社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

[ポスター講演] 音楽音響信号に対する多重音高推定と和音構造学習のための階層ベイズ音響・言語統合モデル

尾島　優太†　　中村　栄太†　　糸山　克寿†　　吉井　和佳†

† 京都大学 大学院情報学研究科
E-mail: †{ojima,enakamura,itoyama,yoshii}@sap.ist.i.kyoto-u.ac.jp

あらまし　本稿では，音楽音響信号から教師なしで和音構造を学習し，同時に音高を推定するための手法を提案する．従来より多重音音高推定には非負値行列因子分解（NMF）が広く用いられてきた．NMF を用いた手法では，ピアノの鍵盤に対応する各音について音量の時間的変化を推定し，それを閾値処理することでピアノロールを出力していた．このような段階的な処理では適切な閾値を決定することが困難であり，さらに同時に出現する音高間の関係が考慮されず，音楽的に不自然な結果となるといった問題があった．本研究ではこの問題を解決するため，スペクトログラムの生成過程を表現する音響モデルに，ピアノロールの生成過程を表現する音楽文法としての言語モデルを統合した階層ベイズモデルを提案する．実験の結果，提案法により音楽文法の一種であるコード構造が正しく学習され，さらに自動採譜への活用の可能性が確認された．

キーワード　多重音音高推定，非負値行列因子分解，教師なし学習，階層ベイズモデル，自動採譜

1. はじめに

計算機による自動採譜の最終的な目標は，楽譜の主要要素である音高，音価を音楽音響信号から獲得することである．本研究ではこの自動採譜問題の一部である多重音の音高推定を扱う．具体的には，音楽音響信号を入力として二値のピアノロール形式の楽譜を出力することが目的である．

音高推定には従来より非負値行列因子分解（NMF）が広く用いられてきた [1–4, 4–6]．NMF は観測音響信号のスペクトログラムを，各音高の周波数スペクトルを表す基底行列と各音の時間的音量変化を表すアクティベーション行列の積の形に近似し分解する手法である．音高推定では，NMF により得られたアクティベーション行列に対し閾値処理や隠れマルコフモデル（HMM）に基づく二値化を行うことにより，各音高の存在を決定する [6, 7].

しかし，このような手法には二つの問題が存在した．一つ目は，曲ごとに適切な閾値を設定することの困難さである．二つ目は，推定結果の音楽的な不自然さである．これは，推定の際に各音の間の関係性が考慮されないために生じるものである．実際の音楽では和声構造が存在し，ある種の音高の組み合わせ（例えば C，G，E）が同時に発音されてコード（C メジャー）を形成する．さらにコードは時間的に変化し，典型的なコード進行を形成する．音高とコードは相互に依存する鶏と卵の関係にあるため，互いの情報を利用して同時に推定する必要がある．

本稿ではこの問題を解くため，コードと音高の依存関係を考慮しつつ，教師なしで音楽音響信号からコードと音高を推定するための統計的手法を提案する．詳細には，生成モデルとして，音高からスペクトログラムが生成される過程を表す音響モデル（NMF に基づく確率モデル）と調及びコード列から音高が生成される過程を表す言語モデル（HMM）を統合した階層ベイズモデルを定式化する．本モデルの特徴は，各音の存在を表す二値変数を NMF の枠組みに導入した点である．これにより言語モデルの HMM は，ピアノロールを表す二値変数を観測としてモデル化することができる．統合されたモデルでは与えられたスペクトログラムに基づき，ギブスサンプリングを用いてすべての隠れ変数（音高とコード）が同時に推定される．

本研究により，音楽音響信号から教師なしでの音楽文法の推論が可能となった．本研究でのモデル統合は自動音声認識（ASR）と同様の試みであるが，本研究は両方のモデルを教師なしで学習するという点で異なる．さらに，ASR のモデルは単語とスペクトログラムの二階層からなるが，本研究のモデルではコード，音高及びスペクトログラムの三階層からなる点でも異なる．この違いは言語モデルとして n-gram モデルであるマルコフモデルではなく隠れマルコフモデルを用いているために生じるものである．

2. 関連研究

本節では多重音音高推定（音響モデル）と音楽理論の実装及び音楽文法推論（言語モデル）についての関連研究を概観する．

2.1 音響モデル

音楽信号解析に対するアプローチは非負値行列因子分解（NMF）によるものが主流である [1–4, 4, 5, 8]．Cemgil ら [8] は NMF に対するベイズ推論の枠組みを示し，それにより様々な事前分布の導入が可能となった．Hoffman ら [3] はガンマ過程 NMF と呼ばれる NMF のノンパラメトリックベイズモデルを提案し，これにより基底数の自動推定が可能となった．Liang ら [5] は各基底の各時間フレームに対し二値変数を導入したベータ過程 NMF を提案した．NMF の別の拡張としては，基

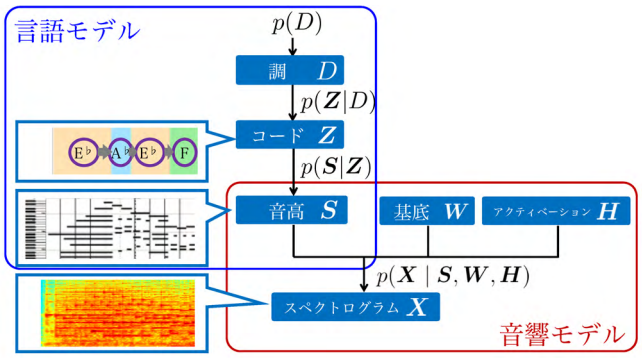

図 1　提案法の全体像.

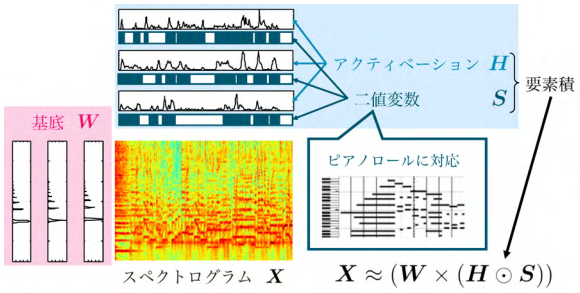

図 2　音響モデルの全体像.

