2章 音楽と統計的記号処理

吉井和佳†

キーワード：音楽情報処理，統計的記号処理，機械学習

1．まえがき

人間と同じように音楽を理解し，創作を行うことができる計算機を実現するためには，音楽の記号的な側面に着目した研究が必要不可欠である．われわれが日常生活で音楽と触れあう機会のうち，最も主要なものは，CDを購入したり，インターネットからダウンロードするなどして音楽を聴くことであり，音楽は音響信号としての形態をとっている．したがって，音楽信号処理技術を改良することが，音楽の理解につながると考えがちである．しかし，本来，音楽（本稿では，調性を持つ標準的な西洋音楽を想定している）は，楽譜，すなわち離散記号である音符の集合で表現され，音響信号なしで存在しうることには注意せねばならない．このことは，言語を理解しようと思うと，音声信号だけではなく，テキスト，すなわち単語や文字の系列を解析することが必須となるのと同じである．音楽も言語もその起源は同じであると言われており，その音響的な側面と，記号的な側面とを区別して取り扱うことが重要である．

音楽情報処理分野では，統計的機械学習技術を導入することで，音楽信号処理技術は目覚ましい発展を遂げているが（別記事「音楽と統計的信号処理」を参照），最近ようやく，楽譜情報解析技術も同様に発展の兆しを見せ始めている．本稿では，音符系列や和音系列など，楽譜上に記載されている離散記号に対する確率モデルについて紹介する．音楽と言語とのアナロジーから，この種の確率モデルは言語モデルと呼ばれている．言語モデルは，音楽的に妥当な音符の配置や和音系列に対しては高い確率を，音楽的に不自然なものに対しては低い確率を与える．すなわち，言語モデルは，楽譜の背後に存在する何らかの文法規則・現象を捉えており，楽曲の作曲・編曲支援に応用したり，音響信号から自動採譜を行う際の制約に用いるなど応用範囲が極めて広い．

† 京都大学 大学院情報学研究科
"Music and Statistical Symbolic Processing" by Kazuyoshi Yoshii (Graduate School of Informatics, Kyoto University, Kyoto)

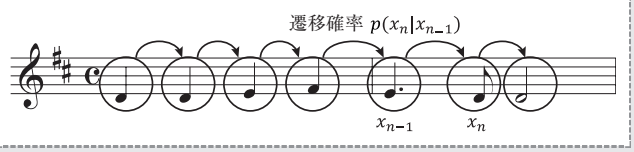

図1　音符マルコフモデル

2．音符系列に対する確率モデル

まず，メロディ（単旋律），すなわち単一の音符系列に対する言語モデルについて，基本的なものから最近のものまで紹介する．解析対象のデータを音符系列$x_{1:N}=x_1\cdots x_N$（Nは音符数）とする．x_nは，n番目の音符の音高あるいは音価，あるいはそれらのペアを表す．ここで，音価とは，楽譜に記される音長の全音符に対する相対的な値として定義する（例えば，四分音符は$x=1/4$，付点二分音符は$x=3/4$など）．また，音高の種類数はK_P，音価の種類数はK_Vであるとする．

2.1　音符マルコフモデル

初期の研究では，自然言語処理分野で一般的な単語系列に対するマルコフモデル（n-gramモデル）に着想を得て，音符系列に対するマルコフモデルが提案されている[1]（図1）．一次のマルコフモデル（バイグラムモデル）では，$x_{1:N}$が生成される確率は，

$$p(x_{1:N})=p(x_1)\prod_{n=2}^{N}p(x_n|x_{n-1}) \tag{1}$$

で与えられる．ここで，$p(x_1)$は初期確率，$p(x_n|x_{n-1})$はx_{n-1}からx_nへの遷移確率を表す．$p(x_n|x_{n-1})$を$p(x_n|x_{n-1:n-P})$に置き換えれば，P次のマルコフモデルが定義できるが，学習データが充分にないとパラメータ（初期確率と遷移確率）の推定が難しくなる．そのため，実用上は$P=1$がよく用いられるが，音高と音価を同時に考える場合は，パラメータ数は$K_PK_V+K_P^2K_V^2$と依然として多い．そのため，音高と音価はそれぞれ独立したマルコフモデルに従うと仮定し，全体のパラメータ数を$K_P+K_V+K_P^2+K_V^2$に制限することも多い．

