4T-06

階層隠れマルコフモデルに基づく音楽音響信号に対する構造解析

柴田 剛 † 錦見 亮 ‡ 中村 栄太 ‡ 吉井 和佳 ‡

† 京都大学 工学部情報学科 ‡ 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

楽曲には，ポピュラー音楽でのサビやAメロのような意味上のまとまり（セクション）が存在する．音楽構造解析とは，(1) 音楽音響信号をセクション単位に分割し（セグメンテーション），(2) 楽曲の繰り返し構造に基づいてセクションをクラス分類し（クラスタリング），(3) 各クラスに「サビ」や「Aメロ」などのラベルを付ける（ラベリング）問題である．これは，サビの検出や楽曲の視聴用音源の生成に有用である．本研究ではこのうち，セグメンテーションとクラスタリングを目的とする．

楽曲のセクション構造は，「新規性」「同質性」「類似性」という3つの性質で典型的に特徴付けられる [1]．これらの性質は，二時刻間の特徴量類似度を各要素にもつ自己類似度行列で視覚化できる（図1）．新規性は，新規のセクションの開始時に直前のセクションからの特徴の変化があるという性質である．同質性は，セクション内部で楽器編成などの特徴が一貫しているという性質である．類似性は，同種のセクションで類似する特徴系列が繰り返されるという性質である．音楽構造解析では，これら3つの性質を統合的に用いることが有効である．

従来研究では，これらのうち単一の性質のみを取り扱うか，複数の性質を多段処理で扱う試みが多く [2, 3, 4]，統合的な手法はまだほとんどない [5]．本研究では，音楽音響信号の階層的生成モデルを構成してセクションの構造を表現することで，同質性と類似性を統合的・統計的に取り扱う音楽構造解析手法を提案する．具体的には，セクション単位の楽曲構造を表す上位モデルと、各セクションの内部構造を表す下位モデルからなる階層隠れセミマルコフモデル（HHSMM）の教師なし学習により，セグメンテーションとクラスタリングを一挙に行う．評価実験によってセグメンテーションが既知の場合にこの手法のクラスタリングに対する有効性を確認する．

2. 提案手法

HHSMM に基づく音楽構造解析手法について述べる．入力は音楽音響信号から得られたビート単位のクロマ特徴系列 $\boldsymbol{X}^c = \{\boldsymbol{x}_1^c, \ldots, \boldsymbol{x}_B^c\} \in \mathbb{R}^{B\times 12}$ と MFCC 特徴系列 $\boldsymbol{X}^m = \{\boldsymbol{x}_1^m, \ldots, \boldsymbol{x}_B^m\} \in \mathbb{R}^{B\times 12}$ であり，出力は各セクションの境界とクラスタリング結果である．ここで，B は4分音符単位のビート数を表す．これらの特徴量を使う理由は，類似性はクロマ特徴量で表されるコード進行に表されることが多く，同質性は MFCC によって表される音響特性により捉えられることが多いからである．

2.1 階層隠れセミマルコフモデル

提案モデルは遷移確率によりセクション単位の繰り返し構造を記述する上位モデルと，セクション内での特徴量系列の生成過程を表す下位モデルからなる（図2）．

上位モデル　上位モデルはセミマルコフモデルで，各状態はセクションのクラスとその継続時間長を表す．潜在

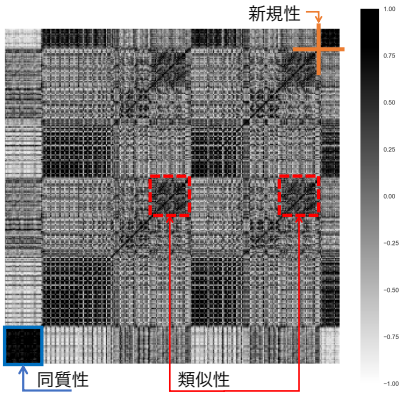

図 1: 自己類似度行列の例

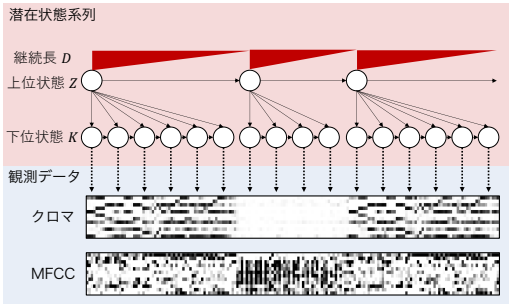

図 2: 提案する階層隠れセミマルコフモデル（HHSMM）

状態系列を $\boldsymbol{Z} = \{z_1, \ldots, z_T\}$ $(z_\tau \in \{1, \ldots, N_Z\})$，$\boldsymbol{D} = \{d_1, \ldots, d_T\}$ $(d_\tau \in \{1, \ldots, N_D\})$ と記す．ここで，T は楽曲中のセクション数，N_Z はセクションのクラス数，N_D は最大継続長を表す．継続長時間 d_τ および状態 z_τ は以下の確率に従い生成されるものとする．

