情報処理学会論文誌 Vol.60 No.11 2034–2047 (Nov. 2019)

楽譜情報を用いた 高時間分解能Audio-to-MIDI変換

保利 武志[1,a)] 中村 和幸[1,2,b)] 嵯峨山 茂樹[1,2,c)]

受付日 2019年2月1日, 採録日 2019年9月11日

概要：本稿は，音楽音響信号に対し，高時間分解能で音楽特徴量（オンセット時刻，音長）を詳細に推定し，MIDI 信号へと変換する手法について述べる．多重音からなる音楽音響信号に対する音楽特徴量の抽出は音楽情報科学分野においてきわめて重要な要素技術であり，演奏解析だけでなく自動演奏や作曲など幅広い音楽活動に寄与できる．特に演奏者の個性（演奏モデル）を獲得するためには，微細な演奏変化をとらえうる高時間分解能な解析が必要とされており，またビッグデータ活用も視野に入れる場合，実演奏に対する詳細な解析の自動化が望まれている．本研究では，演奏者は楽譜どおりに演奏しているという仮定のもと，詳細解析のための多重時間分解能解析手法を提案する．音楽音響信号に対し高周波数分解能にチューニング補正および楽譜情報全体とのマッチングをとり，そのタイムアライメント情報を事前情報として，フレームレベル，波形領域レベルへと時間分解能を上げながらリファインする．また，楽譜との詳細アライメントと音高パターンから逐次解析範囲を限定し，打鍵された音の基本周波数を分離可能な周波数分解能下限を計算することで，リファインされた音楽特徴量を事前分布パラメータとしたベイズモデルによる高時間分解能な音楽特徴量抽出を実現する．評価実験の結果，和音打鍵時のオンセット時刻の揺れの検出に有効であること，また音長に対してもロバストな推定が可能であることが示され，本稿で提案する多重時間分解能解析に基づく演奏の詳細解析への有効性が示唆された．

キーワード：演奏詳細解析，Onset detection，Auto-tuning，HSMM，NMF，階層ベイズ

High-Time-Resolution Audio-to-MIDI Conversion Exploiting Music Score Information

Takeshi Hori[1,a)] Kazuyuki Nakamura[1,2,b)] Shigeki Sagayama[1,2,c)]

Received: February 1, 2019, Accepted: September 11, 2019

Abstract: In this paper, we discuss a method for score-informed audio-to-MIDI conversion, which is aimed at high-time-resolution analysis for music performances. High-time-resolution analysis of music audio signals is expected to be useful for modeling music performance, which is required in automatic performance rendering systems, automatic composition systems, and music information retrieval. Performed music is characterized by fine deviations of music features (changing tempo, asynchronous chord onset timings, varying dynamics and note durations) from the written score, with which musicians increase the expressiveness of their performance. We present a method for precise multi-pitch analysis which is based on the assumption that a music performer plays notes according to a musical score. The algorithm first coarsely estimated note onset times based on matching with the score and then uses this time alignment information as a prior to precisely estimate onsets from high time-resolution spectra and the waveform domain. Based on the music score and its precise time alignment, we can determine the lowest frequency-resolution that still allows to discern the lowest notes that occur at a given position in the music piece, which in turn allows to compute the highest usable time-resolution. The obtained onset and alignment information is used as prior in a Bayesian model to estimate music performance features in high time-resolution. Experimental results showed the effectiveness of our proposed approach to precise music performance analysis, especially, the asynchronous chord onset timings can be also identified in high time resolution.

Keywords: Precise music analysis, onset detection, HSMM, LSAR, NMF, hierarchical bayesian inference

1. はじめに

多重音からなる音楽音響信号に対する音楽特徴量（オンセット時刻，音長）の抽出は音楽情報科学分野においてきわめて重要な要素技術であり，演奏解析や自動演奏，自動採譜など様々な場面での活用が期待されている．特に，近年では計算機の能力向上や情報発信の容易さから個人レベルでの需要も増え，人間と計算機とのセッションや演奏・作曲支援など，音楽活動における人間の新たなパートナとして活用されることにより，これまでにない新たな芸術活動領域も現実味を帯びている．

計算機との協調演奏実現に向け，演奏者の個性を表現する演奏モデル獲得や演奏解析のためには，実演奏の特徴解析が必要であり，これまでにも音楽構造に基づく強弱やアーティキュレーション，テンポ変化などを推定・予測する様々な提案がなされている [1], [2]．しかし熟練した演奏者はより局所的な，個々の音や他の音との響き方に対しても意識しており，代表的な例としては同時音が続くような楽譜に対して，よりメロディ部分の音を強めに，かつ和音としての響きが崩れない範囲で若干早めに弾くことにより，聴取者に対して主題を明示することがある（c.f. 先行音効果 [3], [4]）．すなわち，演奏モデル獲得のためには，和音における個々の音のオンセット時刻揺れを検出できる時間分解能解析が必要であり，本研究ではこれを演奏詳細解析と称して主要課題に位置付けている．一般に MIDI ピアノをはじめとした特殊な機材を使用することができれば，実際に演奏された音高やオンセット時刻などの詳細を容易に得ることは可能であるが，近年のビッグデータによる学習モデルの有効性に鑑みると，大量にデータが存在する録音 CD など音楽音響信号に対して解析可能であることが望ましく，本研究でも解析対象は WAVE 形式から得た音楽音響信号としている．

2. 関連研究

2.1 先行研究

音楽音響信号に対する多重音分離に関する研究は自動採譜や音源分離などの研究分野において特に盛んに議論されており，実用に向け様々な手法が提案されてきた．対数周波数領域において調波構造パターンをモデル化することによる逆畳み込み手法 [5], [6] や，周波数領域における音高情報をマルチエージェントシステムにより MAP 推定する手法 [7]，スペクトルの時間周波数構造を加算重畳した調波時間構造化クラスタリング（HTC）による多重音信号の分離手法 [8] などはその実用的実現に向けた代表例である．また近年では非負値因子行列分解（Non-negative matrix factorization; NMF）[9] や PLCA（Probabilistic latent component analysis）を用いた，各音高に対応する基底スペクトル行列とその強度を表すアクティベーション行列とに分解する手法 [10], [11] や，隠れセミマルコフモデル（Hidden semi-Markov model; HSMM）を用いた確率モデルをベースとした手法 [12]，さらには深層学習ベースとして，再帰構造を有する深層学習（Recurrent neural network; RNN）を用いた手法 [13], [14], [15] などによる高精度な推定の実現例も報告されており，今なお様々なアプローチから研究が続けられている．

以上のような多重音分離のモデルは詳細解析においても適用可能であるが，一方で音楽音響信号に対するすべての音高，オンセット時刻，音長など多くのパラメータ推定が要求されるため，音符イベントやオクターブ誤推定（特に基底スペクトルを用いる手法では，楽器音の調波構造の特性上オクターブ誤推定が起きやすい傾向にある）などが影響し，ノートオンフレーム区間（音が鳴っていると推定された区間）の適合率（precision）や F 値（F-measure）などは近年の研究手法でもおおむね 70%前後であり，詳細データベースとしてはまだ実用可能な精度には至っていない．

また，出現する音高が未知ですべての音高を分離可能な周波数分解能を考える場合，ピアノ 88 鍵盤を考えると，最も低い音である A0，B0 の基本周波数差分 1.6 Hz を分離可能な周波数分解能を保証する必要がある．しかし，フーリエ変換における周波数分解能および時間分解能は不確定性原理の影響下にあることが知られており，理想的には前述の周波数分解能と，出現するすべての音価（実際には演奏される音長）の連続音を分離可能な時間分解能（たとえば Tempo 120 の十六分音符は 0.125 sec の音長であり，その連続する音符列を分離）が必要とされるが，これらのトレードオフな関係が詳細解析のボトルネックとなっている．[16] ではこの問題に対し，高周波数分解能（低時間分解能）なスペクトログラムと高時間分解能なスペクトログラムの 2 つのデータを並列に用いて相互に参照する多重時間分解能解析モデル NMF（並列 NMF，Concurrent NMF）を提案しており，このトレードオフ関係の緩和が報告されている．

一方で，楽譜情報を既知として明示的に扱うことで，音高推定に掛かるコストを抑え高精度な推定を可能とした研究もある．文献 [17] はピアノ音源に対し，MIDI と Audio データを WTW（Windowed time warping）で得られたアライメントを利用し NMF を用いて採譜を行った．文献 [18] も同様に楽譜情報を併用し，複数音源の分離を目標として NMF をベースに．また畳み込みネットワーク

[1] 明治大学大学院先端数理科学研究科
Graduate School of Advanced Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan
[2] 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan
[a)] hori@meiji.ac.jp
[b)] knaka@meiji.ac.jp
[c)] sagayama@ieee.orz

(Convolutional neural network; CNN) を用いた [19] など深層学習によるものも報告されている．以上のような多くの手法では 50–100 msec 前後の範囲において正確なノートオンイベントを推定する手法が述べられているが，和音における個々のオンセットの揺らぎを可能な限り検出するための時間分解能向上に関する議論は不十分であった．

これらの手法に対し，我々もまた楽譜情報を明示的に扱うが，そのうえで不確定性原理に基づく周波数・時間分解能のトレードオフな関係の解消による高時間分解能解析の実現に向け検討を行ってきた．特に，文献 [16] の多重時間分解能解析に着目し，楽譜情報を陽にモデル化した多重時間分解能解析ベースな詳細解析手法を提案した [20], [21]．これらの手法は一定の高精度推定を実現する一方で，楽譜追跡で得られたタイムアライメントに大きく依存しており，より精緻な事前情報が求められていた．