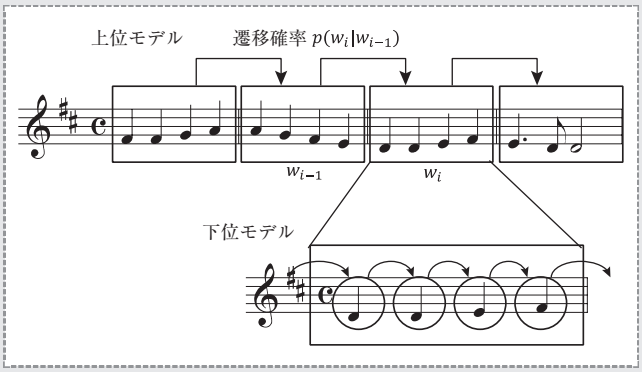

図2　音型マルコフモデル

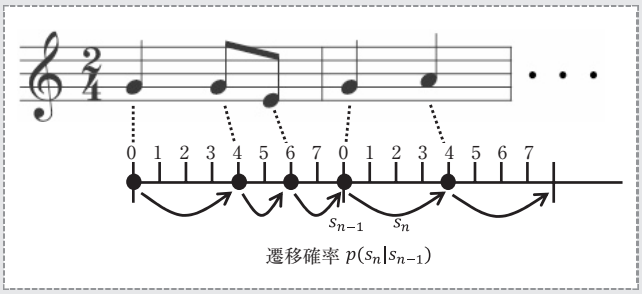

図3　拍節マルコフモデル

音符マルコフモデルは簡潔であり，他の確率モデルに組み入れやすい利点があるが，音価列の論理的な制約を表現できない欠点がある．例えば，三連符は三つ組の音符の組合せで現れるという制約は，高次のマルコフモデルを用いても表現できない．また，音符の局所的な隣接構造のみに着目しているため，音符系列を構成する大域的な拍節構造を表現することができない．

2.2　音型マルコフモデル

一般的な西洋音楽に存在する拍節構造を表現するため，音符の集合である「音型」を基本単位とする確率モデルが提案されている[1)2)]（図2）．音型マルコフモデルでは，ある時間単位（例：小節）ごとの音型を考える[1)]．以下，音型の種類数をKとし，各音型を$B_k = z_{k,1} \cdots z_{k,L}$ $(k = 1, \cdots, K)$と記す．ここで，$z_{k,l}$ $(l = 1, \cdots, L)$は音型kのl番目の音符を表す．音符列B_kはマルコフモデルに従うと仮定するのが一般的である．上位の階層において，音型の系列$w_1 \cdots w_I$ $(w_i \in \{B_k\}_{k=1}^{K})$がマルコフモデルに従うとすると，その生成確率は，式(1)と同様に計算できる．音符列$x_{1:N}$は生成された音型列$w_{1:I}$を結合することにより得られるので，$x_{1:N}$は階層マルコフモデルに従うことになる．一方，音型の系列$w_{1:I}$が確率的文脈自由文法（PCFG）に従うと仮定する音型PCFGモデルも提案されている[2)]．

音型マルコフモデルは，三連符などの論理制約を自然に記述できるが，音型の単位をどう設定すべきかは自明ではない．音楽的な意味を持ち得る音符系列の単位は動機やフレーズと呼ばれ，これらの長さは一定ではないので，本来は音型の長さを可変にすることが望まれる．また，小節線など音型の境界をまたぐ音符を含むシンコペーションの取り扱いが困難である．

2.3　拍節マルコフモデル

拍節構造を表現するための別のモデルとして，音符を小節内の相対的グリッド位置で記述する拍節マルコフモデルが提案されている[3)4)]（図3）．いま，各x_nは音価を表すとして，x_nの小節内の相対的な開始位置を$0 \le s_n < G$で表す．ここで，Gは小節をいくつのグリッドに分割するかを示す．例えば，4/4拍子の楽曲において，$G = 16$とすると，音符の最小時間単位は16分音符となる．このとき，音符の開始時刻の系列$s_{1:N}$がグリッド上のマルコフモデルに従うとすると，その生成確率は，式(1)と同様に計算できる．ただし，音価x_nは，

$$x_n = \begin{cases} s_{n+1} - s_n, & (s_{n+1} > s_n) \\ G + s_{n+1} - s_n, & (s_{n+1} \le s_n) \end{cases} \tag{2}$$