底を，基本周波数を表すソースと音色を表すフィルターにさらに分解するソースフィルタ NMF [4] が存在する．

2.2　言語モデル

音楽の背後に存在する音楽理論の推定及び実装も研究されている [9–11]．例えば，音楽の様々な要素を一つの枠組みで記述した Generative Theory of Tonal Music (GTTM) [12] を計算機向けに定式化する試みがある．Hamanaka ら [9] は計算機による実装を通して GTTM を再定式化し，タイムスパン木と呼ばれる，音楽構造を表す木構造を自動獲得するための手法を提案した．Nakamura ら [10] も確率文脈自由文法を用いて GTTM を再定式化し，その推論アルゴリズムを提示した．

一方で，教師なしで音楽理論を推論する試みも存在する．Hu ら [11] は潜在的ディリクレ配分法を拡張し，同じ調性を持つ曲では同じ音が出やすくなるという知見に基づいて，楽譜及び音響信号から調を決定するための手法を提案した．この手法により，ラベル付けされた教師データなしで，ある調のもとでの各音の出やすさを獲得することが可能となった．

また，コードの概念も音楽文法の一種として考えられる．教師データを用いたコード推定のための統計的手法は広く研究されてきた [13–16]．Rocher ら [13] は与えられた楽譜に対し，ありうるコード遷移を有向グラフで表し，その中の最適経路を計算することでコード認識を試みた．Sheh ら [14] はクロマベクトルと呼ばれる音響特徴量を用いて音楽音響信号からコードを推定した．この手法ではコードラベルを隠れ変数とし，観測がクロマベクトルであるような HMM を構成し，コード列を決定する．Maruo ら [15] はクロマベクトルと NMF の双方を用いてコード推定を行い，コード推定精度の向上を実現した．これらの手法はいずれもラベル付けされた教師データが必要であり，そのアノテーションの際にコードの概念が必要である．さらに，コード列を音高推定に利用する試みも行われている [17,18]．これらの手法では動的ベイジアンネットワークを用いてコード進行と多重音の音高を同時推定し，単純な音響モデルの下でもよい性能を出している．近年では言語モデルとして再帰型ニューラルネットワークを用いて音高間の関係を記述するモデルも提案されている [19,20]．

3.　提　案　法

本節では提案法である，音楽音響信号から時間フレーム単位で音高とコードを同時推定するための手法について説明する．まず，観測される音楽スペクトログラムが生成される過程の確率的生成モデルとしての定式化について述べる．提案する生成モデルは，音響モデルと言語モデルが，各音高の存在を表す二値変数列であるピアノロールにより結び付けられた階層構造となっている（図 1）．ピアノロール，基底スペクトル及び各音の音量の時間的変化から音楽スペクトログラムが生成される過程を音響モデルにより表現し，調，コード列からコード進行及び同時に出現する音高の組み合わせが生成される過程を言語モデルにより表現する．最後に，逆問題として，与えられた音楽スペクトログラムを用いたモデル内の確率変数の推定について述べる．

3.1　問 題 設 定

多重音音高推定の目標は音楽音響信号からピアノロール形式の出力を得ることである．すなわち，周波数ビン数を F，時間フレーム数を T としたときに，音楽音響信号の対数周波数領域のスペクトログラム $\boldsymbol{X} \in \mathbb{R}_+^{F\times T}$ を，K 種類の音高及び T 個の時間フレームからなるピアノロール $\boldsymbol{S} \in \{0,1\}^{K\times T}$ に変換することが目的である．さらに，本手法ではコード列 $\boldsymbol{Z} = \{z_t\}_{t=1}^{T}$ の推定も行う．

3.2　音響モデル定式化

音響モデルは二値変数を持つベータ過程 NMF [5] と同様に定式化される（図 2）．与えられたスペクトログラム $\boldsymbol{X} \in \mathbb{R}_+^{F\times T}$ は基底 $\boldsymbol{W} \in \mathbb{R}_+^{F\times K}$，アクティベーション $\boldsymbol{H} \in \mathbb{R}_+^{K\times T}$ 及び二値変数 $\boldsymbol{S} \in \{0,1\}^{K\times T}$ の積の形として，以下に示すように分解される．

$$X_{ft}|\boldsymbol{W},\boldsymbol{H},\boldsymbol{S} \sim \mathrm{Poisson}\left(\textstyle\sum_{k=1}^{K} W_{fk}H_{kt}S_{kt}\right) \qquad (1)$$

ここで，$\{W_{fk}\}_{f=1}^{F}$ は k 番目の基底スペクトルを，H_{kt} は基底 k の時刻 t における音量を，S_{kt} は基底 k が時刻 t において使われているかどうかを示す二値変数を表す．

基底スペクトル $\boldsymbol{W}$ は，調波構造を表すスペクトルと非調波構造を表すスペクトルの二種類で構成される．本研究では調波構造スペクトルとして K_h 個の異なる音高に対応する K_h 個のスペクトルと，非調波構造スペクトルとして一つのスペクトルを用意する（$K = K_h + 1$）．同じ楽器では音高の異なる調波構造は音高に応じてシフトされただけの関係であり調波構造は変化しないと仮定すると，調波構造 $\boldsymbol{W}$ は，$k = 1, \ldots K_h$ について

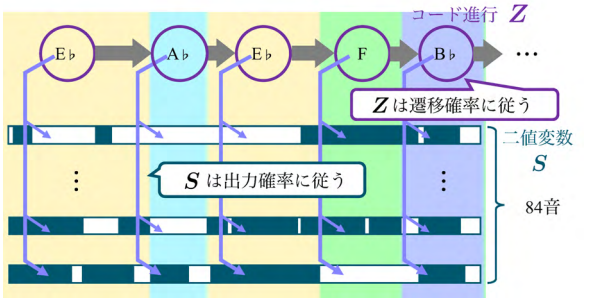

図 3　言語モデルの全体像.

$$\{W_{fk}\}_{f=1}^{F} = \mathrm{shift}\left(\{W_f^{\mathrm{h}}\}_{f=1}^{F},\ \zeta(k-1)\right) \tag{2}$$

となる．ここで，$\{W_f^{\mathrm{h}}\}_{f=1}^{F}$ はどの音高でも共通する調波構造テンプレートであり，$\mathrm{shift}\,(\boldsymbol{x}, a)$ は $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ を $[0, \ldots, 0, x_1, \ldots, x_{n-a}]^T$ へとシフトする演算である．また，ζ は半音の音程に対応する周波数ビン数である．