2.2 提案手法

従来の確率モデルによる手法では，主として観測系列（パワースペクトログラム）全体に対する尤度最大化に基づく定式化がなされているが，演奏詳細解析では個々の和音に焦点を当てた解析が必要であることを考えると，系列全体に対するある程度の “もっともらしさ” を確保しつつも，和音単位で局所的に解析することが望ましいと考える．ここで楽譜情報が既知で，和音のオンセット時刻がおおよそ推定できているのであれば，そのオンセット（からオフセットまで）が起こりうる解析範囲を限定しながら局所的に解析することが可能となる．また，解析範囲を限定することができれば，楽譜に存在する広範囲な音高ではなく，限定的な出現音高に絞った最低限分離のために必要な程度の周波数を確保することで，最大限の時間分解能を適用可能である．さらには，データ量を制限できるために，データサンプル点単位（Waveform domain）での解析も効果的に行うことが可能となり，実質的な時間分解能の向上を期待できる．

すなわち，まず必要とされるのは楽譜の音高パターンと音楽音響信号とのマッチング（楽譜追跡）である．そのためには周波数分解能を大きくとることでノートオンイベントの時系列を可能な限り正確に推定する必要があり，必要な時間分解能の確保は難しい．一方で，楽譜–音響信号間のおおよそのタイムアライメントが得られたならば，音高情報ではなくオンセット推定に焦点を当てた高時間分解能解析が可能となる．また，高精度なオンセット時刻（やテンポ）が得られたならば，先の楽譜追跡結果をリファインし個々の和音ごとに焦点を当てた解析も可能となる．

図 1 に本稿で提案する詳細解析の概要を示す．まず観測される音楽音響信号に対し，CQT を用いて対数周波数基準のパワースペクトログラムを得る．次に，全体的なピッチのずれが生じている場合推定に支障をきたすため，周波数のチューニング補正を行い（3.1 節），補正分ずらしたパワースペクトルに対し楽譜との系列全体に対するアライメントを取る（3.3 節）．

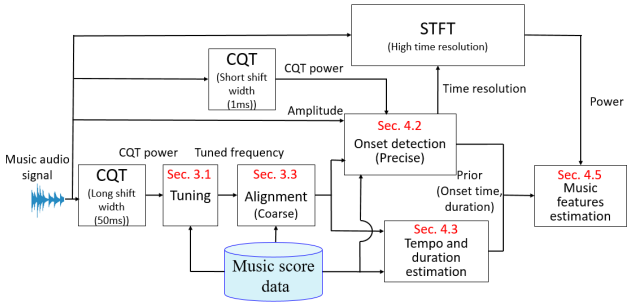

図 1 詳細解析ブロック図．長フレームシフト CQT スペクトルからチューニングを行った後に和音単位の “粗い” 楽譜追跡結果を得る．次に，そのタイムアライメントと短フレームシフト CQT スペクトルを用いて詳細なオンセット時刻および（テンポと）音長を得る．これらの情報を事前分布パラメータとして用い詳細な音楽特徴量を推定する

Fig. 1 A block diagram of high time-resolution analysis of performed music. First, low time-resolution CQT is used to obtain coarse time alignment (coarse onset times), followed by more precise onset detection using CQT with higher time-resolution. Finally, music features are estimated using high time-resolution STFT.

続けて，楽譜追跡で得られたタイムアライメント結果を用いて，より正確なオンセット時刻を再推定する．このとき，ピアノのオンセット時には（音高によらず）パワーの変化が顕著に見られることを利用して，細かいシフト幅による CQT や波形領域での変化点検出手法により，精緻なオンセット時刻を得る（4.2 節）．また，オンセット時刻から得られるオンセット間の時間（Inter-onset interval; IOI）からテンポを推定し，ノート情報（オンセット時刻と音長）を補正する（4.3 節）．

最後に，可変長解析窓（和音単位のオンセットから最長オフセットまでの範囲を抽出）を用いて，出現音高の基本周波数を十分に分離可能な窓幅による STFT を行い，和音ごとの解析を行う．このとき先の解析で得た推定オンセット時刻と推定音長を事前分布パラメータとしたベイズモデルを用いることで，スペクトル情報を考慮した詳細なオンセット時刻および音長を推定する（4.5 節）．

本研究の主たる貢献は，和音単位のオンセット揺らぎ抽出の必要性を提起し，求められる高時間分解能解析を実現するために，音楽音響解析のタスクを音高焦点（高周波数分解能）とオンセット焦点（波形領域）の 2 つに分け，さらにそれらを再統合（高時間分解能）した多重時間分解能解析手法の提案し，その有効性を示した点にある．

なお本稿ではピアノ演奏によるクラシック曲を対象とし，MAPS Database [22] の MAPS_ENSTDkCl および MAPS_StbgTGd2 の一部を解析対象および正解データと

して用いた．また本稿では多重時間分解能解析に焦点を当て演奏ミスの影響を排除するため，楽譜情報の代替として，サンプル曲を三十二分音符単位で量子化してオンセット揺らぎを除去し，音価については前後の和音との順序が変わらない範囲で一様乱数で補正し，テンポについては一定間隔で ±50 の範囲で一様乱数を用いランダムに揺らして滑らかに繋ぎ生成した MIDI データを適用した．

3. 実演奏と楽譜情報間のタイムアライメント

3.1 チューニング補正モデル

実演奏録音データに対する解析においては，録音環境や音源などによってチューニングがずれていることもあり，チューニング補正の有無は推定精度に影響しうる．

一般にピアノにおける単音のスペクトルは，基本周波数とその倍音にピークが立つことが知られている．和音はその重ね合わせによる結果として近似可能であるとするならば，音楽音響信号のパワースペクトル系列に対し，時間方向に総和をとって得られた対数周波数軸におけるスペクトルは，最も低い音高（本研究では音高 A0（= 27.5Hz））および半音ごとにピークを持つような櫛型形状をとると考えられる．ただし，3 倍音以上の周波数は平均律からずれていることがよく知られており，また実際の演奏データにはノイズなども含まれるが，倍音成分は基本周波数から離れるほど減少する傾向にあるため，本研究では先の仮定に基づき櫛型形状モデル化による検証を行った．なお，本稿では櫛型形状関数として混合ガウスモデル（Gaussian mixture model; GMM）を用いて近似した．

3.2 チューニング補正定式化

時間方向に総和を取った対数周波数スペクトルを $\mathbf{y}_c$，パラメータ $\boldsymbol{\Theta}$ を持つ櫛型形状近似関数を $\mathcal{T}_c(\boldsymbol{\Theta})$ とする．櫛型関数 $\mathcal{T}_c(\boldsymbol{\Theta})$ は，そのエネルギー V_ξ と，音高 A0 の基本周波数（27.5 Hz = 2,100 cent 基準）および半音（100 cent）ごとにピークを持つ GMM による形状関数との積としてモデル化できる．したがって，混合数を $M = 88$（一般的なピアノの鍵盤数に対応），周波数のずれの差分を ξ，A0 の基本周波数を μ_0，各ガウシアンの分散を σ_ξ^2 とし，対数周波数にあたる cent の値を c とすると，

$$\mathcal{T}_c(\boldsymbol{\Theta}) = V_\xi \sum_{m=0}^{M-1} \pi_m \mathcal{N}(c|\mu_0 + 100m + \xi, \sigma_\xi^2)$$
$$s.t. \sum_{m=0}^{M-1} \pi_m = 1, \tag{1}$$

となる．ここで，$\sigma_\xi = 25$ cent とした．以上より，

$$\arg\min_{V_\xi,\pi,\xi} \sum_c |\mathbf{y}_c - \mathcal{T}_c(\boldsymbol{\Theta})|^2 \tag{2}$$

として定式化することで，チューニング補正差分 ξ を求められる．式 (2) から，

$$V_\xi = \sum_c \mathbf{y}_c$$
$$\pi_m = \frac{V_\xi - V_\xi \sum_c \sum_{l\neq m} \pi_l \mathcal{N}(c|\mu_0 + 100l + \xi, \sigma_\xi^2)}{\sum_c \mathcal{N}(c|\mu_0 + 100m + \xi, \sigma_\xi^2)}$$
$$= 1 - \sum_c \sum_{l\neq m} \pi_l \mathcal{N}(c|\mu_0 + 100l + \xi, \sigma_\xi^2) \tag{3}$$

が得られる．下式はあらかじめ $\mathbf{y}_c$ を正規化した場合 $V_\xi = 1$ となることを利用した．チューニング差分 ξ は EM アルゴリズムなどを用いて求めることもできるが，予備実験において 6.25 cent 程度のずれ（半音 16 分割）ではオンセットおよび音長推定精度に影響がなかったため，ξ を 6.25 cent ごとにグリッドサーチを行うことで最尤推定の枠組みを用いることなく最適な値を得た．

また，π_m について，（半音ごとにピークを持つという仮定に基づき）仮にそれぞれのガウシアンが急峻であるとすれば，ある対数周波数 c を含むガウシアンの範囲外ではほぼ 0 と見なすことができ，結果としてそれぞれのガウシアンに含まれる対数周波数に対応する範囲のみの積分で近似できる．すなわち，$\pi'_m = \sum_{c\in\text{m-th gaussian}} y_c$ として，$\pi_m = \frac{\pi'_m}{\sum_l \pi'_l}$ で近似できる．

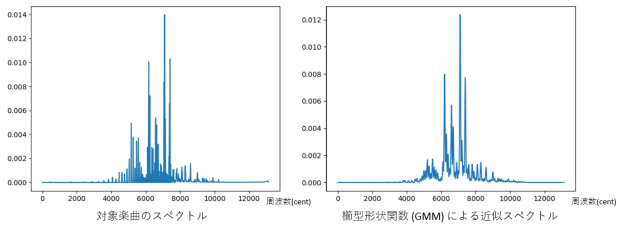

図 2 CQT により得られたパワースペクトル系列（対数周波数軸）に対し，時間方向に総和を取り正規化したスペクトル（左）と櫛型形状関数（GMM）による近似例（右）．解析には MAPS-StbgTGd2 中の楽曲の一部を用いた

Fig. 2 Examples of a piece's normalized power spectrum (left) and estimated tandem function (right), which is approximated using a Gaussian mixture model. Both graphs result from summation over a time frame and are displayed over a logarithmic frequency axis.