で与えられる．

拍節マルコフモデルでは，シンコペーションを自然に表現できる．式(2)において$s_{n+1} \neq 0$かつ$s_n > s_{n+1}$の場合，n番目の音符は小節線をまたいでいることを示す（s_{n+1}は次の小節内の位置）．一方，音価の種類を増やしたり，異なる拍子に対応する際には，モデルを再構成する必要があり，拡張性は高くない．

2.4　変形を考慮した音型マルコフモデル

実際の楽曲では，ある音型が繰り返し使用される際には変形を伴うことがある．例えば，ポピュラー音楽において，1番と2番の歌詞の音素数が異なると，メロディはほとんど同じであるものの，細部の音符配置が異なる場合がある．クラシック音楽においては，ある動機が繰り返し使用され，その際に修飾を受けたりなどして変形を伴うことが一般的である．

この現象を表現するため，2.2項で説明した音型モデルを拡張し，音型の反復と変形を同時に考慮できるモデルが提案されている[5)]（図4）．このモデルは階層構造を持ち，ま

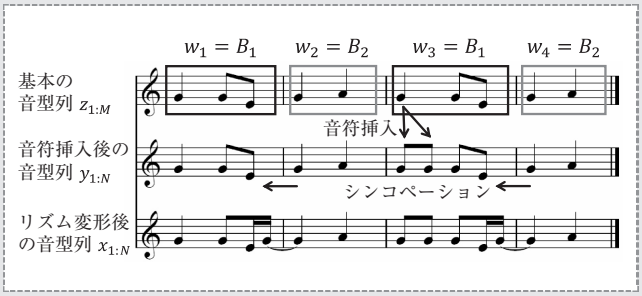

図4　変形を考慮した音型マルコフモデル

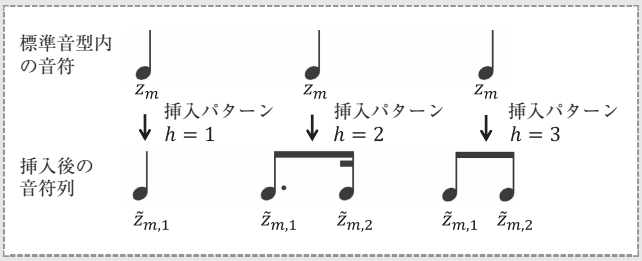

図5　リズムの変形：音符の挿入

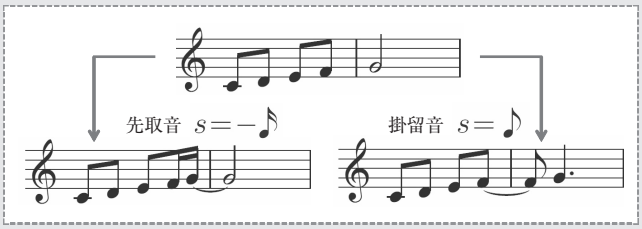

図6　音型の変形：シンコペーション

ず，音型モデルにより標準音型からなる音符系列$z_{1:M}$が生成され，変形が加わることによって最終的な音符系列$x_{1:N}$が生成されると考える．このとき，よく利用される変形操作として，音符の挿入とシンコペーションに着目し，これらの内で音価の総和が変化しないものが考慮されている．

まず，音符z_mに対する挿入操作で得られる音符列を$\tilde{z}_{m,1:Q}$とする（図5）．ここで，Qは挿入後の音符数であり，$\tilde{z}_{m,1}+\cdots+\tilde{z}_{m,Q}=z_m$が成立する．いま，可能な挿入操作の集合を$\{C_h \to D_h\}_{h=1}^{H}$とすると（$H$は挿入パターンの総数），挿入操作の生起確率は$p(\tilde{z}_{m,1:Q}=D_h|z_m=C_h)=p(D_h|C_h)$となる．音符列$z_{1:M}$内の各音符$z_m$に対して挿入パターン$h_m$を適用して得られる音符列を$y_{1:N}$で表す．この音符挿入による変形過程は，2.2項の音型モデルに，下位のマルコフモデルを付け加えることで記述できる．