調波構造テンプレートと非調波構造スペクトルに対し，二種類の事前分布を用意する．まず，調波構造テンプレートは，次式に示すようにガンマ分布を事前分布として用意し，スパースになるように誘導する．

$$W_f^{\mathrm{h}} \sim \mathcal{G}\left(a^{\mathrm{h}}, b^{\mathrm{h}}\right) \tag{3}$$

ここで，a^{h} と b^{h} はハイパーパラメータである．一方，非調波構造スペクトルは，次式に示すように逆ガンマ連鎖事前分布 [21] を事前分布として用意し，周波数方向に滑らかになるように誘導する．

$$\begin{aligned} G_f^W | W_{f-1}^{\mathrm{n}} &\sim \mathcal{IG}\left(\eta^W, \frac{\eta^W}{W_{f-1}^{\mathrm{n}}}\right), \\ W_f^{\mathrm{n}} | G_f^W &\sim \mathcal{IG}\left(\eta^W, \frac{\eta^W}{G_f^W}\right) \end{aligned} \tag{4}$$

ここで，η^W は滑らかさを決定するハイパーパラメータで，G_f^W は W_{f-1}^{n} と W_f^{n} が正の相関を持つように導入した補助変数である．

アクティベーション行列 $\boldsymbol{H}$ も基底行列 $\boldsymbol{W}$ と同様に定式化される．H_{kt} がほぼ 0 となってしまうと S_{kt} の値が NMF に影響を与えず，$\boldsymbol{S}$ がマスクとしての機能を果たさない．H_{kt} の事前分布として逆ガンマ分布をおくことで H_{kt} が常にある程度の値を持つように誘導すればこの問題は回避できる．さらに，時間方向の滑らかさを導入するため，$\boldsymbol{H}$ に対し下式に示す逆ガンマ連鎖事前分布を与える．

$$\begin{aligned} G_{kt}^H | H_{k(t-1)} &\sim \mathcal{IG}\left(\eta^H, \frac{\eta^H}{H_{k(t-1)}}\right), \\ H_{kt} | G_{kt}^H &\sim \mathcal{IG}\left(\eta^H, \frac{\eta^H}{G_{kt}^H}\right) \end{aligned} \tag{5}$$

ここで，η^H は滑らかさを決定するハイパーパラメータで，G_{kt}^H は $H_{k(t-1)}$ と H_{kt} が正の相関を持つように導入した補助変数である．

3.3　言語モデル定式化

言語モデルは，マルコフ性を持つコード列 $\boldsymbol{Z} = \{z_1, \ldots, z_T\}$ $(z_t \in \{1, \ldots, I\})$ を隠れ変数に持ち，二値変数 $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_T\}$ $(\boldsymbol{s}_t \in \{0,1\}^{K_h})$ を出力する HMM として定式化される（図 3）．ここで I は隠れ状態の種類，すなわちコードの種類であり，K_h は出現する可能性がある音高の数を表す．また，HMM のパラメータの一部（遷移確率，初期確率のハイパーパラメータ）は調により決定されるモデルとすることで，調についても同時推定を行う．なお，提案モデル全体で考えると，$\boldsymbol{S}$ は実際には隠れ変数である．調の総数を J，曲全体の調を表す番号を D（$D \in \{1, \ldots, J\}$）とすると，HMM は以下に示すように定式化される．

$$z_1 | \boldsymbol{\phi}_D \sim \mathrm{Categorical}(\boldsymbol{\phi}_D), \tag{6}$$

$$z_t | z_{t-1}, \boldsymbol{\psi}_{D,z_{t-1}} \sim \mathrm{Categorical}(\boldsymbol{\psi}_{D,z_{t-1}}), \tag{7}$$

$$S_{kt} | z_t, \pi_{z_t k} \sim \mathrm{Bernoulli}(\pi_{z_t k}) \tag{8}$$

ここで $\boldsymbol{\psi}_{D,i} \in \mathbb{R}^I$ は調 D の下でのコード i からの遷移確率，$\boldsymbol{\phi}_D \in \mathbb{R}^I$ は調 D の下での初期確率，$\pi_{z_t k}$ はコード z_t の下で k 番目の音高が出力される確率を表す．

これらのパラメータに対し，共役事前分布

$$\boldsymbol{\psi}_{D,i} \sim \mathrm{Dir}(\mathbf{1}_I), \quad \boldsymbol{\phi}_D \sim \mathrm{Dir}(\mathbf{1}_I), \quad \pi_{z_t k} \sim \mathrm{Beta}(e, f) \tag{9}$$

をおく．ここで $\mathbf{1}_I$ は全要素が 1 の I 次元ベクトルであり，e と f はハイパーパラメータである．

実際には出力確率には 1 オクターブ内に出現する 12 音高（C, C♯, …, B）分だけを用意し，これをすべてのオクターブで用いることで K_h 種類の音高を表現する．さらに，コードのうち，種類が同じで根音が異なるものについては出力確率を共有し，根音の位置に応じて巡回シフトしたものとする．本稿では簡単のため，コードの種類として 2 種類のみを考える（$I = 2 \times 12$）．これは，メジャーコードとマイナーコードを想定したものである．

次に，調に関するモデル化を考える．調 D は以下に示す事後分布に従う．

$$D \sim \mathrm{Categorical}(\boldsymbol{\delta}), \quad \boldsymbol{\delta} \sim \mathrm{Dir}(\mathbf{1}_J) \tag{10}$$

ここで，$\mathbf{1}_J$ は全要素が 1 の J 次元ベクトルであり，$\boldsymbol{\delta}$ は各調の選ばれやすさを表す．また，HMM における遷移確率及び初期確率については，データを効率的に利用するため，長調，短調の二種類のみを用意し，調の主音に応じてシフト巡回することで決定する．すなわち，A の音と主音の音程が $b \in \{0, \ldots, 11\}$，種類が $k \in \{\mathrm{major}, \mathrm{minor}\}$ なる調 D の下での初期確率及び遷移確率 $\boldsymbol{\phi}_D, \boldsymbol{\psi}_{D,i}$ は，下式に従い決定される．

$$\boldsymbol{\phi}_D = \mathrm{rot}(\boldsymbol{\phi}_k, b), \quad \boldsymbol{\psi}_{D,i} = \mathrm{rot}(\boldsymbol{\psi}_{k,\lceil i+b,12 \rceil}, b) \tag{11}$$

