GTTMに基づくメロディ音符列の確率的木構造モデル

Tree-Structured Probabilistic Model of Melodic Musical Notes Based on GTTM

中村栄太[*1] 浜中 雅俊[*1] 平田 圭二[*2] 吉井 和佳[*1]
Eita Nakamura Masatoshi Hamanaka Keiji Hirata Kazuyoshi Yoshii

[*1]京都大学 Kyoto University [*2]はこだて未来大学 Future University Hakodate

This paper presents a probabilistic formulation of music language modelling based on the generative theory of tonal music (GTTM). GTTM is a well-known music theory that describes the tree structure of written music in analogy with the phrase structure grammar of natural language. To develop a computational music language model incorporating GTTM and a machine-learning framework for data-driven music grammar induction, we construct a generative model of monophonic music based on probabilistic context-free grammar, in which the time-span tree proposed in GTTM corresponds to the parse tree. We derive supervised and unsupervised learning algorithms based on the maximal-likelihood estimation, and a Bayesian inference algorithm based on the Gibbs sampling. We found that the model automatically acquires music grammar from data and reproduces time-span trees of written music as accurately as an automatic analyser that required elaborate manual parameter tuning.

1. はじめに

音楽情報処理の中心的課題の一つに，自動採譜がある．これを目的として音楽音響モデルに関する研究が広く行われているが [1]，音響的変動による採譜の曖昧性を完全に取り除くことは難しく，楽譜に対する事前知識を用いる必要がある．音声認識と同様に，従来はマルコフモデルなどに基づく音楽言語モデルが研究されてきたが，音楽理論の知識が十分に組込まれているとは言えない．そこで本研究では，音楽理論を組込んだ楽譜の文法を記述するモデルを構築し，その学習手法を開発する．

音楽理論 GTTM (Generative Theory of Tonal Music) [2] は，音楽の階層構造を記述するモデルである．即ち，音符は，動機・小楽節・大楽節・セクションといったより大きな構造にグループ化され，一部の音符は周りの音符よりも目立って聴かれるという楽曲の構造のモデル化を試みている．GTTM では，タイムスパン木と呼ばれる木構造により各音符の相対的な重要度が記述される．タイムスパン木は，音符列の簡約化の手順を表すものとして解釈できる．タイムスパン木の導出には，和声 [3] やシェンカー分析 [4] などの音楽知識が必要とされるが，GTTM はそれに必要な規則の提案も行っている．

GTTM の規則を計算論的に定式化する研究がこれまで行われてきた [5–8]．文献 [5] では GTTM の規則をパラメータ化し，ATTA と呼ばれるタイムスパン木の自動決定手法の導出を行っているが，46 個のパラメータの調整が課題として残っている．最近，σGTTM III と呼ばれる確率モデルに基づくタイムスパン木の分析法が提案されており，パラメータをラベル付きデータから学習することが可能となった [8]．この手法はグルーピング分析など他の分析結果 [5,6] を入力として必要とするが，さらに拡張して，生成モデルとして定式化することによりスタンドアローンで動作するタイムスパン木分析器の構築および自動採譜などへの応用可能な言語モデルの構築が可能となるであろう．また，近年自然言語処理で発達している機械学習手法を取り入れることにより，教師なし文法学習などの手法の開発が可能となっている [9–12]．

連絡先: 中村栄太，京都大学，606-8501 京都府京都市吉田本町，enakamura@sap.ist.i.kyoto-u.ac.jp

そこで，本研究では GTTM に基づく確率的生成モデルの定式化を行う．タイムスパン木を自然言語の構文木と同様に，音符の生成過程を表すものと解釈することにより，確率文脈自由文法 (probabilistic context-free grammar; PCFG) [13,14] に基づくモデルを構築する．PCFG はこれまで，σGTTM III の他に，多重音解析 [15] や和声解析 [16] などで音楽に応用されており，文献 [15] では教師なし学習手法も議論されている．本研究では，単純化のため単旋律音楽に限って議論する．

2. 提案法

2.1 GTTM のタイムスパン木

タイムスパン木は，楽譜内の各音符の相対的な重要度を記述する二分木であり，木のリーフからルートに向かって，楽譜を簡略化する過程を表すものと解釈される [2]．簡略化の過程では，2 つの隣り合う音符（子ノードに対応）が一つの音符（親ノードに対応）にまとめられる．この際，隣り合う音符間の主従関係により，いずれかの音符の音高が簡略化された音符の音高として使われる．また，簡略化の性質から，子ノードの音価の和は親ノードの音価に一致する必要がある．