図 2 は実際の演奏に対し，CQT により得られた対数周波数軸によるパワーを時間方向に総和をとったスペクトル（左図）と，GMM で近似（右図）した例である．元のスペクトルに見られる複数のピークを GMM においてもとらえることができており，先の仮説はチューニング補正機能において適用可能であると推察できる．

以上の手法を用いて，International Piano-e-Competition [23] の MIDI 楽曲に対し半音の周波数ビン範囲内において一様乱数を用いランダムにピッチ全体を変化させ Cubase（HALion Sonic SE）を用いて WAVE 音源としたものを 15 曲，MAPS Database より 5 曲（ただしこちらはチューニングずれがないものとした）に対し実

験した結果，前者 15 曲についてはすべて正解し，後者については 1 曲周波数 1 ビン分負の方向のずれ（−6.25 cent）として検出されたが，その他はずれなし（0 cent）という結果を得，本手法の有効性が示された．

3.3 楽譜モデル

ある音楽音響信号がピアノの楽譜に基づき演奏された結果であるとするならば，各時刻のパワースペクトルは楽譜に記載された音高情報に基づき生成されたものであると仮定できる．また各単音のスペクトルを表現する関数がモデル化可能ならば，和音表現は単音の重ね合わせ，すなわち加法的に和音のスペクトルを近似できる（ただし厳密にはパワースペクトルの加法性は成り立たない）．

したがって，四分音符や八分音符といった音価に対応する音長を継続長確率，音高遷移（音高パターン）を遷移確率として表現するならば，音楽音響信号は和音スペクトルを期待値とした多変量正規分布を出力確率密度として持つ隠れセミマルコフモデル（HSMM）として近似できる．

ここで，短時間フーリエ変換（STFT）により得られるパワースペクトルを直接観測系列とした場合，次元が大きくなり次元の呪いの影響を過度に受けるため，適切に次元削減を行う必要がある．音楽音響信号において最も容易かつ代表的な手法としては 12 次元のピッチクラスに集約した Chroma vector [24] が考えられるが，オクターブ情報を極力保存することを考えると，一般的なピアノにおける最大出現音高数 88 次元を確保することが望ましい．したがって，本研究ではまずチューニング補正により得られた周波数差分に応じてピッチを補正し，続けて CQT により MIDI ノート番号に対応した周波数ビンを持つ 88 次元ベクトルを得，これを観測系列（入力）とした．

3.4 楽譜モデルの定式化

図 3 に本研究における楽譜モデルを示す．チューニング補正済み CQT で得られる各時刻ステップごとの観測データに対して，音量変化にロバストにするために正規化したベクトルを，3 状態 left-to-right でモデル化された単音 HSMM のいずれかの状態におけるスペクトル期待値の和で表現される正規化された和音ベクトルから生成されたものであると見なし，これを多変量正規分布でモデル化する．また，分散共分散行列は対角要素に定数 0.01 を持つ対角行列とした．なお単音 HSMM はあらかじめ MIDI で作成した各単音に対し，Cubase の HALion Sonic SE を用いて WAVE 音源として生成し事前学習して得られたパラメータを用いた．

また継続長確率は対応する音の音価に等しい音長（たとえばテンポ 120 の四分音符であれば 0.5 秒）を期待値とし，同様に MIDI から計算した音長分散を適用した正規分布でモデル化し，音高間の遷移確率は楽譜から得られる音高パターンに基づき，（隣り合う）次状態に対してのみ 1，それ以外を 0 として与えた．以上より，観測データを $\mathbf{Y}$，状態 i, j 間の遷移確率を a_{ij}，時刻 t における状態 i からの出力確率を $b_i(\mathbf{Y}_t)$，状態 i において継続長が d である確率を $p_i(d)$ とすると，時刻 t，状態 j，継続長 d における局所的な最尤確率 $\delta_t(j,d)$ は，

$$\delta_t(j,d) = \max_{i \in S_{\neq j}, d' \in D} \delta_{t-d'}(i,d') a_{ij} p_j(d') \prod_{s=t-d'+1}^{t} b_j(\mathbf{Y}_s). \tag{4}$$

なお，S および D はそれぞれ各状態，継続長集合を表す．式 (4) を用いた Viterbi アルゴリズムにより，和音単位における音楽音響信号と楽譜とのタイムアライメントが得られる．なお，本研究では継続長 d の最大フレーム数を 1.5 sec として与えた．

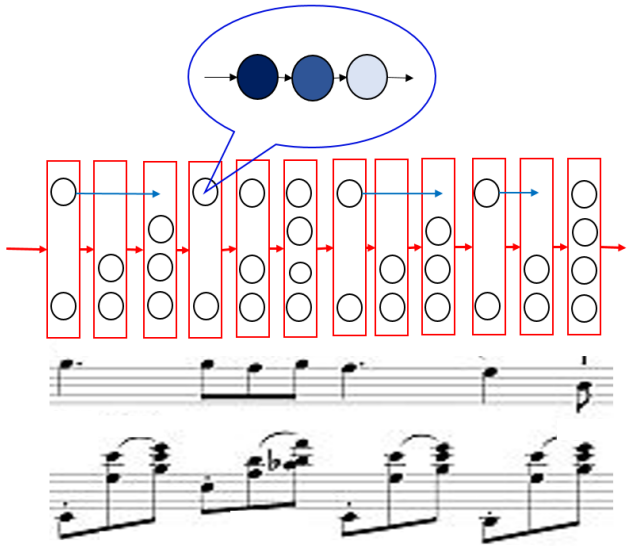

図 3 HSMM に基づく楽譜モデル．楽譜情報は原則として四分音符や八分音符などの音価を単位として量子化されており，和音は単音の重ね合わせとして近似できる．上図の白丸はそれぞれ単音 HSMM を示し，赤枠は和音単位，青矢印は複数の状態にまたがって音長が継続していることを示している．本研究ではさらにそれぞれの単音（白丸）を三状態（3 状態 left-to-right）でモデル化している

Fig. 3 The music score model based on HSMM. Single tones (white circles) are modeled with three states (left-to-right) and organized in approximated chords (red rectangles). The long-duration notes across multiple states (blue arrow).

4. 演奏詳細解析

4.1 高時間分解能解析に向けて

前章の手法で得られた楽譜追跡結果を事前情報とし，時間分解能を上げてリファインすることを考える．

ピアノ演奏では新たなオンセットが入らない限り，（微小区間のビブラートや反響などを考慮しなければ）音高に依らず音量は単調減少する．したがって，ピアノ音源にお

いてはパワー変化に着目することでより正確なオンセット時刻を得ることが期待できると同時に，正確なオンセット時間が分かれば，IOI と楽譜情報とを比較することでテンポ変化を推定できる（たとえば楽譜再現 MIDI のテンポが 120 で IOI が 0.5 秒である場合，これに対し観測される IOI が 1.0 秒であった場合のテンポは 60 だと推測できる）．

ここで，演奏者は原則楽譜に示された音価に従い演奏すると仮定すれば，その IOI 情報に基づき次時刻のテンポ変化を予測することで，現在着目しているオンセット周辺の音長の修正および，次のオンセットが配置されている時刻の推定が可能となる．さらには，詳細オンセット推定により実演奏と楽譜間の正確なマッチングができるのであれば，ある時刻の和音のみに着目した局所的な高時間分解能音楽特徴量解析を逐次的に行うことも可能となる．

4.2 パワー量に基づくオンセット推定

オンセット推定に関する手法として，NMF により得られる推定アクティベーションに対する閾値処理などを行う周波数領域の特徴量に着目した手法のほかに，波形領域や位相情報などを用いた手法も提案されている [25], [26]．特にピアノ演奏においては，打鍵時のパワーやその変化量に顕著な特徴が見られるため，オンセット推定においてこれらの特徴量を併用することは有効であると考えられる．一方でバイオリンなどの弦楽器に対しては位相変化（位相の連続性）に着目したオンセット検出も報告されているが，本研究では位相特徴量を用いた際の有意性を確認できなかったため除外した．

また音楽演奏（音楽音響信号）を，新しい音が追加されるたびに（時系列的に）確率構造が変化する確率過程であるととらえるならば，その構造が変化する点を検出することで，波形領域でのオンセットを推定することが可能となる．そこで，本研究では波形領域の定常信号に対し赤池情報量基準（AIC）で確率構造変化点を検出する局所定常 AR モデル（Local stationary autoregressive model; LSAR）を用いたオンセット時刻推定を提案する．一般に AR モデルではデータがある分布に従うことを仮定し，（時）系列データの予測分布を求めることで推定を可能とするが，LSAR では何らかの要因が付加されたときにその分布形状が変わりモデル評価指標である AIC が低下することを利用して変化点を検出できる．