ここで，$\boldsymbol{\phi}_k$ は調の種類が k の下での初期確率，$\boldsymbol{\psi}_{k,i}$ は調種類が k の下での状態 i からの遷移確率であり，$\lceil a, x \rceil$ は $a \bmod x$ を表す．また，$\mathrm{rot}(\boldsymbol{x}, a)$ は $\boldsymbol{x} = [x_1, \ldots, x_n]^T$ を $[x_{\lceil 1-a,n \rceil}, \ldots, x_{\lceil n-a,n \rceil}]^T$ へと巡回シフトする演算である．

3.4　事後分布推論

以上のモデルについて，観測データ $\boldsymbol{X}$ が与えられた下での事後分布 $p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{S}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\psi} | \boldsymbol{X})$ を推論する必要があるが，解析的に計算することは不可能である．そのため，[22] にあるように，マルコフ連鎖モンテカルロ法を用いて推論を行う．音響

モデルと言語モデルは二値変数のみを共有するため，二値変数が与えられるとそれぞれのモデルは独立に更新できる．これら二つのモデルと二値変数をサンプリングにより交互に更新し，最後に言語モデルの隠れ変数（コード進行）はビタビアルゴリズムにより推定する．また，二値変数（ピアノロール）の最終的な出力は，尤度が最大となるパラメータを用いて決定される．

3.4.1 二値変数の推論

二値変数 $\boldsymbol{S}$ は音響モデルと言語モデルの双方を結びつけるパラメータであり，各音の使われやすさはコードにより決定され，各音が使われたかどうかが再構成されたスペクトログラムに影響する．そのため，音響モデルを尤度関数，言語モデルを事前分布とみなし，ベイズ則に基づき計算される事後分布を用い，二値変数をサンプルする．これは以下に示すように定式化される．

$$S_{kt} \sim \text{ Bernoulli}\left(\tfrac{P_1}{P_1+P_0}\right) \tag{12}$$

ここで P_1 と P_0 は下式により計算される．

$$\begin{aligned} P_1 &= p(S_{kt}=1|S_{\neg k,t},\boldsymbol{x}_t,\boldsymbol{W},\boldsymbol{H},\boldsymbol{\pi},\boldsymbol{z}) \\ &\propto \pi_{z_k}^{\alpha}\textstyle\prod_f\left(\hat{X}_{ft}^{\neg k}+W_{fk}H_{kt}\right)^{X_{ft}}\exp\{-W_{fk}H_{kt}\}, \end{aligned} \tag{13}$$

$$\begin{aligned} P_0 &= p(S_{kt}=0|S_{\neg k,t},\boldsymbol{x}_t,\boldsymbol{W},\boldsymbol{H},\boldsymbol{\pi},\boldsymbol{z}) \\ &\propto (1-\pi_{z_k})^{\alpha}\textstyle\prod_f\left(\hat{X}_{ft}^{\neg k}\right)^{X_{ft}} \end{aligned} \tag{14}$$

ここで，$\hat{X}_{ft}^{\neg k}\equiv\sum_{l\neq k}W_{fl}H_{lt}S_{lt}$ は k 番目の基底を用いずに再構成された振幅スペクトログラムの周波数ビン f，時間フレーム t における値を表し，α は言語モデルの重みを決定づけるパラメータである．このような言語モデルの重み付けは ASR でも行われる．α が 1 以外の値をとるとき，正規化項を解析的に計算することは不可能であり，ギブスサンプリングを用いることはできない．そのため，代わりに式 (12) を提案分布としたメトロポリス・ヘイスティング法を用いてサンプルする．

3.4.2 音響モデルの更新

音響モデルのパラメータ $\boldsymbol{W}^{\mathrm{h}}$，$\boldsymbol{W}^{\mathrm{n}}$，$\boldsymbol{H}$ はギブスサンプリングによりサンプルされる．これらのパラメータは，事前分布としてガンマ分布を持つ $\boldsymbol{W}^{\mathrm{h}}$ と逆ガンマ分布を持つ $\boldsymbol{W}^{\mathrm{n}}$ および $\boldsymbol{H}$ に大別される．ベイズ則に基づき計算すると，$\boldsymbol{W}^{\mathrm{h}}$ の条件付き事後分布は

$$W_{fk}^{\mathrm{h}}\sim\mathcal{G}\left(\textstyle\sum_t X_{ft}\lambda_{ftk}+a^{\mathrm{h}},\ \sum_t H_{kt}S_{kt}+b^{\mathrm{h}}\right) \tag{15}$$

となる．ここで，λ_{ftk} は最新のサンプル値 $\hat{\boldsymbol{W}}$，$\hat{\boldsymbol{H}}$，$\hat{\boldsymbol{S}}$ を用いて計算される正規化項であり，下式で定義される．

$$\lambda_{ftk}=\frac{\hat{W}_{fk}\hat{H}_{kt}\hat{S}_{kt}}{\sum_l\hat{W}_{fl}\hat{H}_{lt}\hat{S}_{lt}} \tag{16}$$

一方，残りのパラメータは補助変数を用いてサンプルされる．$\boldsymbol{H}$ については，式 (5) に示すように $\boldsymbol{G}^H$ と相互依存関係にあるため，同時にサンプルすることはできない．そのため，$\boldsymbol{H}$ と $\boldsymbol{G}^H$ を交互にサンプルする．観測 $\boldsymbol{X}$ の影響を考慮しない場合，$\boldsymbol{G}^H$ の条件付き事後分布は

$$G_{kt}^H\sim\mathcal{IG}\left(2\eta_H,\eta_H\left(\tfrac{1}{H_{kt}}+\tfrac{1}{H_{k(t-1)}}\right)\right) \tag{17}$$

であり，同様に $\boldsymbol{H}$ の条件付き事後分布は，

$$H_{kt}\sim\mathcal{IG}\left(2\eta_H,\eta_H\left(\tfrac{1}{G_{k(t+1)}^H}+\tfrac{1}{G_{kt}^H}\right)\right) \tag{18}$$