シンコペーションは，音型の境界をまたいである音型の最後の音符と次の音型の最初の音符の音価とが同時に変形されたものである．音符列$y_{1:N}$に対してシンコペーションを適用して得られる音符列が$x_{1:N}$である．シンコペーションは，その度合を表す変数sによりパラメータ化でき，次の音符の変化に対応する（図6）．

$$y_n \to x_n = y_n + s \qquad y_{n+1} \to x_{n+1} = y_{n+1} - s \tag{3}$$

ここでx_{n+1}はある音型の最初の音符を表す．変数sは，正の場合は掛留音（Suspension），負の場合は先取音（Anticipation）を表す．それらが生起する確率を$p(s)$とする．モデルパラメータ（音型の遷移確率，$p(s)$や$p(D_h|C_h)$など）は，大量の音符系列データから教師なしで学習することができる．

このモデルは，MIDI演奏の楽譜化（リズム採譜）に応用され，良好な結果が得られている．図7に採譜例を示す．人間の演奏には大きなテンポ変動が含まれているため，各音符の音長を単純に量子化しただけでは，音楽的に不自然な楽譜が得られる．そこで，反復と変形を考慮した音型マルコフモデルを楽譜の事前分布とし，観測できない楽譜をベイズ推論することにより，音楽特有の構造を考慮した楽譜推定が可能になる．

2.5　木構造モデル

音符系列の階層構造を記述するモデルとして，音楽理論Generative Theory of Tonal Music（GTTM）[6]が知られている．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仮定のもと，音符系列は，動機・楽節・セクションといった構造に再帰的にグループ化される．タイムスパン木は，音符系列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，音符系列を簡略化する過程を表す（図8）．簡略化の過程では，二つの隣り合う音符（子ノード）が一つの音符（親ノード）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，いずれかの音高が簡略化された音符の音高として使われる．また，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．

GTTMに基づく解析を行ううえで，多数ある簡約化規則の優先度を決定する必要があるため，PCFGに基づく音符系列の確率的生成モデルが提案されている[7]．このモデル

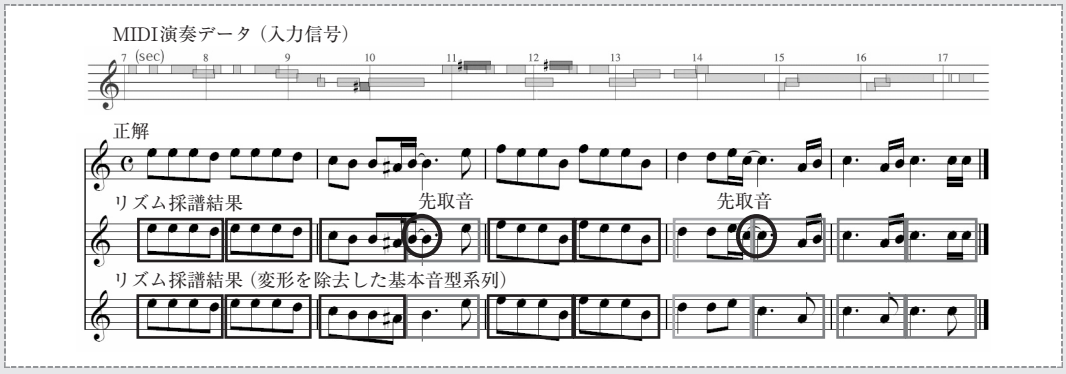

図7　MIDI演奏に対するリズム採譜例

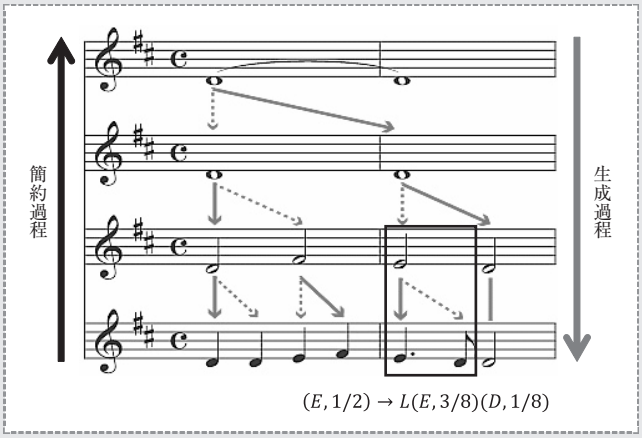

図8　GTTMに基づくPCFGモデル

では，1つの音符が再帰的に分割されて音符系列が生成されると考える（図8）．一般的なPCFG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），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$A \rightarrow \alpha\,(A \in \Omega_N,\ \alpha \in \Omega_N \times \Omega_N \cup \Omega_T)$の形で表され，確率$p\,(A \rightarrow \alpha)$で使用される．与えられた非終端記号系列$\boldsymbol{w} = w_1 \cdots w_N\ (w_n \in \Omega_T)$に対して，それを導出する一連の生成規則の適用は木構造をなす．