2.2 基本モデル

以下，音符を音高 p と音価 r のペアにより表し，音符列 $(p_n, r_n)_{n=1}^N$ として楽譜を表す．ただし，休符を表す「R」も音高の集合（Ω_p と記す）に含まれるものとし，音価は全音符からの相対長で表現する（例：四分音符は $r = 1/4$）．

PCFG モデル [13] は，終端記号の集合 Ω_T，非終端記号の集合 Ω_N（開始記号 S を含むものとする），そして生成規則の集合で定義される．チョムスキー標準形では，生成規則は $A \to \alpha$ ($A \in \Omega_N$, $\alpha \in \Omega_N \times \Omega_N \cup \Omega_T$) の形で表されれ，確率値 $P(A \to \alpha)$ が付与される．以下，記号 A のことを親，α の中の記号を子と呼ぶ．与えられた非終端記号の列 $\boldsymbol{w} = w_1 \cdots w_N$ ($w_n \in \Omega_T$) に対して，それを導出する一連の生成規則の集まりは木構造で表され，$\boldsymbol{w}$ の導出木と呼ばれる．

GTTM の確率的式化のために，通常の PCFG に対して次の変更と拡張が必要となる．まず，タイムスパン木では各ノードはリーフと同様に音符によって表されるため，非終端記号は終端記号と同じく音符により表される．(開始記号についての例

外については後述する.) 次に，タイムスパン木により表される音符の主従関係を記述するため，L と R の 2 値をとる確率変数 s を導入する．生成規則は $(p,r) \to s(p_L,r_L)(p_R,r_R)$ （ただし，$p,p_L,p_R \in \Omega_p$，$r,r_L,r_R \in \Omega_r$）の形で表され，(p_s,r_s) が主となる音符を表す（$p=p_s$）．音価の和に関する制約条件は $r=r_L+r_R$ と表現される．

以上に基づき音符の生成モデルを定式化する．生成過程は音価 r_S を持つ開始記号 S から始まる．生成規則は $(S,r_S) \to s(p_L,r_L)(p_R,r_S-r_L)$ 又は $(p,r) \to s(p_L,r_L)(p_R,r-r_L)$ の形を持つ（ただし，$p,p_L,p_R \in \Omega_p$，$r,r_L \in \Omega_r$，$s=L,R$）．生成確率の規格化条件は以下の通り．

$$\sum_{s,p_L,p_R,r_L} P\big((S,r_S) \to s(p_L,r_L)(p_R,r_S-r_L)\big) = 1, \quad (1)$$

$$\sum_{s,p_L,p_R,r_L} P\big((p,r) \to s(p_L,r_L)(p_R,r-r_L)\big) = 1. \quad (2)$$

なお，休符は親ノードには現れないものとする．

生成確率は調に依存すると考えられるが，この状況は 12 の長調と 12 の短調それぞれに対して個別の生成モデルを作り，これらの混合モデルを考えることで記述できる．本稿では，状況を単純化し，調は予め与えられた状況を考える．即ち，全ての楽譜はハ長調またはハ短調に移調されているものとする．

2.3 モデルの単純化と改良

現実的に扱える計算量のモデルとするため以下の単純化をする．まず．生成確率は音高と音価に関して独立であり，次のようにファクトライズされるものとする．

$$P\big((p,r) \to s(p_L,r_L)(p_R,r-r_L)\big) = P^s(s)P^p(p \to sp_Lp_R)P^r(r \to sr_L(r-r_L)). \quad (3)$$

ここで，$\sum_s P^s(s)=1$，$\sum_{p_L,p_R} P^p(p \to sp_Lp_R)=1$，$\sum_{r_L} P^r(r \to sr_L(r-r_L))=1$ を満たす（S に対する生成規則も同様）．次に，音高はオクターブを無視し，12 のピッチクラスで表されるものとする．即ち，$\Omega_p=\{\mathrm{C},\mathrm{C\#},\cdots,\mathrm{B},\mathrm{R}\}$ (ただし，C#=D♭ など)．各確率は次のように表記するものとする：$\phi_s=P^s(s)$，$\theta_{sp_Lp_R}=P^p(p \to sp_Lp_R)$，$\rho_{sr_L}=P^r(r \to sr_L(r-r_L))$，$\Theta=(\phi_s,\theta_{sp_Lp_R},\rho_{sr_L})_{s,p,p_L,p_R,r,r_L}$．