となる．これと同様に $\boldsymbol{G}^W$，$\boldsymbol{W}^{\mathrm{n}}$ の条件付き事後分布は

$$G_f^W\sim\mathcal{IG}\left(2\eta_W,\eta_W\left(\tfrac{1}{W_f^{\mathrm{n}}}+\tfrac{1}{W_{f-1}^{\mathrm{n}}}\right)\right), \tag{19}$$

$$W_f^{\mathrm{n}}\sim\mathcal{IG}\left(2\eta_W,\eta_W\left(\tfrac{1}{G_{f+1}^W}+\tfrac{1}{G_f^W}\right)\right) \tag{20}$$

となる．式 (18) を事前分布とみなし，式 (15) と同様にベイズ則とイェンゼンの不等式を用いることで，観測 $\boldsymbol{X}$ を考慮した，$\boldsymbol{H}$ の条件付き事後分布は以下に示すように計算される(注1)．

$$H_{kt}\sim\mathrm{GIG}\left(2S_{kt}\textstyle\sum_f W_{fk},\delta_H,\sum_f X_{ft}\lambda_{ftk}-\gamma_H\right)$$

ここで，$\gamma_H=2\eta_H$，$\delta_H=\eta_H(\frac{1}{G_{k(t+1)}^H}+\frac{1}{G_{kt}^H})$ とおいた．全く同様に，$\boldsymbol{W}^{\mathrm{n}}$ の条件付き事後確率は

$$W_{fk}^{\mathrm{n}}\sim\mathrm{GIG}\left(2\textstyle\sum_t H_{kt}S_{kt},\delta_W,\sum_t X_{ft}\lambda_{ftk}-\gamma_W\right)$$

となる．ただし，$\gamma_W=2\eta_W$，$\delta_W=\eta_W(\frac{1}{G_{f+1}^W}+\frac{1}{G_f^W})$ である．

3.4.3 言語モデルの更新

言語モデルの隠れ変数 $\boldsymbol{Z}$ は以下の条件付き事後分布に従いサンプルされる．

$$p(z_t|\boldsymbol{S},D,\boldsymbol{\pi},\boldsymbol{\phi},\Psi)\propto p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_t,z_t|D) \tag{21}$$

ここで，$\boldsymbol{\pi}$ は出力確率を，$\boldsymbol{\phi}$ は初期確率を，$\Psi=\{\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_I\}$ は各状態からの遷移確率を表す．$\boldsymbol{Z}$，$\boldsymbol{S}$ は条件付き独立なので式 (21) の右辺は更に分解され，

$$\begin{aligned} &p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_t,z_t|D) \\ &=p(\boldsymbol{s}_t|z_t)\textstyle\sum_{z_{t-1}}p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_{t-1},z_{t-1}|D)p(z_t|z_{t-1},D), \end{aligned} \tag{22}$$

$$p(\boldsymbol{s}_1,z_1|D)=p(z_1|D)p(\boldsymbol{s}_1|z_1)=\phi_{D,z_1}p(\boldsymbol{s}_1|\pi_{z_1}) \tag{23}$$

と表せる．式 (22) と式 (23) より，$p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_T|z_T)$ は再帰的に計算され（フォワードフィルタリング），$z_T\sim p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_T|z_T)$ に従い z_T をサンプルする．また，$z_{t+1},\ldots,z_T$ が与えられた下で

$$p(z_t|\boldsymbol{S},z_{t+1},\ldots,z_T,D)\propto p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_t,z_t|D)p(z_{t+1}|z_t,D) \tag{24}$$

に従い z_t をサンプルする．$p(\boldsymbol{s}_1,\ldots,\boldsymbol{s}_t,z_t)$ は式 (22) で計算されるので，このサンプルも再帰的に行われる（バックワードサンプリング）．

出力確率 $\boldsymbol{\pi}$ の事後分布はベイズ則より，

$$p\left(\boldsymbol{\pi}|\boldsymbol{S},\boldsymbol{z},\boldsymbol{\phi},\Psi,D\right)\propto p\left(\boldsymbol{S}|\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{\phi},\Psi,D\right)p\left(\boldsymbol{\pi}\right) \tag{25}$$

となる．$p\left(\boldsymbol{\pi}\right)$ は $p\left(\boldsymbol{S}|\boldsymbol{\pi},\boldsymbol{z},\boldsymbol{\phi},\Psi,D\right)$ の共役事前分布なのでこ

(注1)：$\mathrm{GIG}(a,b,p)\equiv\frac{(a/b)^{\frac{p}{2}}}{2K_p(\sqrt{ab})}x^{p-1}\exp(-\frac{ax+\frac{b}{x}}{2})$ は一般化逆ガウス分布を表す

の事後分布は解析的に計算できる．C_i を $\boldsymbol{Z}$ 内でのコード $i \in \{1 \dots I\}$ の出現回数，$\boldsymbol{c}_i \equiv \sum_{t \in \{t|z_t=i\}} \boldsymbol{s}_t$ を $z_t = i$ なる時間フレーム t における $\boldsymbol{s}_t$ の総和を表す K_h 次元ベクトルとすると，パラメータ $\boldsymbol{\pi}$ は以下の条件付き事後分布に従いサンプルされる．

$$\boldsymbol{\pi}|\boldsymbol{S}, \boldsymbol{z}, \boldsymbol{\phi}, \Psi \sim \mathrm{Beta}\left(e + c_{ik}, f + C_i - c_{ik}\right). \tag{26}$$

同様に，遷移確率 $\boldsymbol{\psi}_i$ 及び初期確率 $\boldsymbol{\phi}$ の事後分布は

$$p(\boldsymbol{\phi}|\boldsymbol{S}, D, \boldsymbol{z}, \boldsymbol{\pi}, \Psi) \propto p(z_1|\boldsymbol{\phi}, D)\ p(\boldsymbol{\phi}) \tag{27}$$

$$p(\boldsymbol{\psi}|\boldsymbol{S}, D, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\phi}) \propto \textstyle\prod_t p(z_t|z_{t-1}, \boldsymbol{\psi}_{z_{t-1}}, D)\ p(\boldsymbol{\psi}_{z_{t-1}}) \tag{28}$$