$$P(z_1, d_1) = \rho_{z_1}\psi_{d_1} \tag{1}$$

$$P(z_\tau, d_\tau | z_{\tau-1}) = \pi_{z_{\tau-1}z_\tau}\psi_{d_\tau} \tag{2}$$

ここで，ρ_z は状態 z の初期確率，$\pi_{zz'}$ は状態 z から状態 z' への遷移確率，ψ_d は継続長確率を表す．

下位モデル　下位モデルは状態種類数 N_K の left-to-right（LR）型の隠れマルコフモデルであり，各上位状態の開始時刻から継続長時間が経過するまでビート単位で遷移を繰り返す．潜在状態系列 $\boldsymbol{K}_\tau = \{k_1, \ldots, k_{d_\tau}\}$ $(k_t \in \{1, \ldots, N_K\})$ は，対応する上位状態のクラス z_τ に依存する遷移確率 $\phi^{(z_\tau)}_{k_{t-1}k_t}$ により生成される．

$$P(k_t | z_\tau, k_{t-1}) = \phi^{(z_\tau)}_{k_{t-1}k_t} \tag{3}$$

ここで，z_τ と d_τ はそれぞれ対応する上位状態と継続長時間を表す．初期状態 k_1=1，終了状態 k_{d_τ}=N_K を満たし，時刻 t_1, t_2 $(t_1$<$t_2)$ において，$k_{t_1} \leq k_{t_2}$ を満たす．

音響モデル　モデルは各時刻 b でクロマ特徴量 $\boldsymbol{x}_b^c \in \mathbb{R}^{12}$ と MFCC 特徴量 $\boldsymbol{x}_b^m \in \mathbb{R}^{12}$ を出力する．コード進行の繰り返し構造を捉えるためにクロマ出力確率分布 $\chi^c_{z_\tau,k_t}(\boldsymbol{x}_b^c)$ は上位状態と下位状態両方に依存させ，音響特性のセク

Music Structure Analysis Based on a Hierarchical Hidden Markov Model: *Go Shibata, Ryo Nishikimi, Eita Nakamura, and Kazuyoshi Yoshii (Kyoto Univ.)*

ション内での同質性を捉えるため MFCC 出力確率分布 $\chi^m_{z_\tau}(\boldsymbol{x}^m_b)$ は上位状態のみに依存させる．

$$P(\boldsymbol{x}^c_b, \boldsymbol{x}^m_b) = \chi^c_{z_\tau,k_t}(\boldsymbol{x}^c_b)\chi^m_{z_\tau}(\boldsymbol{x}^m_b) \quad (4)$$

ここで，z_τ と k_t は時刻 b における上位状態と下位状態を表し，各出力確率分布は多次元正規分布を用いる．

2.2 推論

本手法では，モデルパラメータ推定と潜在状態推定の二段階の統計的推論によって音楽構造解析を行う．パラメータ推定では，$\rho_z, \pi_{zz'}, \psi_d, \phi^{(z)}_{kk'}$ と出力確率分布の平均・分散パラメータに事前分布を置き，ベイズ推論に基づく教師なし学習を行う．ベイズ学習によって，状態数を十分大きく設定すれば，不要な状態が縮退し，最適な状態数のモデルが選択される．また，上位モデルの遷移確率をスパースに誘導することで，繰り返し構造が表現できる．推論方法としてギブスサンプリングを用いた．

潜在状態推定では，入力のクロマ特徴系列 $\boldsymbol{X}^c$ と MFCC 特徴系列 $\boldsymbol{X}^m$ に対して，ビタビ探索で最尤の潜在状態系列 $\boldsymbol{Z}$，$\boldsymbol{D}$，$\{\boldsymbol{K}_1, \ldots, \boldsymbol{K}_T\}$ を得る．これにより上位状態の遷移に現れるセクション分割と上位状態の値により表されるクラスタリング結果が一挙に得られる．

3. 評価実験

ポピュラー音楽を用いた実験で，提案法の有用性を検証する．本稿ではセグメンテーションは既知とした上でのクラスタリングに対する評価を行う．一般にセクションの長さは一定ではなく，さらにアノテータによって異なる解釈が生じやすい．そこでここでは，セクションより短いまとまりであるフレーズに注目する．フレーズはポピュラー音楽では 4 小節程度の長さが一般的であり，解釈の曖昧性も生じにくい．簡単のため本実験ではフレーズの長さを 4 小節に固定し，そのクラスタリングを行う．RWC 音楽データベース [6] の中で曲を通して 4 小節単位で区切ることのできる楽曲 3 曲を評価に用いる．事前分布のハイパーパラメータについて，上位状態の初期確率，継続長確率，遷移確率，下位状態の遷移確率に対応するハイパーパラメータはそれぞれ $\mathbf{0.1}$，$\boldsymbol{b}$，$\mathbf{1}$，$\mathbf{1}$，クロマと MFCC の出力分布の平均，精度行列の分布のハイパーパラメータ β^c，$\boldsymbol{m}^c$，ν^c，$\boldsymbol{W}^c$，β^m，$\boldsymbol{m}^m$，ν^m，$\boldsymbol{W}^m$ はそれぞれ 10，$E[\boldsymbol{X}^c]$，24，$(\nu^c Cov[\boldsymbol{X}^c])^{-1}$，10，$E[\boldsymbol{X}^m]$，60，$(\nu^m Cov[\boldsymbol{X}^m])^{-1}$ を用いた．ここで，$\mathbf{0.1}$ と $\mathbf{1}$ はそれぞれ全ての要素が 0.1 と 1 のベクトル，$\boldsymbol{b} \in \mathbb{R}^{16}$ は，第 16 要素が 1 の単位ベクトル，$E[\cdot]$ と $Cov[\cdot]$ はそれぞれデータの平均と共分散である．また，上位状態数と下位状態数はそれぞれ 8 と 6 を用いた．