以下に本章のオンセット推定で用いた特徴量をあげる．

- Amplitude：$AMP(t) = |x(t)|$
  本研究ではフレームごとの期待値を用いた．
- Spectral power：$POW(t) = \sum_{\omega} |X(\omega, t)|^2$
  ここで，$|X(\omega, t)|$ は CQT で得られるパワーを示す．
- Spectral difference：
  $SD(t) = \sum_{\omega} (H(|X(\omega, t)| - |X(\omega, t-1)|))^2$
  ここで，$H(x) = \frac{x+|x|}{2}$．[26]

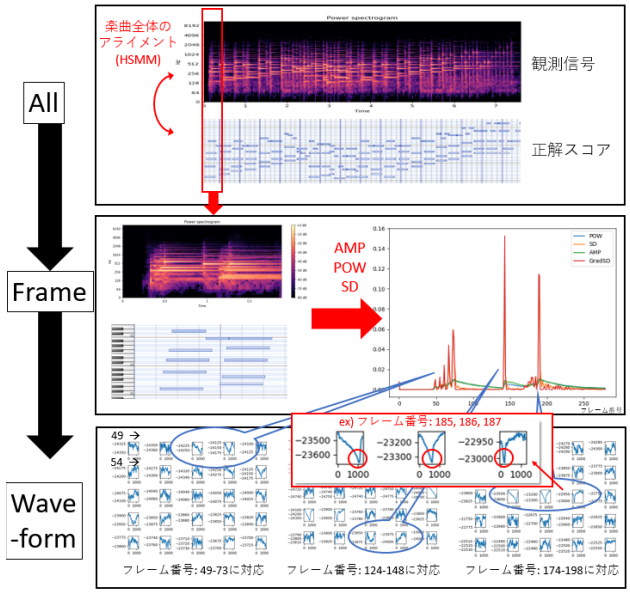

図 4 オンセット時刻推定における多重時間分解能解析の概要．上段は楽曲全体に対し和音単位で粗くアライメントを取る高周波数分解能解析．中段はそのアライメント結果を事前情報としてパワーやその変化量などを用いてフレーム単位での解析を示し，下段は LSAR を用いた波形領域レベルによる高時間分解能なオンセット時刻推定を示す

Fig. 4 For onset detection, the audio data is first aligned using the HSMM model. Next, onset deviations are estimated in spectrum frame resolution. Onset estimation precision is then increased using LSAR.

- GradSD：
  AMP，POW，SD を統合した特徴量（後述）
- AIC（LSAR モデルにより計算）

図 4 はその概要および解析例であり，上から順に解析範囲を制限しながら時間分解能を上げていく様子を示している．最上部のレベルでは 3 章で述べた HSMM による実演奏音楽音響信号全体と楽譜全体どうしのタイムアライメント，すなわち，ここでは系列全体に着目した高周波数解能解析による楽譜追跡が行われる．

中段は数フレームのレベルに着目してオンセット時刻を推定する様子を示す．最上部の HSMM で得た和音単位における楽曲全体のアライメントを事前情報とし，シフト幅を狭めてより正確なオンセットイベントを探す．中段右図は先に示した特徴量中，AIC を除いた 3 種の特徴量（AMP，POW，SD）とそれらを統合した特徴量（GradSD）が示されている．AMP（緑線），POW（青線）を確認すると，和音の（最後の）オンセット位置にピークが来てなだらかに減少する形状をとっており，微細なオンセット揺れの検出には不向きだが和音単位のイベント検出としては有用であることが分かる．

フレーム間におけるパワーの変化量である SD（黄線）は微細なオンセットを検出するのに有用であり，図中のフレーム番号 50 付近や 150 付近のオンセットは明確であ

る．一方で正解スコアと比較すると分かるように，180 から 200 付近のオンセットは大きなピークを持つ 1 点を除いて分離は困難であることが分かる．そこで我々は新たに AMP，POW，SD の 3 つの特徴量を融合した GradSD を導入した．これは，AMP や POW が和音中最後のオンセットを中心になだらかに減衰することを利用し，AMP および POW の変化量（傾き）が正であれば SD に 1 以上の正の重み（本研究では各時刻の AMP と POW の積に 10 を乗じた）をつけ，傾きが負であれば負の重み（同様にして −10 を乗じた）を掛けることにより，オンセットを強調するものである．

最下段ではさらに LSAR モデルを用いて，波形領域でのオンセット時刻推定を行っている．最下段の 5 × 5 の画像はそれぞれが LSAR モデルにより計算された AIC の軌跡であり，それぞれの図は左上から右下に向けて順に，3 フレームシフト分（3 msec）のデータ点を抽出して LSAR モデルを適用した結果を示している．図中の拡大した 3 つの AIC 推移（フレーム番号 185–187）を確認すると，新しいオンセットが入ると AIC の値が極端に減少し，その該当サンプル点がシフト幅（1 msec = 441 サンプル点）程度移行していることが見て取れ，LSAR の有効性を示唆している（一方で新規オンセット箇所以外（たとえばオフセットから無音区間に入る箇所でも減衰する傾向が見られた）でも減衰が生じることもあり，LSAR は単品ではなく，ある程度オンセット位置を推定し範囲を限定したうえで適用すると効果的に活用できる）．AIC 軌跡はメディアンフィルタ（本研究では 15 次とした）で平滑化し，着目しているフレームに対応した軌跡のピークを検出することで，より正確なオンセット時刻を検出できる．実際にシフト幅 1 msec で解析し，前後 1 フレーム内に正解オンセットがあるフレームを適切に推定できたデータに対し，さらに LSAR でどの程度オンセット推定ができているかを調査した結果，正解オンセット時刻（MAPS Dataset は 0.01 msec の分解能でオンセット時刻を記載したファイルが付属しておりそのデータを Ground truth とした）から ±0.5 msec の範囲内にピークを検出したものが 76.2%であったことから，詳細なオンセット推定に有効であることが示唆された．

なお本稿では和音解析にフォーカスするために基本演奏ミスがなくすべてのオンセットが存在するとしているが，たとえば簡易的には本手法においてピーク検出に閾値を設け，楽譜データと照合することでオンセットの false negative および false positive を推定することも可能である．

4.3 テンポ推定

楽譜上の音価が既知であることから，正確なオンセット時刻と対応する音価を得ることができれば，和音間の IOI を用いてテンポ推定することができ，HSMM で得たマッチング結果のリファインが可能となる．ここで，テンポ変化は滑らかかつ連続的であり，急激な変化は原則起こらないと仮定すると，ある時刻 t におけるテンポ $M(t)$ は二階差分と分散 σ_m^2 を持つノイズ ϵ_m を用いて，

$$M(t) = 2M(t-1) - M(t-2) + \epsilon_m,\ \epsilon_m \sim \mathcal{N}(0, \sigma_m^2), \tag{5}$$

また，楽譜情報に基づく $n-1$ 番目と n 番目の和音における IOI を $\hat{I}(n)$，時刻 t におけるテンポを $\hat{M}(t)$ とすると，同様に分散 σ_I^2 を持つノイズ ϵ_I を用いて，

$$I(n) = \frac{\hat{M}(t)\hat{I}(n)}{M(t)} + \epsilon_I,\ \epsilon_I \sim \mathcal{N}(0, \sigma_I^2), \tag{6}$$

と定式化できる．式 (5) および式 (6) は状態空間モデルと見なすことができるので，テンポ $M(t)$ の予測および事後パラメータは粒子フィルタ（Particle filter; PF）を用いて計算できる．本研究ではハイパーパラメータも同時に最適化可能な自己組織化型状態空間モデル [27] に PF を適用した．

図 5 は実際の楽曲（Chopin, Preludes op28, no.1）に対し PF を用いて推定したテンポ推移の例であり，楽譜を再現した MIDI（Score-based MIDI）は Classical archives [28] から，実演奏データとして Cubase から HALion Sonic SE を用いて WAVE 変換した International Piano-e-Competition [23] のデータを用いた．上図のオレンジ線は Score-based MIDI で示された IOI，青線は実際に推定したオンセット時刻から得られる IOI である．また，下図オレンジ線は（局所的な）テンポで，前節で示した実演奏データに対する詳細オンセット時刻推定で得た IOI から式 (6) を用いて $\frac{\hat{M}(t)\hat{I}(n)}{I(n)}$ で計算した結果である．青線は Score-based MIDI のテンポ，緑線は推定 IOI 観測後の事後推定テンポを示す．横軸は和音オンセット（実演奏 IOI は和音オンセット中最も早いオンセット時刻のものを採用した）のノートを順に並べたインデックスである．

図 5 で示されるとおり，表情付き演奏では多くの “溜め” や逆に瞬間的に高速な演奏を適宜挟むため，局所的なテンポ変動は比較的大きくなりやすい傾向にある．したがって本研究では外れ値に影響を受けにくいコーシー分布を観測ノイズとして採用しており，PF でも変動に対し過度にフィットしない期待値を得られている．

以上により，着目している時刻のテンポと，楽譜とのアライメント用いて得られる音価を用いて，HSMM で得られていたタイムアライメント情報を以下のように修正する．高時間分解能に解析したオンセット時刻および推定したテンポと楽譜で示される音価により得られる仮のノートオンフレーム（音が鳴っていく区間）から得られた音長を T_{tempo}，HSMM でノートオンとして採用した際のフレーム数に対応した時間長を T_{est} として，任意の定数 a を用いて $aT_{\text{est}} + (1-a)T_{\text{tempo}}$ とし決定した．なお a の値は予備実験から 0.7 とした．