となる．$p(\boldsymbol{\phi})$，$p(\boldsymbol{\psi}_i)$ はそれぞれ $p(z_1|\boldsymbol{\phi}, D)$，$p(z_t|z_{t-1}, \boldsymbol{\psi}_{z_{t-1}}, D)$ の共役事前分布なので，簡単に事後確率が計算できる．$\boldsymbol{e}_i$ を i 番目の要素が 1 である単位ベクトル，$\boldsymbol{a}_i$ を j 番目の要素が状態 i から状態 j への遷移の回数を表す I 次元ベクトルとすると，$\boldsymbol{\phi}$ 及び $\boldsymbol{\psi}_i$ は以下の事後分布に従いサンプルされる．

$$\begin{aligned}\boldsymbol{\phi}|\boldsymbol{S}, D, \boldsymbol{z}, \boldsymbol{\pi}, \Psi &\sim \mathrm{Dir}\left(\mathbf{1}_I + \boldsymbol{e}_{z_1}\right), \\ \boldsymbol{\psi}_i|\boldsymbol{S}, D, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\phi} &\sim \mathrm{Dir}\left(\mathbf{1}_I + \boldsymbol{a}_i\right).\end{aligned} \tag{29}$$

実際には $\boldsymbol{\psi}_i, \boldsymbol{\phi}$ は調に応じて巡回シフトしているので，観測された回数も適切に巡回シフトする必要がある．

調の事後分布は，ベイズ則に基づき，下式のように表される．

$$p(D|Z, \boldsymbol{\phi}, \Psi, \boldsymbol{\delta}) \propto p(Z|D, \boldsymbol{\phi}, \Psi) p(D|\boldsymbol{\delta}) \tag{30}$$

ここで，$\boldsymbol{Z}, \boldsymbol{\phi}, \Psi$ は既知であるので，$p(Z|D, \boldsymbol{\phi}, \Psi)$ は全ての D について解析的に計算できる．また，式 (10) より $p(D|\boldsymbol{\delta}) = \delta_D$ なので，式 (30) に従い D をサンプルする．また，調のパラメータである $\boldsymbol{\delta}$ の事後分布は

$$p(\boldsymbol{\delta}|D) \propto p(D|\boldsymbol{\delta}) p(\boldsymbol{\delta}) \tag{31}$$

であり，$p(\boldsymbol{\delta})$ は $p(D|\boldsymbol{\delta})$ の事前分布なので，結局

$$\boldsymbol{\delta}|D \sim \mathrm{Dir}(\mathbf{1}_J + \boldsymbol{e}_D) \tag{32}$$

となる．

4. 評価実験

提案法の音高推定精度を評価するため，比較実験を行った．まず事前実験として，正しいピアノロールが与えられたときに言語モデルが正しくコード進行及び出力確率を推定することを確認した．次に，音響モデルのみを用いて音高推定した場合と，提案法である統合モデルを用いて音高推定した場合の推定精度を比較した．

4.1 実験条件

実験には MAPS データベース [23] から，“ENSTDkCl” のラベルが付いている 30 曲を用いた．いずれの曲もモノラル信号に変換した後，冒頭 30 秒を切り出して使用した．振幅スペクトログラムは変 Q 変換 [24] により得られた 926×10075 の行列を MATLAB の resample 関数により 926×3000 へと変換したものを用いた．さらに，事前処理として調波・非調波音分離（HPSS）[25] を行った．なお，元の論文とは違い HPSS は対数周波数領域に対して行っており，メディアンフィルタ幅は時間フレーム方向は 50，周波数方向は 40 とした．ハイパーパラメータは $I = 24, a^{\mathrm{h}} = 1, b^{\mathrm{h}} = 1, a^{\mathrm{n}} = 2, b^{\mathrm{n}} = 1, c = 2, d = 1, e = 5, f = 80, \alpha = 1300, \eta_W = 800000, \eta_H = 15000$ とし，これは実験的に決定した．出力確率は 12 音に対して用意され，これを全てのオクターブで共有することで 84 音高分の出力確率とした．更に遷移確率は，自己遷移確率（$p(z_{t+1} = z_t|z_t)$）を $1 - 8.0 \times 10^{-8}$ で固定し，他状態への遷移が 3. 4. 3 節で示したディリクレ分布に従うものとした．

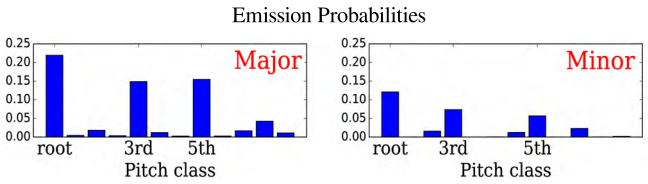

図 4　事前実験の結果得られた出力確率．

4.2 ピアノロールに対するコード推定

まず最初に，言語モデルが出力確率及びコード進行を正しく推定できることを確認するため，予備実験を行った．予備実験では入力として，MIDI 番号 21–104 に対応する 84 音に対する正解のピアノロールを 30 曲分連結した 84×90000 の行列を用いた．これに対し，言語モデルのうち調推定部分を除くコード推定部分を用いて，コード推定精度及び出力確率の推定を行った．コード推定精度は，推定コードと正解コードが一致した時間フレーム数の割合で評価した．コード種類としてはメジャーとマイナーの 2 種類のみを想定して用意したため，正解コードにおける「メジャー」と「メジャーセブンス」を「メジャー」，「マイナー」と「マイナーセブンス」を「マイナー」として評価した．この他の種類のコードは評価の際には無視した．また，教師なしでコードを推定しているため，具体的なコードラベルについてはシフト関係・コード種類を考慮した上で推定精度が最大になるものを採用した．MAPS データベースにはコード情報が含まれていないため，コード情報は筆者の 1 人が人手で与えたものを正解とした(注2)．

図 4 に示した結果より，調性音楽で頻出するメジャーコードとマイナーコードが出力確率として得られていることが分かる．このことは，ピアノロールのみに基づき，事前知識無しでコードの概念が自動獲得できていることを示しており，興味深い．コード認識率は 61.33%であり，教師なしの状況下でもコードの認識が可能であることが分かる．一方，他のコード認識に関する研究 [14, 15] ではこれよりも高い精度を達成している．これらの手法ではラベル付けされた教師データを使っており，かつコード認識に使ったデータが，コード構造がはっきりしてい

(注2)：コード情報は http://sap.ist.i.kyoto-u.ac.jp/members/ojima/mapschord.zip から入手可能

表 1　30 曲に対する音高推定結果.