同じクラスのセクションのフレーズ数は等しいという仮定の下で，次の 2 つの評価尺度を用いる．

フレーズ一致率　正解データで同じクラスであるセクション s と s' について，推定されたフレーズをそれぞれ $\{z_1, \ldots, z_n\}$ と $\{z'_1, \ldots, z'_n\}$ とする．この時各 z_τ と z'_τ が一致する割合を次で定義する．

$$E^{\text{一致}} = \frac{\sum_{s\neq s'}\sum_\tau^n \delta_{z_\tau z'_\tau}\delta_{\mathrm{cls}(s)\mathrm{cls}(s')}}{\sum_{s\neq s'}\sum_\tau^n \delta_{\mathrm{cls}(s)\mathrm{cls}(s')}} \quad (5)$$

ここで，$\mathrm{cls}(s)$ はセクション s のクラスを表す．

フレーズ分離度　正解データで異なるクラスであるセクション s と s' について，推定されたフレーズをそれぞれ $\{z_1, \ldots, z_n\}$ と $\{z'_1, \ldots, z'_{n'}\}$ とする時，各 z'_τ が $\{z_1, \ldots, z_n\}$ の全てと異なる場合の割合を次で定義する．

表 1: 評価結果

手法	$E^{\text{一致}}$	$E^{\text{分離}}$	E^{HM}
HHSMM	0.632	0.743	0.658
MFCC のみ	0.677	0.729	0.671
クロマのみ	0.862	0.526	0.620

図 3: 正解セクションデータとフレーズ推定結果の例

$$E^{\text{分離}} = \frac{\sum_{s\neq s'}\sum_\tau^{n'}(1-\delta_{\mathrm{cls}(s)\mathrm{cls}(s')})\prod_v^n(1-\delta_{z_v z'_\tau})}{\sum_{s\neq s'}\sum_\tau^{n'}(1-\delta_{\mathrm{cls}(s)\mathrm{cls}(s')})} \quad (6)$$

また，$E^{\text{一致}}$ と $E^{\text{分離}}$ の調和平均を E^{HM} と記す．

表 1 に 3 つの楽曲に対する評価値の平均を示す．提案手法の結果の他，MFCC およびクロマ特徴量のみを用いる場合の結果も示す．正解データとしてセクションアノテーションデータ [7] を用いた．類似性と同質性を考慮した提案手法によって，比較的高い精度でクラスタリングができることが確認できる．セグメンテーション既知の条件では，特徴量として MFCC のみ用いる場合が最も精度が高かったが，可変なセクションのセグメンテーションも行う場合にはクロマ特徴量も用いるのが有効だと考えられる．またより多くのデータに対して評価実験を行う必要がある．また，図 3 に本手法を用いた推定結果の一例を示す．最上図が正解セクション系列，下図 3 つが推定フレーズ系列を示している．推定結果は尤度が最大のものを選択した．これにより提案法によって繰り返し構造が捉えられていることが確認できた．

4. おわりに

本稿では，階層隠れマルコフモデルを用いた音楽音響信号に対する統計的構造解析手法を提案した．評価実験の結果，本手法が音楽構造解析に有効であることが確認された．今後は，長さの制約を取り除きより一般的な楽曲での実験を行うとともに，セクションとフレーズを別の階層で表現するモデルへの拡張に取り組む．さらに，現在考慮できていない「新規性」を含む，3 つの性質全てを統合的に扱うモデルへの拡張を目指す．

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

参考文献

[1] J. Paulus *et al.*: “Audio-based music structure analysis,” *ISMIR*, 625–636, 2010.

[2] J. Foote : “Automatic audio segmentation using a measure of audio novelty,” *ICME*, 452–455, 2000.

[3] M. Cooper *et al.*: “Summarizing popular music via structural similarity analysis,” *WASPAA*, 127–130, 2003.

[4] M. Goto : “A chorus section detection method for musical audio signals and its application to a music listening station,” *ASLP*, vol.14, no.5, pp.1783–1794, 2006.

[5] B. McFee *et al.*: “Analyzing song structure with spectral clustering,” *ISMIR*, 405–410, 2014.

[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.