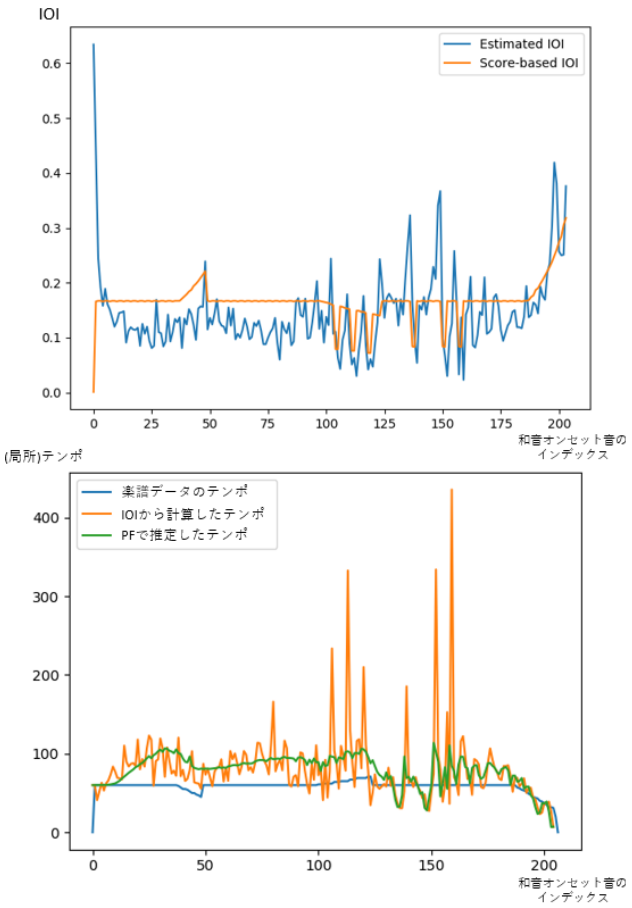

図 5 粒子フィルタ（PF）を用いたテンポ推定の例．上図は Score-based な MIDI から得られる IOI（オレンジ線）および推定したオンセットより計算された IOI（青線）を示し，下図は同様に Score-based MIDI のテンポ（青線）と IOI からのみ計算したテンポ（オレンジ線），そしてノイズなどを考慮して得られたテンポカーブ（緑線）を示す

Fig. 5 An example of tempo estimation using a particle filter. The upper graph displays IOI (inter-onset-intervals), which are used to estimate the tempo shown in the lower graph.

4.4 高時間分解能 STFT と CQT

演奏詳細解析では（楽譜上の）同時音に対する単音レベルでの分離が必要とされるため，可能な限り高い時間分解能で解析する必要がある．しかしフーリエ変換における時間・周波数分解能の不確定性原理から，特に連続する低音を解析する場合，時間分解能を高くするために周波数分解能を下げると，低周波数域が互いに重なってしまい分離が不可能となる．基本周波数ではなく倍音成分などに着目する手法も考えられるが，調波構造は音高や弾き方，楽器や環境などによって変化しやすく，安定した要素の推定は難しいため，ここでは汎用的かつロバストな特徴量である基本周波数を用いた音楽特徴量推定を前提として進める．

周波数分解能 Δf と時間分解能 Δt の不確定性原理は以下の式で表される．

$$\Delta f \times \Delta t \geq \frac{1}{2}. \tag{7}$$

詳細解析において確保すべき周波数分解能は解析範囲における音符集合をすべて分離可能とする分解能であり，解析範囲およびその範囲に出現する音高集合が既知であれば，理論上解析可能限界である時間分解能を陽に計算できる．

仮に解析範囲における音符集合の最下音 2 つが “D3”（146.8 Hz）と “A3”（220.0 Hz）であった場合，その周波数差分は 73.2 Hz であり，周波数分解能はこれを確保すればよい．したがって，式 (7) より，サンプリング周波数 $f_s = 44.1$ kHz と解析サンプル数（解析幅）N を用いて $N = \frac{f_s}{73.2} \simeq 602$ とすることで解析に必要な周波数分解能を決定できる．

一方で CQT を用いることでも高い周波数分解能を保持しながら高い時間分解能解析を可能である．CQT では Q 値を固定させ，解析周波数に応じて窓幅を決定する．特に低周波数域においては周波数分解能を高く（窓幅を大きく），高周波数域においては時間分解能を高く（窓幅を小さく）することで，バランスよく高い精度でパワースペクトルの変化を解析できる．

図 6 は MAPS database の MAPS_MUS-ty_mai_StbgTGd2 の冒頭部分に対しそれぞれ STFT（上図），CQT（下図）をしたものである．図から明らかなように低周波数域，A4（440 Hz）付近から少しずつぼやけ，A3（220 Hz）以下ではオンセットの判別が難しくなる．実際に図 6 の範囲の出現音高の幅が最も小さいのは D3 と A3 で，提案する手法による窓幅が 602 であるのに対し，D3 に対応する周波数における CQT の窓幅は 5,051 であり，時間分解能をより向上させる余地があると考えられる．

そこで実際に推定精度に差が出るのか，**図 7** に時間分解能を変えて解析した例を示す．音楽特徴量解析は 4.5 節で提案する NMF-based モデルによる．オンセット時刻や音長に関するパラメータは前節までの手法で求めたものを使用した．図上段と下段は STFT を用いてそれぞれフレーム長 8,192，602（出現音高の下 2 つの周波数 146.8（D3）と 220.0（A3）から上式で計算）として解析した例，中段は CQT を用いて解析した例である．低い時間分解能によるケースでは推定したオンセット時刻が大きくずれてしまっていることが分かる．一方で CQT と STFT（フレーム長 602）では図では判断つかないが，実際に推定したオンセット時刻を示した**表 1** を見ると，STFT のほうが精度よく推定できていることが分かる．一方で CQT でも比較的高精度な推定が可能であった背景としては，NMF は基本周波数だけではなく倍音などの調波構造や非調波成分を考慮するモデルであることから，高精度にパワーの抽出ができている他の成分の寄与があったものと考えられる．

表 1 オンセット時刻の推定誤差と正解オンセット時刻（MAPS_MUS-ty_mai_StbgTGd2 の冒頭部分）

Table 1 Result of onset time estimation (Begining part of MAPS_MUS-ty_mai_StbgTGd2).

出現音高	50	57	62	66	74
STFT(N=8192)	−0.0703 [sec]	−0.0293 [sec]	−0.0208 [sec]	−0.0323 [sec]	−0.0152 [sec]
CQT	−0.0089 [sec]	−0.0067 [sec]	−0.0063 [sec]	−0.0157 [sec]	−0.0064 [sec]
STFT(N=602)	−0.0017 [sec]	−0.0011 [sec]	−0.0016 [sec]	−0.0027 [sec]	−0.0018 [sec]
Ground truth	0.5283 [sec]	0.5921 [sec]	0.6502 [sec]	0.7082 [sec]	0.7663 [sec]

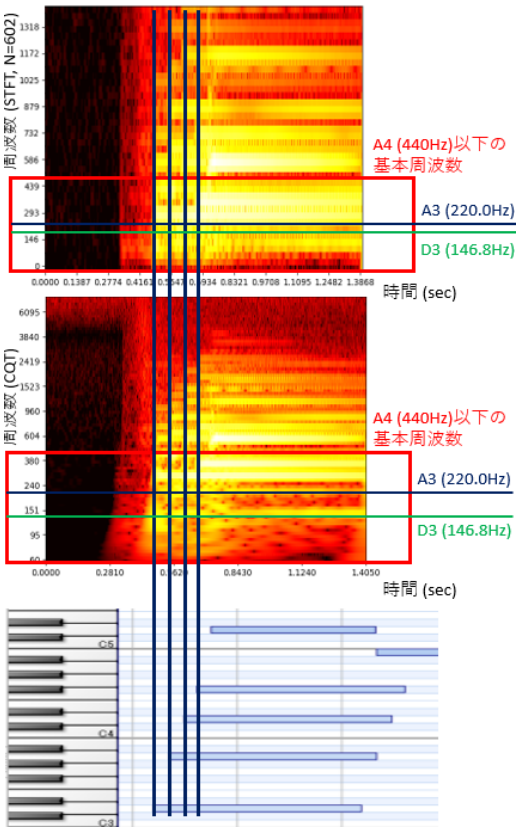

図 6 MAPS database の MAPS_MUS-ty_mai_StbgTGd2 の冒頭部分に対し，上図は窓幅 $N = 602$ で STFT による結果，下図は CQT による結果を示す．特に A3（220 Hz）付近以下において CQT の時間分解能十分ではない

Fig. 6 Results of STFT (with a window size of 602 frames) and CQT of the file MAPS_MUS-ty_mai_StbgTGd2 of the MAPS data set. In case of CQT, the time resolution is especially low below 220 Hz (A3).

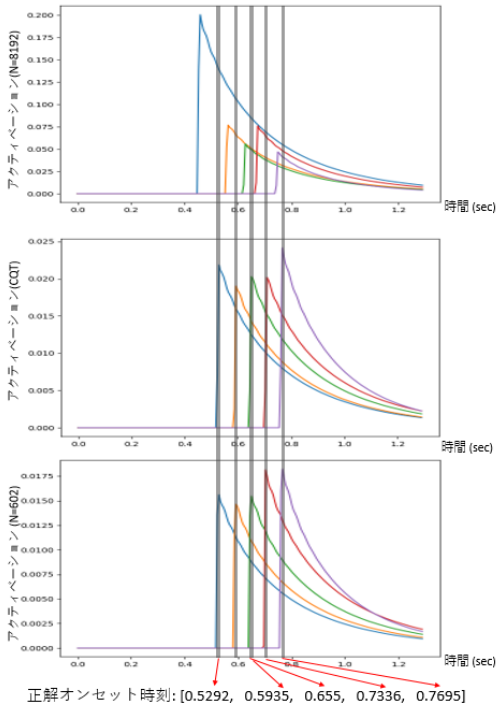

図 7 本稿で提案する音響モデル（4.5 節）で推定したアクティベーション．出現 MIDI ノート番号は（正解の）オンセット順に $\{50, 57, 62, 66, 74\}$．上図はフレーム長 8,192，中央図は CQT，下図は 602 で解析した例．全体にまたがる黒線は正解のオンセット時刻を表す

Fig. 7 Activations estimated using the proposed model (Sec. 4.5). The MIDI note numbers of the observed notes are 50, 57, 62, 66, 74. The upper graph was computed using STFT with a window size of 8,192 frames, and the lower graph with a window size of 602 frames. The graph in the middle was obtained using CQT. The black lines indicate the ground truth onset times.