実験条件	$\mathcal{F}$	$\mathcal{R}$	$\mathcal{P}$
言語モデル＋音響モデル（提案法）	65.0	67.3	62.8
コード遷移なし（条件 1）	64.7	64.7	64.7
コードを用いて事前学習（条件 2）	65.5	65.3	65.6
コード無しで事前学習（条件 3）	65.0	65.5	64.6

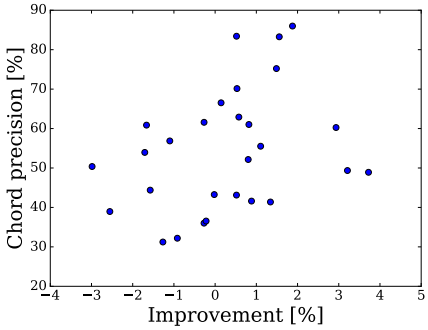

図 5　コード推定精度と f 値の向上幅の相関.

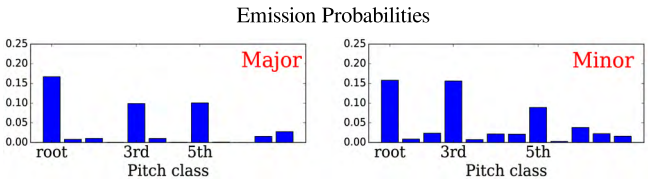

図 6　音楽音響信号から推定された出力確率.

るポピュラー音楽であるため，より高い精度を実現していると考えられる.

4.3　音楽音響信号の音高推定

次に，下式で定義される，フレーム単位での再現率・適合率・f 値により音高推定精度を評価した.

$$\mathcal{R} = \frac{\sum_t c_t}{\sum_t r_t},\ \mathcal{P} = \frac{\sum_t c_t}{\sum_t e_t},\ \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}} \qquad (33)$$

ここで，r_t，e_t，c_t はそれぞれ，t 番目の時間フレームにおける正解データの音高の数，推定された音高の数，正解データと一致した推定音高の数を表す．なお，曲全体を通したオクターブずれは許容した．比較実験として，以下の 3 条件のもとで音高推定を行った.

（1）　一曲の間でコード遷移が起きない

（2）　正解ピアノロール・コードに基づき言語モデルを事前学習し，学習されたパラメータを用いる

（3）　正解ピアノロールのみに基づきコードを推定しつつ言語モデルを事前学習し，学習されたパラメータを用いる

条件 2，3 については交差検定により評価を行った.

表 1 に示すように，教師なしでの音高推定精度（65.0%）は音響モデルのみでの音高推定精度（64.7%）よりも高かった．また，図 5 に示すように，言語モデルを統合することによる f 値の向上幅とコード推定精度の間には正の相関が見られた（相関係数 $r = 0.33$）．このことは，言語モデルの精度向上が音高推定の精度向上につながることを示唆している．さらに図 6 に示すように，音楽音響信号のみに基づいて事前知識無しで図 4 と

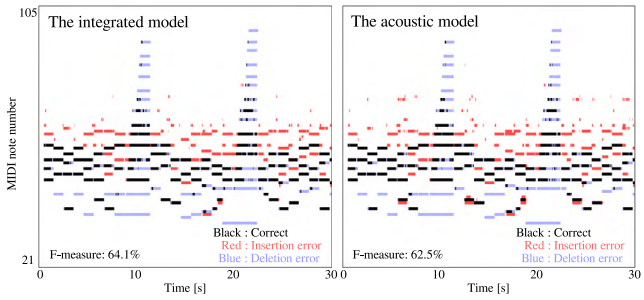

図 7　MUS-bk_xmas5_ENSTDkCl に対する音高推定結果の例.

同様のコード構造が出力確率として獲得された．この結果より頻出するコードの種類が音楽音響信号から自動的に獲得できることが分かる．このようなコード種類の獲得は音楽分類や類似度判定に有用である．また，事前学習した場合の結果（65.5%）は教師なしでの結果よりも高かった．コード情報を含むピアノ譜は多数出版されており，この条件は現実的であると考えられる．また，標準誤差は 1.5 ポイントであり精度向上は統計的には有意な差ではないが，f 値は 30 曲中 25 曲で向上し，さらにそのうち 15 曲では 1 ポイント以上向上した.

音高推定結果の一例を図 7 に示す．この図より，言語モデルを統合することで低音域の挿入誤りが減少していることが確認できる．一方，挿入誤りの総数は統合モデルでは増加している．これは調波構造においたシフト不変の条件が強すぎるものであり，その結果各音のスペクトルが正しく推定できずに，倍音を存在する音高として誤って推定していることが原因であると考えられる.

本手法には，十分な性能向上の余地がある．まず，音響モデルは上述したように調波構造について強い制約があり，この制約はソースフィルタ NMF [4] を用いることで緩和できると考えられる．ソースフィルタ NMF では基底行列が音高を表すソースと音色を表すフィルタにさらに分解される．提案モデルはこのフィルタが 1 つだけの場合に対応し，フィルタを増やすことで例えば高音と低音の音色の違いを表現することが可能になると考えられる．一方，言語モデルは現在は縦方向（同時刻の音高方向）の関係のみをモデル化しているが，横方向（音高の時間的遷移）をモデル化することで，オクターブ誤りや倍音誤りといった直前の音から大きく離れた位置に存在する挿入誤りの減少が期待でき，精度向上につながると考えられる.

5.　おわりに

本稿では，音高・コードを音楽音響信号から同時推定するためのモデルについて提案した．提案モデルは NMF に基づく音響モデルとベイジアン HMM に基づく言語モデルから構成され，両モデルの情報を用いて音高が決定される．実験結果から，音楽音響信号からの教師なしでの音高推定及び音楽文法推論の可能性が示された．一方，音響モデルは調波構造に対する制約が大きく，言語モデルはコード構造を記述するのみで音楽理論を表現するには不十分であるなど，いずれのモデルも十分に改善の余地がある.