GTTM の規則 TSRPR 1 では，音符の主従関係は拍重みの大小関係に依存すると記されている [2]．ここで拍重みとは拍位置の相対的な強さを表すものである [17]．この規則の効果を取り込むため，確率 ϕ_s は子ノードに対応する音符の拍重みに依存するものとする．具体的には，音符 (p_L,r_L) と (p_R,r_R) の拍重みを ω_L と ω_R 表す時，ω_L/ω_R が 1 か，1 より大きいか小さいかにより，ϕ_s が異なる値をとり得るものとする．

2.4 ベイズ拡張

言語処理分野では，PCFG などの確率モデルの推論にベイズ推定手法が近年多く用いられている [9]．ベイズ拡張は，モデルの確率パラメータに事前分布を与えることにより行える．生成規則確率は離散分布に従うため，事前分布には共役であるディリクレ分布を用いられる．確率パラメータを $\boldsymbol{\eta}=(\eta_s)_s$，$\boldsymbol{\lambda}_{sp}=(\lambda_{sp_Lp_R})_{p_L,p_R}$，$\boldsymbol{\nu}_{sr}=(\nu_{sr_L})_{r_L}$ と記し，対応するディリクレパラメータを $\boldsymbol{\phi}=(\phi_s)_s$，$\boldsymbol{\theta}_{sp}=(\theta_{sp_Lp_R})_{p_L,p_R}$，$\boldsymbol{\rho}_{sr}=(\rho_{sr_L})_{r_L}$ と記すことにする．ディリクレ分布を P_{Dir} を表すと，事前分布は $P_{\mathrm{Dir}}(\boldsymbol{\phi}|\boldsymbol{\eta})$ などと表される．

2.5 推論

動的計画法に基づく PCFG の効率的な推論アルゴリズムが提案されており，これらを用いて提案法に対する推論アルゴリズムを導出することができる．まず，与えられた音符列 $W=(p_n,r_n)_{n=1}^N$ に対する，確率最大のタイムスパン木は CYK-Viterbi アルゴリズム [13] により求まる．

最尤法に基づくパラメータ学習には，EM アルゴリズムに基づく反復法 [18] を用いられる．更新式は次の通り．

$$\phi_s^{\mathrm{new}} \propto \sum_{p,p_L,p_R,r,r_L} \mathscr{C}\big((p,r) \to s(p_L,p_R,r_L);\Theta^{\mathrm{old}}\big),$$

$$\theta_{sp_Lp_R}^{\mathrm{new}} \propto \sum_{r,r_L} \mathscr{C}\big((p,r) \to s(p_L,p_R,r_L);\Theta^{\mathrm{old}}\big),$$

$$\rho_{sr_L}^{\mathrm{new}} \propto \sum_{p,p_L,p_R} \mathscr{C}\big((p,r) \to s(p_L,p_R,r_L);\Theta^{\mathrm{old}}\big).$$

ここで，

$$\mathscr{C}\big((p,r) \to s(p_L,p_R,r_L);\Theta^{\mathrm{old}}\big) = \sum_{T\in\mathbb{T}_W} \frac{P(T|\Theta^{\mathrm{old}})c((p,r) \to s(p_L,p_R,r_L);T)}{P(W|\Theta^{\mathrm{old}})} \quad (4)$$

は導出木の確率重み付きの生成規則の出現数を表し，内側変数 $\beta_{nmp}(W)$ と外側変数 $\alpha_{nmp}(W)$ を用いて次の式で与えられる．

$$\phi_s^{\mathrm{old}}\theta_{sp_Lp_R}^{\mathrm{old}}\rho_{sr_L}^{\mathrm{old}} \sum_{n=1}^{N}\sum_{m=n+1}^{N}\sum_{k=n}^{m-1} \delta_{r_n^m,r}\delta_{r_n^k,r_L}\delta_{r_{k+1}^m,r-r_L} \cdot \alpha_{nmp}^{\mathrm{old}}(W)\beta_{nkp_L}^{\mathrm{old}}(W)\beta_{(k+1)mp_R}^{\mathrm{old}}(W). \quad (5)$$