4.5 音楽スペクトログラムの生成モデル

本節では前節までに得られたパラメータ（詳細なオンセット時刻と音長）を事前知識として活用した音楽特徴量推定手法について述べる．

STFT で抽出したパワースペクトルに対し，以下の生成モデルを考える．パワースペクトログラム（パワースペクトル系列）を非負値行列 $\mathbf{Y} \in \mathbb{R}^{W \times T}$ とすると，各音高のスペクトルパターン（基底行列）$\mathbf{H} \in \mathbb{R}^{W \times K}$ とそのアクティベーション $\mathbf{U} \in \mathbb{R}^{K \times T}$ を用いて，

$$\mathbf{Y} \approx \mathbf{HU}, \tag{8}$$

として，低ランク行列積で近似できるため，任意の距離尺度を $D[\cdot||\cdot]$ として，

$$\mathbf{H}, \mathbf{U} = \arg\min_{\mathbf{HU}} D[\mathbf{Y}||\mathbf{HU}]$$
$$s.t. \ \forall_k \quad H_{w,k} \geq 0, \ U_{k,t} \geq 0 \tag{9}$$

と定式化できる．特に，その距離基準に Kullback–Leibler divergence を適用すると，ポワソン分布を用いて，

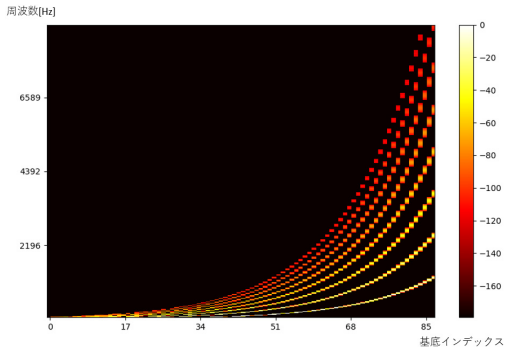

図 8 基底行列テンプレート．基底の数値は log 変換後の値を表示している．基底インデックス番号は MIDI ノート番号と対応しており，基底インデックス番号に 21 を加えた値が MIDI ノート番号と一致する

Fig. 8 The basis matrix template, which is the logarithmized. Each index on the x axis corresponds to a MIDI note number.

$$Y_{w,t} \sim \mathcal{P}o\left(\sum_k H_{w,k}U_{k,t}\right). \tag{10}$$

とした生成モデルと等価となる．

ここで，基底行列 $\mathbf{H}$ の非負値性およびポワソン分布との共役構造を考慮すると，ガンマ分布を用いて，

$$H_{w,k} \sim \mathcal{G}a((b^H_{w,k})^{-1}\bar{H}_{w,k}, b^H_{w,k}), \tag{11}$$

としてモデル化できる．ここで，

$$\mathcal{G}a(x|a,b) = \frac{b^a}{\Gamma(a)}x^{(a-1)}\exp(-bx), \tag{12}$$

である．$\bar{H}_{w,k}$ はガンマ分布の Mode であり，本研究では倍音にピークが立つように，各倍音の周波数位置に個々のガウス分布が配置されるよう設定した GMM による各音高基底のテンプレートを用いた（**図 8**）．なお調波構造を表す各混合重みは，

$$\frac{h(f_n)}{h(f_0)} = (n+1)^{-\alpha}$$

で与えた．f_n は第 n 倍音，f_0 は基本周波数を示し，$h(\cdot)$ はその重みを示す．また，本研究では $\alpha = 1.5$ として第 8 倍音までを考慮した．

アクティベーション $\mathbf{U}$ も同様にモデル化できるが，演奏詳細解析では単音単位での解析が必要とされるため，まず各音高のアクティベーションを単音レベルで分解することを考える．単音のインデックスを r とし，それぞれの単音が持つエネルギーを V_r，ある時刻 t のアクティベーション形状を $\nu_{r,t}$ とすると，

$$U_{k,t} = \sum_r \delta(\kappa_r = k)V_r\nu_{r,t}$$
$$s.t. \sum_t \nu_{r,t} = 1 \tag{13}$$

とモデル化できる．ただし，κ_r は単音インデックス r が対応する音高を示す．

エネルギー V_r は基底 $H_{w,k}$ と同様に非負値性から，

$$V_r \sim \mathcal{G}a(a^V_r, b^V_r), \tag{14}$$

とモデル化できる．

次に単音のアクティベーションの形状 $\nu_{r,t}$ に対して，オンセット推定により得られたパラメータを陽にモデルに組み込むことを考えると，$\nu_{r,t}$ は，理想的にはオンセット時刻にピークが立つようなオンセットピーク分布 $O_{r,t}$ と，音長を表す時定数 τ を用いたパワーの減衰を表現する単音形状分布 $G_r(\tau)$ との畳み込みとして近似できる．本研究ではオンセットピーク分布として，オンセット時刻 s^{on}_r をパラメータとして持つ正規分布を用いて $O_{r,t} = \mathcal{N}(t|s^{on}_r, \sigma^2_{on})$ とし，またパワーの減衰は単調減少であると仮定して，パラメータ λ_r を持つ指数分布を用いて，$G_r(\tau) = Exp(\tau|\lambda_r)$ とし，$\sum_t \nu_{r,t} = 1$ を考慮して，

$$\nu_{r,t} = \int_\tau \frac{\mathcal{N}(t-\tau|s^{on}_r, \sigma^2_{on})}{\sum_{t'}\mathcal{N}(t'-\tau|s^{on}_r, \sigma^2_{on})}Exp(\tau|\lambda_r)d\tau, \tag{15}$$

としてモデル化した．分散 σ^2_{on} は小さくなるほどオンセットの立ち上がりにおけるアクティベーションが急峻となり，理想的にはオンセットをインパルスで表現するため，その標準偏差を 1 msec として設定した．

オンセット時刻 s^{on}_r は前節までに述べた詳細オンセット推定により得られた推定オンセット時刻 s^{est}_r 付近にあると仮定して，

$$s^{on}_r \sim \mathcal{N}(s^{est}_r, \sigma^2_{est}), \tag{16}$$

とした．ただし，オンセット時刻はすでに高精度な値が得られているものであるとして，σ^2_{est} においてもその標準偏差を 5msec として設定した．減衰パラメータは非負値性と楽譜演奏から得られる音長に対応するパラメータ $\bar{\lambda}_r$ が Mode となるよう考慮して

$$\lambda_r \sim \mathcal{G}a((b^\lambda_r)^{-1}\bar{\lambda}_r, b^\lambda_r), \tag{17}$$

としてモデル化した．b^λ_r はどの程度事前に推定した音長に寄せるかを決定するパラメータであり，実験的に繰り返し決定した．また，MIDI で与えられた IOI に対し，WAVE データから推測される実際に音が鳴っている区間（アクティベーションが有効な値を取り続けている区間）は必ずしも一致しない．そこで本研究では，アクティベーション形状が（一方の畳みこまれる正規分布が急峻な形状であるため）ほぼ単純な指数分布の形で与えられる．したがって，指数分布パラメータが直接音長を表現するとして，$\bar{\lambda}_r$ は事前に得られた推定音長 l_r から，指数分布の累積分布関数を用いて，本研究ではヒューリスティックに累積分布関数 $F(x) = 0.9$ となるときを音長だと見なし，

$$\bar{\lambda} = -\frac{\log(0.1)}{l_r}, \tag{18}$$

として与えた．これは，音長 l_r が長くなると，指数分布のパラメータ $\bar{\lambda}$ が小さくなる，つまり指数分布形状の傾きがなだらかになることを意味している．

5. 評価実験

5.1 実験設定

提案モデルによる音楽特徴量の推定精度を評価することを目的とする．実験には MAPS Database より 10 曲，楽譜情報を Classical archives [28] とした実演奏 MIDI [23] を Cubase の HALion Sonic SE で WAVE 変換した曲を 10 曲，冒頭から 30 秒〜1 分程度を抽出して行った．後者の楽曲セットはチューニング機能評価に利用し，40 – 60 cent の範囲でランダムにずらした．また，MAPS Database における楽譜情報となる参照用データは，実験対象とする（実演奏）MIDI ファイルに対して仮想的な（演奏ミスが存在しない）楽譜再現 MIDI データを作成するため，近いオンセット同士を量子化して和音としたうえで，1 章で述べた任意のテンポ変化および音長変化をランダムに与えた MIDI データを作成し適用した．MAPS Database の WAVE ファイルはステレオ音源であるため，Audacity を用いてモノラル信号へと変換した．なおサンプリングレート 44.1 kHz であるため，本研究においてもそのサンプリングレートをそのまま使用した．また，HSMM における解析のシフト幅は 50 msec，オンセット推定と NMF 解析時は 1 msec とした．