GTTMの確率モデルを定式化するには，通常のPCFGに対する拡張が必要となる．いま，音符を音高pと音価rのペアにより表し，音符系列を$\{x_n = (p_n \in \Omega_p,\ r_n \in \Omega_r)\}_{n=1}^{N}$とする．タイムスパン木の各ノードはリーフと同様にある音符であるから，非終端記号は終端記号と同じく音符の集合である．また，タイムスパン木で表される音符の主従関係を記述するため，LとRの二値をとる確率変数sを導入し，生成規則を$(p, r) \rightarrow s\,(p_L, r_L)\,(p_R, r_R)$の形で表す．ただし，$p, p_L, p_R \in \Omega_p,\ r, r_L, r_R \in \Omega_r$であり，$(p_s, r_s)$が主となる音符を表す．音価の和の保存に関する制約条件は$r = r_L + r_R$で与えられる．この生成規則の確率は，大量の音符系列データから教師なしで学習することができ，計算言語学的な手法により，音楽構造の統計的な解析が可能となった．

3．和音系列に対する確率モデル

次に，和音系列に対する言語モデルについて紹介する．和音系列も音符系列と同様に，離散記号の系列である$x_{1:N}$として表せることから，基本的には2節で紹介したのと同様の確率モデルが利用できる．ただし，x_nは和音シンボルであるが，和音の語彙をどう設定するかは必ずしも自明ではない．従来は多くの場合，和音のルート音12種類と和音の種類としてmajorあるいはminorの組合せを考えることで，和音の語彙サイズは24としていた．実際には，7 thコードなど複雑な和音も多く含まれているため，最近は10種類以上の和音から構成される大規模語彙を用いる研究が増えつつある．

3.1　マルコフモデル

和音系列に対する言語モデルとして最も広く用いられているのは，マルコフモデルである[8) 9)]．例えば，音楽音響信号に対して和音認識を行う際には，隠れマルコフモデル（HMM）を用いることが一般的であり，HMMの潜在空間はマルコフ連鎖をなす和音系列となっている[8)]．基本的なマルコフモデル（n-gramモデル）の拡張として，音符の任意の組合せを和音として許容することで，和音の語彙を事前に限定する必要がなく，系列中の和音ごとに最適な次数（過去の何個の和音に依存しているか）を推定することができる無限語彙可変長n-gramモデルが提案されている[9)]．

3.2　潜在変数モデル

最近では，和音系列の生成モデルとして，HMMやPCFGなどの潜在変数モデルを仮定する試みがなされている[10)]．大量の和音系列データに対して教師なし学習を行うことで，HMMでは潜在変数系列として，PCFGではリーフノードに，和声理論における「機能」に相当する概念が自動獲得できることが報告されている．PCFGに基づく和音系列の生成モデルは，メロディに対する自動和声付けに応用され，従来のHMMに基づく自動和声付けより優れた結