ただし，$W_n=(p_n,r_n)$，$W_n^m=W_n\cdots W_m$，$r_n^m=r_n+\cdots+r_m$ であり，β^{old} と α^{old} は Θ^{old} を用いて計算した内側変数と外側変数を表す．内側変数と外側変数の計算アルゴリズムは文献 [13] を参照されたい．

ベイズ推定は，ギブスサンプリングに基づく手法 [9] を用いて行える．この手法では，ハイパーパラメータ $\Lambda=(\boldsymbol{\eta},\boldsymbol{\lambda}_{sp},\boldsymbol{\nu}_{sr})$ とタイムスパン木 T を $P(\Theta|T,W,\Lambda)$ と $P(T|\Theta,W,\Lambda)$ からサンプルすることで推論を行う．前者の事後分布はディリクレ分布となり，これからのサンプリングを行う．後者のタイムスパン木のサンプリングは，各ノードに関する逐次サンプリングにより行える．各ノードはペア (p,r_n^m) $(1\le n\le m\le N,\ p\in\Omega_p)$ により表される（p はタイムスパン r_n^m を表す音高とする）．ルートノード (S,r_1^N) から以下の分布に基づいてノードを展開することによりリーフノードまでサンプルできる．

$$P\big((p,r_n^m) \to s(p_L,r_n^{k-1})(p_R,r_k^m)|\Theta,W\big), \propto \phi_s\theta_{sp_Lp_R}\rho_{s(r_n^m)(r_n^{k-1})}\frac{\beta_{n(k-1)p_L}\beta_{kmp_R}}{\beta_{nmp}}. \quad (6)$$

3. 評価

提案法を，専門家による 300 曲のタイムスパン木解析のデータベース [7] を用いて評価した．モデルの学習能力を比較するため，オープン及びクローズドの教師あり学習，EM アルゴリズム及びギブスサンプリングに基づく教師なし学習の 4 つの学習条件の比較評価を行った．推定結果に対して，タイムスパン木の全てのノード中で親及び子ノードが完全に一致するものの割合を計算することで正解率とした．

評価結果を表 1 に示す．比較のため，ATTA [5] 及び σGTTM III [8] による結果も示す．ただし，σGTTM III は正解データのグループ境界を用いているため，他の手法との平等な比較はできない．教師なし学習の結果は，ATTA と同等であった．オープンとクローズドで値が同等であることにより，データ量の増加によるさらなる正解率の向上は少ないと考えられる．教師なし学習の結果は，EM アルゴリズムとギブスサンプリング

表 1: タイムスパン木解析の精度

学習条件	正解率 (%)
教師あり (オープン)	44.1
教師あり (クローズド)	44.9
教師なし (EM)	32.3
教師なし (ギブス)	31.4
ATTA	44
σGTTM III	76

表 2: 誤り解析の結果

高さ	ノード数	正解率 (%)	マッチした子の数 (%)
1	4237	60.8	80.4
2	2548	43.6	52.7
3	1578	35.0	45.8
4	998	22.8	35.6
5	606	13.2	21.9
6	358	3.9	8.9
7≥	253	3.6	9.1

で同等であり，教師あり学習の結果よりも正解率が低かった．

誤り解析の結果を表 2 に示す．ここで，タイムスパン木のノードに対する高さをその子孫のリーフからの最大距離として定義し，ノードの高さごとに正解率を示した（おオープンの教師あり学習の結果を示すが，その他の結果も同様であった）．また，子ノードは正解と一致しているが，親ノードは必ずしも一致していないノードの割合も併記している．この値と正解率との差異は，親ノードが正しく推定されなかったノードの割合を表している．結果より，高さが小さいノードほど高い正解率となっていることが分かる．同様の傾向は，タイムスパン木の推定結果の一例（図 1）においても確認できる．この例では楽節の最も重要な音である終止音が正しく選択されておらず，同様の誤りは推定結果において典型的であった．

手動でのパラメータ調節が必要であった ATTA と同等の精度で動作したことから，提案モデルの有効性が確認できた．しかし，結果から精度向上のためには改良が必要であることも示唆された．まず，言語処理の場合と同様に [11]，生成確率はコンテキストに強く依存するため，連続した数音符にまたがる依存性を取り入れる必要があるであろう．また文法を記述する記号の選択も重要であろう．今回は非終端記号を実質的に用いていないため，生成規則はタイムスパン木の中の全ての位置と高さにおいて同じであった．高さの小さなノードでは，通常拍重みが大きい程より音符の重要度が高くなるが，終止音はしばしば弱拍に現れることからも，潜在的な文法カテゴリーに対応する記号の導入の必要性が示唆される．この拡張には言語処理で用いられる symbol refinement が有効な可能性がある [14,19].