評価はオンセット時刻推定と音長，フレームごとの音の有無（ここではノートオンフレームと呼ぶ）の 3 つに分けて行った．Ground truth は MAPS Database に付属している MIDI データを使用した．オンセット時刻推定は推定オンセット時刻が Ground truth から ±20 msec，±50 msec，±100 msec の範囲内にあるかどうかで評価し，音長についても ±50 msec の範囲内かどうかで判断した．ノートオンフレームの評価については，推定ノートオンフレーム内に対し Ground truth において一致している場合を TP，Ground truth の該当フレームにおいてノートオフ状態である場合に FP とし，また逆に Ground truth のノートオンフレームに対してノートオフ状態であると推定した場合を FN とし，以下の評価量，

$$\begin{aligned}
\text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}} \\
\text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\
\text{F}-\text{measure} &= \frac{2\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}
\end{aligned} \tag{19}$$

で求めた．また評価はチューニング補正機能の有無によるノートオン推定精度（3 章），および詳細オンセット推定と音楽特徴量推定の有無によるノートオン推定精度（4 章）の 2 点に焦点を当てて行った．

- チューニング補正機能の影響（3 章）
  チューニング補正有無別の HSMM による楽譜追跡（状態系列）精度評価
- 詳細オンセット・テンポ推定による音長推定（4 章）
  - リファインせず HSMM で得られた状態系列を使用して最終的に推定した結果の精度評価
  - リファインした状態系列を使用して最終的に推定した結果の精度評価

5.2 実験結果

表 2，**表 3** にオンセット，音長，そしてオンセットフレームに関する実験結果を示す．表に記された章節番号は使用した手法（c.f. 図 1）と対応している．個々の傾向としては，HSMM によってノートオンフレームの推定精度が 65%前後と大まかなタイムアライメントをとることはできていたが，一方でオンセット時刻推定は 100 msec 範囲を見ても 50%程度であることから，状態の境界を上手くとらえることができていないと推察できる．

一方で，パワーやその変化に着目したオンセット推定は，HSMM のみの場合と比較して 40%程度の精度向上がみられていることから，その有効性に対する仮説の検証ができたといえる．ただし，上記の音高情報を用いず時間分解能を上げ推定したオンセット時刻および音長と，それらをパラメータとして再度音高情報（スペクトル）も考慮した NMF による推定結果とでは有意な差は得られなかった．

5.3 考察

HSMM による楽譜追跡において状態の境界（オンセット位置）をとらえることが困難であった背景として，今回単音についてのみ事前学習を行いそのパラメータを利用したこと，また音量に対してロバストにするために正規化したことで，和音としての出力ベクトルが観測データとフィットしなかった可能性がある．また，継続長に関し今回の HSMM ではテンポ変化を考慮にいれなかったため，今後は HSMM 自体の最適化やテンポに関するパラメータの導入によって改善する可能性がある．

また詳細オンセット推定においては，特に従来使用された特徴量である AMP や POW，SD のみをそれぞれ単品で見ると，図 4 で示したとおり AMP や POW のみでは和音に対する個々の音の微細なオンセット揺らぎをとらえることは難しい．また SD のみに着目した場合にはパワー減衰時にみられる複数のピークをオンセットのピークと区別することが難しく，総じてそれぞれを独立に考えた場合のオンセットピーク検出は困難であった．一方で今回提案した GradSD を用いることでオンセットにおける微細なオンセット時刻の揺らぎに対しても検出可能であり，詳細解析に必要な時間分解能の高さにも応えうることが示されてい

表 2 オンセット時刻・音長の推定精度評価 (%)
Table 2 Evaluation of onset times and durations.

	オンセット時刻			音長
Ground truth からのずれ (±)	100 msec	50 msec	20 msec	50 msec
Tuning 無 / HSMM (Section 3.3)	44.2	38.9	17.9	35.5
Tuning 有 / HSMM (Section 3.1, 3.3)	53.2	47.1	23.8	46.2
Tuning 有 / HSMM+LSAR (Section 3, 4.2)	92.7	90.5	81.5	33.0
Tuning 有 / HSMM+LSAR+PF (Section 3, 4.2, 4.3)	93.0	91.5	82.2	70.6
Tuning 有 / HSMM+NMF (Section 3, 4.5)	50.9	46.0	27.0	52.9
Tuning 有 / HSMM+LSAR+PF+NMF (Section 3, 4)	93.5	91.7	81.1	71.7

表 3 ノートオンフレームの推定精度評価
Table 3 Evaluation of note-on frames.

	Precision	Recall	F-measure
Tuning 無 / HSMM (Section 3.3)	0.644	0.756	0.695
Tuning 有 / HSMM (Section 3.1, 3.3)	0.685	0.778	0.728
Tuning 有 / HSMM+LSAR (Section 3, 4.2)	0.876	0.928	0.901
Tuning 有 / HSMM+LSAR+PF (Section 3, 4.2, 4.3)	0.909	0.931	0.920
Tuning 有 /HSMM+NMF (Section 3, 4.5)	0.678	0.835	0.748
Tuning 有 / HSMM+LSAR+PF+NMF (Section 3, 4)	0.917	0.939	0.928

る．また，LSAR を用いた詳細オンセット推定も同様に，AIC の減衰ピークを追うことでサンプル点単位でのオンセット検出が可能であり，これらの特徴量が和音の微細なオンセット揺れ検出に有効であることが実際の実験結果からも示された．ただし LSAR は 2 つの構造間の変化点を検出する手法であるため，解析区間に複数の変化点が存在する場合に AIC の大きな減少ピークを明確に検出することは難しく，また高音のオフセットのように音が比較的早く途切れノートオフに入るような場合にも AIC が下がる傾向が見られたことから，より信頼性を向上するためにはさらなる研究調査が必要である．

テンポ推定による推定音長の補正もまた実験結果から推定精度の向上に寄与しているものと考えられる．特にテンポ変化が激しい曲（あるいは緩急が大きい曲）では HSMM による楽譜追跡推定精度は大きく減少する傾向があり，また平滑化なしの IOI から直接求めたテンポも同様に IOI の乱高下の影響から音長も不安定となることから，オンセット推定と平滑化したテンポ推定を活用して楽譜追跡結果をリファインすることは有効であると考えらえる．

次に，音楽特徴量推定に用いた NMF は自由度の少ないパラメータ数で構成されており，アクティベーションの減衰形状も指数的な単調減少であるため，ロバストである反面，ノイズ（あるいはアクティベーションの振動）などの影響により，急峻な形状をとるなどしてパラメータの推定精度が落ち込むケースも見られた．したがって今後は乗法ノイズなどによる振動のモデル化が必要かと思われる．

また，今回音高情報を用いない詳細オンセット（および音長）推定において高精度な結果が得られ，NMF によるそれらの推定精度との有意な差はみられなかったが，誤り検出にまで視野に入れた場合に音高情報との照合は重要な

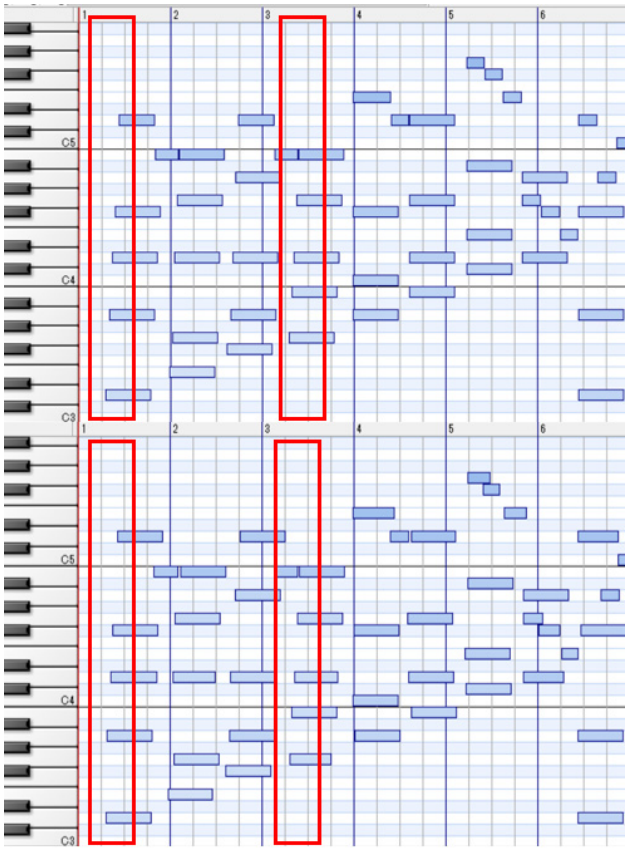

図 9 MAP データベースの MAPS_MUS-ty_mai_StbgTGd2 の正解 MIDI（上図）と推定ピアノロール（下図）．赤枠は（ほぼ同時音ながら）微細なオンセットの揺れがある箇所．推定したピアノロール（下図）でもその微細なオンセットを上手くとらえることができている（c.f. 図 7，表 1）

Fig. 9 An example of an estimated (bottom) piano roll in comparison with the ground truth (top, MAPS_MUS-ty_mai_StbgTGd2, MAPS data set). The algorithm succeeds at detecting asynchronous chord onset times.