提案法は文法推論の観点から，言語獲得と深く関係がある．自然言語処理の分野では，単語列に基づく教師なしの文法獲得や文字列に基づく教師なしの単語分割が研究されてきた [26, 27]．提案法は楽譜（離散記号列）や音楽音響信号から音楽文法を教師なしで推論することが可能であり，音声音響信号からの言語獲得 [28] などへの応用も期待される．

謝辞 本研究の一部は，JSPS 科研費 24220006，26700020，26280089，16H01744，15K16054，16J05486 と JST OngaCREST プロジェクトおよび栢森財団の支援を受けた．

文　献

[1] P. Smaragdis and J.C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp.177–180, 2003.

[2] K. Ohanlon, H. Nagano, N. Keriven, and M. Plumbley, “An iterative thresholding approach to L0 sparse hellinger nmf,” International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.4737–4741, 2016.

[3] M. Hoffman, D.M. Blei, and P.R. Cook, “Bayesian nonparametric matrix factorization for recorded music,” Proceedings of the 27th International Conference on Machine Learning (ICML), pp.439–446, 2010.

[4] J.L. Durrieu, G. Richard, B. David, and C. Févotte, “Source/filter model for unsupervised main melody extraction from polyphonic audio signals,” IEEE Transations on Audio, Speech, and Language Processing (TASLP), vol.18, no.3, pp.564–575, 2010.

[5] D. Liang, M.D. Hoffman, and D.P. Ellis, “Beta process sparse nonnegative matrix factorization for music.,” International Society for Music Information Retrieval Conference (ISMIR), pp.375–380, 2013.

[6] E. Vincent, N. Bertin, and R. Badeau, “Adaptive harmonic spectral decomposition for multiple pitch estimation,” IEEE Transactions on Audio, Speech, and Language Processing (TASLP), vol.18, no.3, pp.528–537, 2010.

[7] G.E. Poliner and D.P. Ellis, “A discriminative model for polyphonic piano transcription,” EURASIP Journal on Applied Signal Processing, pp.154–154, 2007.

[8] A.T. Cemgil, “Bayesian inference for nonnegative matrix factorisation models,” Computational Intelligence and Neuroscience, vol.2009, pp.1–17, 2009.

[9] M. Hamanaka, K. Hirata, and S. Tojo, “Implementing “ a generative theory of tonal music ”,” Journal of New Music Research, vol.35, no.4, pp.249–277, 2006.

[10] E. Nakamura, M. Hamanaka, K. Hirata, and K. Yoshii, “Tree-structured probabilistic model of monophonic written music based on the generative theory of tonal music,” International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.276–280, 2016.

[11] D. Hu and L.K. Saul, “A probabilistic topic model for unsupervised learning of musical key-profiles.,” International Society for Music Information Retrieval Conference (ISMIR)Citeseer, pp.441–446 2009.

[12] R. Jackendoff and F. Lerdahl, A generative theory of tonal music, MIT Press, 1985.

[13] M. Rocher, T.and Robine, P. Hanna, and R. Strandh, Dynamic Chord Analysis for Symbolic Music, Ann Arbor, MI: MPublishing, University of Michigan Library, 2009.

[14] A. Sheh and D.P. Ellis, “Chord segmentation and recognition using EM-trained hidden Markov models,” International Society for Music Information Retrieval Conference (ISMIR), pp.185–191, International Symposium on Music Information Retrieval, 2003.

[15] S. Maruo, K. Yoshii, K. Itoyama, M. Mauch, and M. Goto, “A feedback framework for improved chord recognition based on NMF-based approximate note transcription,” International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.196–200, 2015.

[16] Y. Ueda, Y. Uchiyama, T. Nishimoto, N. Ono, and S. Sagayama, “HMM-based approach for automatic chord detection using refined acoustic features,” International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp.5518–5521, 2010.

[17] S. Raczynski, E. Vincent, F. Bimbot, and S. Sagayama, “Multiple pitch transcription using DBN-based musicological models,” International Society for Music Information Retrieval Conference (ISMIR), pp.363–368, 2010.

[18] S.A. Raczynski, E. Vincent, and S. Sagayama, “Dynamic bayesian networks for symbolic polyphonic pitch modeling,” IEEE Transactions on Audio, Speech, and Language Processing (TASLP), vol.21, no.9, pp.1830–1840, 2013.

[19] S. Sigtia, E. Benetos, and S. Dixon, “An end-to-end neural network for polyphonic piano music transcription,” IEEE Transactions on Audio, Speech, and Language Processing (TASLP), vol.24, no.5, pp.927–939, 2016.

[20] S. Sigtia, E. Benetos, S. Cherla, T. Weyde, A. Garcez, and S. Dixon, “An RNN-based music language model for improving automatic music transcription,” International Society for Music Information Retrieval Conference (ISMIR), pp.53–58, 2014.

[21] A.T. Cemgil and O. Dikmen, “Conjugate Gamma Markov random fields for modelling nonstationary sources,” Independent Component Analysis and Signal Separation, pp.697–705, Springer, 2007.

[22] M. Davy and S.J. Godsill, “Bayesian harmonic models for musical signal analysis,” Bayesian Statistics, vol.7, pp.105–124, 2003.

[23] V. Emiya, R. Badeau, and B. David, “Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle,” IEEE Transations on Audio, Speech, and Language Processing (TASLP), vol.18, no.6, pp.1643–1654, 2010.

[24] C. Schörkhuber, A. Klapuri, N. Holighaus, and M. Dörfler, “A Matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution,” Audio Engineering Society Conference, pp.2–5, 2014.

[25] D. Fitzgerald, “Harmonic/percussive separation using median filtering,” International Conference on Digital Audio Effects (DAFx), pp.1–4, 2010.

[26] M. Johnson, “Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure.,” Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL), pp.398–406, 2008.

[27] D. Mochihashi, T. Yamada, and N. Ueda, “Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling,” Proceedings of the 47th Annual Meeting of the Association of Computational Linguistics (ACL)Association for Computational Linguistics, pp.100–108 2009.

[28] T. Taniguchi and S. Nagasaka, “Double articulation analyzer for unsegmented human motion using Pitman-Yor language model and infinite hidden markov model,” IEEE/SICE International Symposium on System Integration (SII)IEEE, pp.250–255 2011.