4. 結論

音楽理論 GTTM に基づき，単旋律音符列の確率的木構造モデルを構成した．PCFG の拡張モデルにより定式化を行い，教師なし及び教師あり学習手法の導出を行った．データからの学習により求めたパラメータを用いた提案モデルでのタイムスパン木の自動解析により，従来のルールベースの手法と同等の精度を達成した．今後，多声部音楽へのモデル拡張の他，自動採譜や編曲への応用を行う予定である．

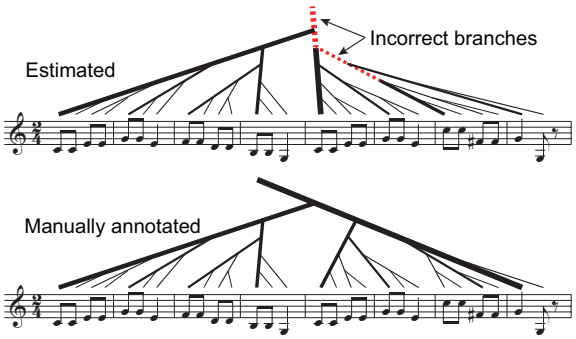

図 1: タイムスパン木の解析結果の例

参考文献

[1] A. Klapuri and M. Davy (eds.), *Signal Processing Methods for Music Transcription*, Springer, 2006.

[2] F. Lerdahl and R. Jackendoff, *A Generative Theory of Tonal Music*, MIT Press, Cambridge, 1983.

[3] S. Kostka *et al.*, *Tonal harmony* (7th ed.), McGraw-Hill, New York, 2004.

[4] A. Cadwallader and D. Gagné, *Analysis of Tonal Music: A Schenkerian Approach* (3rd ed.), Oxford University Press, 2011.

[5] M. Hamanaka *et al.*, "Implementing 'A Generative Theory of Tonal Music'," *Journal of New Music Research*, vol. 35 no. 4, pp. 249–277, 2006.

[6] Y. Miura *et al.*, "Use of Decision Tree to Detect GTTM Group Boundaries," *Proc. ICMC*, pp. 125–128, 2009.

[7] M. Hamanaka *et al.*, "Music Structural Analysis Database based on GTTM," *Proc. ISMIR*, pp. 325–330, 2014.

[8] M. Hamanaka *et al.*, "σGTTM III: Learning Based Time-Span Tree Generator Based on PCFG," *Proc. CMMR*, pp. 303–317, 2015.

[9] M. Johnson *et al.*, "Bayesian Inference for PCFGs via Markov Chain Monte Carlo," *Proc. HLT-NAACL*, pp. 139–146, 2007.

[10] M. Post and D. Gildea, "Bayesian Learning of a Tree Substitution Grammar," *Proc. ACL-IJCNLP*, pp. 45–48, 2009.

[11] T. Cohn *et al.*, "Inducing Tree-Substitution Grammars," *Journal of Machine Learning Research*, vol. 11, pp. 3053–3096, 2010.

[12] H. Shindo *et al.*, "Bayesian Symbol-Refined Tree Substitution Grammars for Syntactic Parsing," *Proc. ACL*, pp. 440–448, 2012.

[13] C. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

[14] M. Johnson, "PCFG Models of Linguistic Tree Representations," *Computational Linguistics*, **24**, pp. 613–632, 1998.

[15] M. Nakano *et al.*, "Bayesian Nonparametric Music Parser," *Proc. ICASSP*, pp. 461–464, 2012.

[16] W. Granroth and M. Steedman, "Statistical Parsing for Harmonic Analysis of Jazz Chord Sequences," *Proc. ICMC*, pp. 478–485, 2012.

[17] D. Temperley, *Music and Probability*, MIT Press, 2006.

[18] R. Neal and G. Hinton, "A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants," in *Learning in Graphical Models*, Springer Netherlands, pp. 355–368, 1998.

[19] T. Matsuzaki *et al.*, "Probabilistic CFG with Latent Annotations," *Proc. ACL*, pp. 75–82, 2005.