ファクターとなりうる．したがって，本稿で我々が提案した詳細推定情報を陽に扱いつつ高い時間分解能を保持したまま，音高情報までを統一的に扱う枠組みは不可欠であると考えられる．

最後に本システムで推定したピアノロール**図 9** を示す．図 9 の上図は解析対象 WAVE データの正解 MIDI のピアノロール表示であり，下図は本システムで推定したピアノロールであり，精度良く推定できていることが分かる．特に始めの赤枠箇所など和音部は図 7 や表 1 でも示したとおり，高精度に分離できている．

6. おわりに

本稿では同時音も分離可能なレベルの高時間分解能な演奏分析および Audio-to-MIDI 変換を実現するために，楽譜情報を既知として，まず音高チューニング補正および HSMM を用いた系列全体に対するアライメントを行った．次にフレームレベルにより着目したパワーやその変化をとらえることによる高精度なオンセット推定と，さらに個々のデータサンプル点にまで時間分解能を上げて LSAR を行う詳細オンセット推定を行い，続けて，その結果を利用したテンポ推定により，先に HSMM で得られた状態系列のリファインを行った．また，解析範囲を限定することによる，高時間分解能解析を実現するための解析範囲窓長の計算方法を明示し，同時に，波形領域までを考慮して得た推定詳細オンセット時刻および音長を事前分布として用いた，単音畳み込みモデルに基づく階層ベイズ推定による音楽特徴量推定方法を提案した．本稿は人間の演奏者による演奏モデルの獲得を念頭に，和音打鍵の微細なオンセット揺れの検出にまで着目する必要性を提示し，実際にその解析手法として，複数の時間分解能による解析を段階的に用いた多重時間分解能解析の提案と，解析範囲を適応的に限定しながら個々のイベントについて詳細に解析するフレームワークとその有効性を示した．

一方で個々の手法に関してはまだ改善の余地があり，たとえば HSMM による全体的なアライメントと，LSAR や PF を用いたリファインについては，これらを統合的なモデルへ拡張することも視野に入れることができる（特に楽譜追跡におけるテンポ推定は重要であると考える）．また今回のモデルでは含まれていないオンセットの誤り検出（Extra note や音抜け，音高誤りなど）やペダル情報の活用なども考えていく必要があり，たとえば誤り検出については NMF における個々の単音エネルギーにガンマ過程などのスパースなモデルを組み込むことによるモデル拡張なども考えらえる．

本稿では誤り検出を考慮しておらず従来手法との十分な比較評価はなされていないが，今後はそれらも考慮したうえで評価し本手法をさらに改良していくとともに，実演奏音楽音響信号を MIDI 変換したデータベースを構築し，演奏モデルの学習と自動演奏表情付与の実現を目指したい．

謝辞 本研究は JSPS 科研費 17H00749 の助成を受けて行われた．

参考文献

[1] Widmer, G. and Goebl, W.: Computational models of expressive music performance: The state of the art, *Journal of New Music Research*, Vol.33, No.3, pp.203–216 (2004).
[2] 奥村健太，酒向慎司，北村 正：楽譜と表情を関連付けた統計モデルに基づく鍵盤楽器演奏の自動生成手法，知能と情報，Vol.28, No.2, pp.557–569（オンライン），DOI: 10.3156/jsoft.28.557 (2016).
[3] Brown, A.D., Stecker, G.C. and Tollin, D.J.: The precedence effect in sound localization, *Journal of the Association for Research in Otolaryngology*, Vol.16, No.1, pp.1–28 (2015).
[4] Haas, H.: On the influence of a single echo on the intelligibility of speech, *Acustica*, Vol.1, pp.49–58 (1951).
[5] 高橋佳吾，西本卓也，嵯峨山茂樹ほか：対数周波数逆畳み込みによる多重音の基本周波数解析，情報処理学会研究報告音楽情報科学（MUS），Vol.2003, No.127 (2003-MUS-053), pp.61–66 (2003).
[6] Saito, S., Kameoka, H., Takahashi, K., Nishimoto, T. and Sagayama, S.: Specmurt analysis of polyphonic music signals, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.16, No.3, pp.639–650 (2008).
[7] Goto, M.: A real-time music-scene-description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals, *Speech Communication*, Vol.43, No.4, pp.311–329 (2004).
[8] Kameoka, H., Nishimoto, T. and Sagayama, S.: A multipitch analyzer based on harmonic temporal structured clustering, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.15, No.3, pp.982–994 (2007).
[9] Lee, D.D. and Seung, H.S.: Algorithms for non-negative matrix factorization, *Advances in Neural Information Processing Systems*, pp.556–562 (2001).
[10] O'Hanlon, K. and Plumbley, M.D.: Polyphonic piano transcription using non-negative matrix factorisation with group sparsity, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp.3112–3116, IEEE (2014).
[11] Benetos, E., Weyde, T., et al.: An efficient temporally-constrained probabilistic model for multiple-instrument music transcription (2015).
[12] Maezawa, A. and Okuno, H.G.: Bayesian audio-to-score alignment based on joint inference of timbre, volume, tempo, and note onset timings, *Computer Music Journal*, Vol.39, No.1, pp.74–87 (2015).
[13] Zhang, J., Tang, J. and Dai, L.-R.: RNN-BLSTM Based Multi-Pitch Estimation, *INTERSPEECH*, Vol.2016, pp.1785–1789 (2016).
[14] Sigtia, S., Benetos, E. and Dixon, S.: An end-to-end neural network for polyphonic piano music transcription, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, Vol.24, No.5, pp.927–939 (2016).
[15] Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S. and Eck, D.: Onsets and frames: Dual-objective piano transcription, arXiv preprint arXiv:1710.11153 (2017).
[16] 落合和樹，中野允裕，小野順貴，嵯峨山茂樹ほか：時間周波数分解能の異なるスペクトログラムの並列 NMF によ

る多重音解析，研究報告音楽情報科学（MUS），Vol.2011, No.5, pp.1–6 (2011).
[17] Benetos, E., Klapuri, A. and Dixon, S.: Score-informed transcription for automatic piano tutoring, *2012 Proc. 20th European Signal Processing Conference* (*EUSIPCO*), pp.2153–2157, IEEE (2012).
[18] Miron, M., Carabias-Orti, J.J. and Janer, J.: Improving Score-Informed Source Separation for Classical Music through Note Refinement, *ISMIR*, pp.448–454 (2015).
[19] Miron, M., Janer, J. and Gómez, E.: Monaural score-informed source separation for classical music using convolutional neural networks, *18th International Society for Music Information Retrieval Conference*, pp.55–62 (2017).
[20] 保利武志，中村和幸，嵯峨山茂樹：多重解像度 NMF に基づく音響信号演奏詳細解析，研究報告音楽情報科学（MUS），Vol.2018, No.16, pp.1–6 (2018).
[21] Hori, T., Nakamura, K. and Sagayama, S.: Multiresolutional Hierarchical Bayesian NMF for Detailed Audio Analysis of Music Performances, *Proc. APSIPA Annual Summit and Conference* (*ASC*) (2018).
[22] Emiya, V., Badeau, R. and David, B.: Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.18, No.6, pp.1643–1654 (2009).
[23] International Piano-e-Competition: International Piano-e-Competition (online), available from ⟨http://www.piano-e-competition.com⟩ (accessed 2019-06-01).
[24] Fujishima, T.: Real-time chord recognition of musical sound: A system using common lisp music, *Proc. ICMC*, pp.464–467 (1999).
[25] Bello, J.P., Duxbury, C., Davies, M. and Sandler, M.: On the use of phase and energy for musical onset detection in the complex domain, *IEEE Signal Processing Letters*, Vol.11, No.6, pp.553–556 (2004).
[26] Bello, J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. and Sandler, M.B.: A tutorial on onset detection in music signals, *IEEE Trans. Speech and Audio Processing*, Vol.13, No.5, pp.1035–1047 (2005).
[27] 市村直幸ほか：自己組織化型状態空間モデルを用いた運動軌跡のフィルタリング，情報処理学会論文誌コンピュータビジョンとイメージメディア（CVIM），Vol.43, No.SIG11 (CVIM5), pp.92–104 (2002).
[28] Classical Archives: Classical Archives (online), available from ⟨https://www.classicalarchives.com⟩ (accessed 2019-06-01).

保利 武志 （学生会員）

2008 年東京学芸大学教育学部環境教育自然環境科学科卒業．2017 年明治大学先端数理科学研究科博士前期課程修了．同年同大学総合数理学部助手．2019 年同大学先端数理科学研究科博士後期課程在籍．専門は音楽の信号処理と記号処理．

中村 和幸

2002 年東京大学工学部計数工学科卒業．2004 年同大学大学院情報理工学系研究科修士課程修了．2007 年総合研究大学院大学複合科学研究科博士後期課程修了．JST CREST 研究員，統計数理研究所特任研究員，明治大学特任講師を経て，2013 年明治大学総合数理学部准教授．2018 年同教授．博士（学術）．専門は時空間データのベイズ分析，データ同化とその応用．日本統計学会，日本応用数理学会，日本シミュレーション学会，American Statistical Association 等の会員．

嵯峨山 茂樹 （正会員）

1948 年生．1972 年東京大学工学部計数工学科卒業．1974 年同大学大学院修士課程修了．同年日本電信電話公社入社，武蔵野電気通信研究所にて音声情報処理の研究に従事．1990 年 ATR 自動翻訳電話研究所音声情報処理研究室長．1993 年 NTT ヒューマンインタフェース研究所主幹研究員．1998 年北陸先端科学技術大学院大学情報科学研究科教授．2000 年東京大学大学院工学系研究科計数工学専攻（現，情報理工学系研究科システム情報学専攻）教授．2014 年明治大学総合数理学部先端メディアサイエンス学科教授．2019 年電気通信大学客員研究員．音楽・音声・音響・文字の信号処理と情報処理の教育研究に従事．博士（工学）．発明協会発明賞，科学技術庁長官賞，電子情報通信学会論文賞，情報処理学会論文賞等を受賞．電子情報通信学会フェロー，日本音響学会，IEEE，各会員．