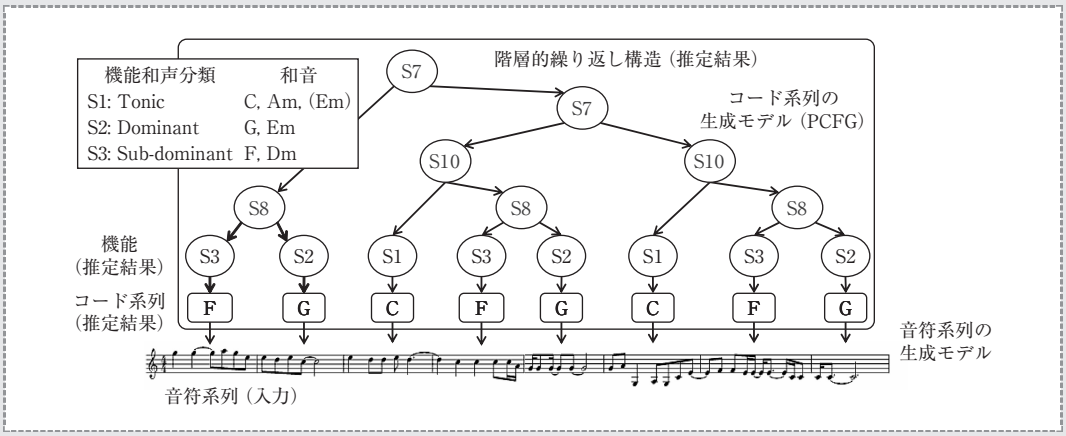

図9　和音系列に対するPCFGに基づく自動和声付け

果が得られている（図9）．具体的には，PCFGに基づく和音系列の生成モデルと，和音が与えられた上でのメロディの生成モデルを階層的に統合することで，メロディが与えられたときに，最尤な和音系列を効率的に求めることができる．この手法では，和音系列の背後に存在する階層的な繰り返し構造と，和音の機能を同時に考慮しつつ，音楽的に妥当な和音系列の生成が可能になっている．

4．むすび

本稿では，音符系列や和音系列に対する言語モデルを紹介した．これらの言語モデルの応用として，MIDI演奏に対するリズム採譜（音符系列モデル＋演奏モデル）や，メロディに対する自動和声付け（和音系列モデル＋音符系列モデル）を紹介した．この他にも，適切な音響モデルと組合せることにより，音楽音響信号に対する和音認識や自動採譜にも応用可能である．このように，さまざまな応用において，音楽的に妥当な出力を得るため，言語モデルを事前分布とした階層ベイズモデルを構築する方式は有望である．

楽譜に対する完全な言語モデルを定式化する上で，重要な課題が残されている．まず，多声楽曲（複数の声部からなる楽曲）を扱えるように，単一の音符系列に対する確率モデルを拡張する必要がある．このとき，声部間である程度リズムを同期する機構が必要となる．最終的には，テンポやリズムの生成過程を考慮しつつ，和音系列と音符配置を統一的に記述できる確率モデルの定式化を目指す必要がある．

謝辞　本稿の執筆に際し，中村栄太氏（京都大学/学振PD）から有益な助言をいただいた．また，JSPS科研費No. 26700020，No.16H01744およびJST ACCEL No.JPMJAC1602の支援を受けた．

（2017年5月8日受付）

〔文　献〕

1）H. Takeda, T. Otsuki, N. Saito, M. Nakai, H. Shimodaira and S. Sagayama: "Hidden Markov Model for Automatic Transcription of MIDI Signals", MMSP, pp.428-431（2002）
2）M. Tsuchiya, K. Ochiai, H. Kameoka and S. Sagayama: "Probabilistic Model of Two-Dimensional Rhythm Tree Structure Representation for Automatic Transcription of Polyphonic MIDI Signals", APSIPA, pp.1-6（2013）
3）M. Hamanaka, M. Goto, H. Asoh and N. Otsu: "A Learning-Based Quantization: Unsupervised Estimation of the Model Parameters", ICMC, pp.369-372（2003）
4）C. Raphael: "Automated Rhythm Transcription", ISMIR, pp.99-107（2001）
5）E. Nakamura, K. Itoyama and K. Yoshii: "Rhythm Transcription of MIDI Performances Based on Hierarchical Bayesian Modelling of Repetition and Modification of Musical Note Patterns", EUSIPCO, pp.1946-1950（2016）
6）F. Lerdahl and R. Jackendoff: "A Generative Theory of Tonal Music", MIT Press, Cambridge（1983）
7）E. Nakamura, M. Hamanaka, K. Hirata, K, Yoshii: "Tree-Structured Probabilistic Model of Monophonic Written Music Based on the Generative Theory of Tonal Music", ICASSP, pp.276-280（2016）
8）K. Lee and M. Slaney: "Automatic Chord Recognition from Audio Using an HMM with Supervised Learning", ISMIR, pp.133-137（2011）
9）K. Yoshii and M. Goto: "A Vocabulary-Free Infinity-Gram Model for Nonparametric Bayesian Chord Progression Analysis", ISMIR, pp.645-250（2011）
10）津島啓晃，中村栄太，糸山克寿，吉井和佳：“ベイズ文脈自由文法に基づく和音系列の教師なし構文解析と自動生成”，情処全大（2017）

吉井（よしい）　和佳（かずよし）　2008年，京都大学大学院情報学研究科博士後期課程修了．同年，産業技術総合研究所情報技術研究部門に入所．2014年，京都大学大学院情報学研究科講師に着任．音楽情報処理や統計的音響信号処理の研究に従事．博士（情報学）．