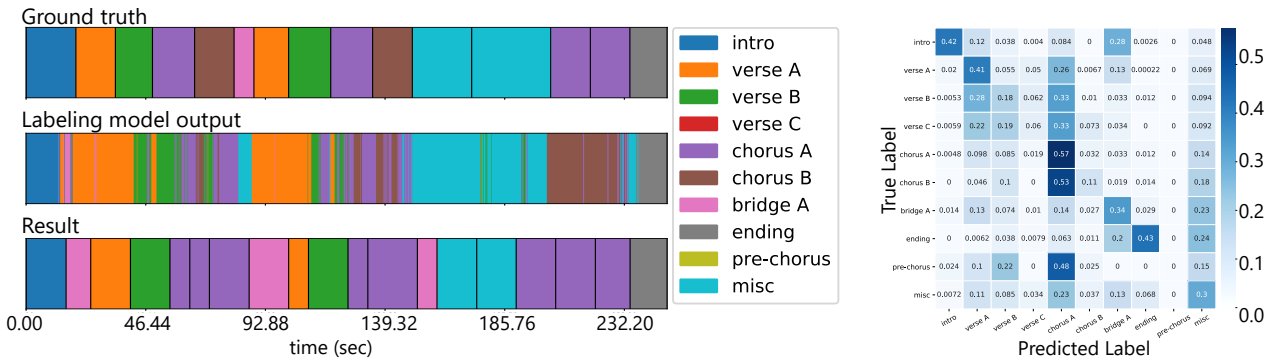

図 2: （左）結果の例（RWC-MDB-P-2001 No.42）（右）混同行列

ターが対応するラベル候補を α 個に絞り，全探索を行う．

3. 評価実験

3.1 実験条件

RWC ポピュラー音楽データベース中の 100 曲 [6] と，正解データ [7] を用いて 10 分割交差検証を行った．この際，ラベリング用 RNN の学習には 80 曲，検証に 10 曲を用いた．統合用 HMM の言語モデルには trigram を用い，学習には上記と同様の 90 曲を用いた．

RNN の入力には 128 次元のメルスペクトログラムを使用した．出力は，[4] の結果との比較のため，全体の 90% をカバーするラベル (verse A/B/C，chorus A/B，bridge A，intro，ending，pre-chorus) とその他を表すラベル (misc) の $N_Z = 10$ 個のラベルとした．BILSTM の隠れ層は 2048×2 次元とし，$\alpha = 3$，$\beta = 6.0$ とした．最適化には確率的勾配降下法（学習率 10^{-3}）を用いた．評価には学習時に検証データで最高精度を出したモデルを用いた．

フレーム単位でのラベル正解率を評価尺度として用いる．HMM を用いた HHSMM と RNN の統合手法の他，比較のため，RNN により得られたセクションラベルをクラスターごとにカウントし，もっともよく現れたラベルを推定結果とする統合手法（多数決法），および従来法 [4] を比較した．また，音響モデルと言語モデルの個別の効果を検証するため，RNN によるラベリング結果と HHSMM のクラスタリング結果を基に trigram 言語モデルにおける最大確率ラベル系列を求めた結果も評価した．

3.2 実験結果

実験結果を表 1 に示す．RNN の精度は比較的高いが，統合手法によりセクション単位に量子化を行う過程で精度が低下した．また，言語モデルのみを用いた結果は精度が低く，統合手法においても言語モデルによる精度向上は見られなかった．提案した統合手法は，従来法 [4] よりもわずかに精度が高かった．

推定結果例を図 2（左）に示す．intro と ending の位置や，verse A→verse B→chorus A の構造を正しく認識できている一方，境界位置の誤推定により部分的なラベルの誤推定が起きている．別例として，最後のセクションと同じクラスターに属するセクションが楽曲の中間に現れた場合，それらを全て ending としてしまう場合があった．このように，セグメンテーションとクラスタリングの結果からの誤差伝播の影響が大きく，言語モデルがうまく機能しなかった例が多く見られた．こうした誤差伝播は，セグメンテーションとクラスタリングとラベリングを同時推定する方法で避けられる可能性がある．

表 1: 評価結果

手法	精度 (%)
言語モデルあり統合手法 (HMM)	35.6
言語モデルなし統合手法（多数決法）	36.7
音響モデル (RNN) のみ	44.7
言語モデル (Trigram) のみ	17.8
Paulus 2009 [4]	34.4

また，ラベルの推定結果と正解データの混同行列を図 2（右）に示す．intro，ending，chorus A は比較的高い精度を出している．また verse A/B/C は verse A と，chorus A/B は chorus A と推定されやすい．しかし，chorus A や misc が他のラベルから推定されやすいことによる精度の低下が大きく，この問題の解消は今後の課題である．

4. おわりに

評価実験の結果，HHSMM と深層学習に基づいて楽曲セクションの境界推定とラベル付けを行う手法の有効性が確認された．今回，ラベリング手法に RNN を用いたが，ラベリングにはより広域な情報も重要であるため，楽曲全体を考慮できる Transformer でのモデルの改良が考えられる．また，本手法ではセグメンテーション・クラスタリングとフレーム単位のラベリングを独立に行ったが，後者の情報は前者の推定に有効だと考えられる．よって，今後はセグメンテーション，クラスタリング，ラベリング全てを同時に扱うモデルの考案を目指す．

謝辞 本研究の一部は，科研費 No. 19H04137，No. 19K20340，No. 16H01744 および JST ACCEL No. JPMJAC1602 の支援を受けた．

参考文献

[1] O. Nieto *et al.*: “Convex Non-negative Matrix Factorization for Automatic Music Structure Identification,” *ICASSP*, 236–240, 2013.

[2] B. McFee *et al.*: “Analyzing Song Structure with Spectral Clustering,” *ISMIR*, 405–410, 2014.

[3] N.C. Maddage *et al.*: “Content-based music structure analysis with applications to music semantics understanding,” *ACM MM*, 112–119, 2004.

[4] J. Paulus *et al.*: “Music structure analysis using a probabilistic fitness measure and a greedy search algorithm,” *TASLP*, vol.17, no.6, 1159–1170, 2009.

[5] G. Shibata *et al.*: “Statistical Music Structure Analysis Based on a Homogeneity-, Repetitiveness-, and Regularity-Aware Hierarchical Hidden Semi-Markov Model,” *ISMIR*, 268–275, 2019.

[6] M. Goto *et al.*: “RWC Music Database: Popular, Classical and Jazz Music Databases,” *ISMIR*, 287–288, 2002.

[7] M. Goto *et al.*: “AIST Annotation for the RWC Music Database,” *ISMIR*, 359–360, 2006.