# 発音時刻の楽譜上の位置を確率モデルにより推定するクォンタイズ手法

浜中雅俊[†] 後藤真孝[††,†††]
麻生英樹[†††] 大津展之[†,†††]

本論文では，ジャムセッション等で伴奏に合わせて弾いた演奏の発音時刻から，元々演奏者が弾こうとした正規化された楽譜上の発音時刻を推定する手法について述べる．本研究の目的は，演奏を再利用しやすい形でデータベース上に蓄積するために，演奏記録の発音時刻を量子化された位置に整列させることである．従来のビート・リズムの認識に関する手法の多くが拍位置の予測や推定に主眼を置いていたのに対し，本手法では，テンポ一定で伴奏の拍位置が既知という条件下で，発音時刻のゆらぎを取り除く問題，すなわちクォンタイズを扱う．和音を含むジャムセッションの MIDI 演奏記録をクォンタイズするため，我々は，発音時刻の遷移とゆらぎを隠れマルコフモデルでモデル化する手法を提案する．本手法の特長は，モデルパラメータを演奏記録から統計的に学習することにより，各演奏に適した確率モデルを使ってクォンタイズすることができる点である．演奏記録を学習し実験した結果，市販のシーケンスソフトウェアの機械的なクォンタイズより性能が良く，モデルが有効に機能したことが示された．

# A Probabilistic-model-based Quantization Method for Estimating the Position of Onset Time in a Score

Masatoshi Hamanaka,[†] Masataka Goto,[††,†††] Hideki Asoh[†††] and Nobuyuki Otsu[†,†††]

This paper describes a method for organizing onset times performed along a jam-session accompaniment into the normalized (quantized) positions in a score. The purpose of this study is to align onset times of a session recording to the quantized positions in order to store the performance data in a reusable form. Unlike most previous beat-tracking-related methods focusing on predicting or estimating beat positions, our method deals with the problem of eliminating the onset-time deviations under the condition that the beat positions are given. To quantize polyphonic MIDI recordings of jam session, we propose a method of using Hidden Markov Model for modeling onset-time transition and deviation. The main advantage of this method is to quantize a performance by using a model that is statistically learned from session recordings of the same player. Experimental results showed that the model that statistically learned MIDI recordings was effective enough to surpass the performance of semi-automatic quantization of commercial sequence software.

## 1. はじめに

我々は，ジャムセッションシステムの研究[1)]において，実在する人間の演奏者の振る舞いを模倣できる仮想演奏者を生成することを目指してきた．仮想演奏者は，事前に人間の演奏者から学習したモデルに基づき演奏意図を決定し，フレーズデータベースの中からそのときの意図に対応したフレーズを次々と接続することで演奏を生成していた．フレーズデータベースには，1 小節から 8 小節の長さの演奏が収められているが，それらは手作業で作成していたため，仮想演奏者は演奏者固有のフレーズまでは模倣できなかった．

そのような模倣を可能とする方法として，ここでは，模倣したい演奏者の即興演奏からフレーズを切り出して，データベースを自動作成することを考えている．演奏の発音時刻には意識的・無意識的なゆらぎがあり，フレーズ切り出しの際，ゆらぎを除去せずに行うと，

† 筑波大学大学院工学研究科
Doctoral Program in Engineering, University of Tsukuba
†† 科学技術振興事業団さきがけ研究 21「情報と知」領域
"Information and Human Activity," PRESTO, JST
††† 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

生成時にフレーズを接続する際に，つなぎ目が不自然となる部分が生じる．たとえば，はしったリズムのフレーズともたったリズムのフレーズが接続された部分では，拍の間隔が不規則となり聞き苦しい演奏となる．このようなことを解決するためには，発音時刻のゆらぎ・ずれを除去するクォンタイズ（quantize）が必要となる．

本研究では，このような即興演奏のフレーズ分割を自動化するため，テンポ一定の伴奏演奏（たとえばドラムスによる伴奏）に合わせて弾かれた演奏の発音時刻を，楽譜上の正規化された位置へとクォンタイズすることを目的とする．つまりここで扱うのは，ゆらぎを含むような即興演奏からゆらぎを取り除く問題である．ただし，即興演奏の場合には楽譜を演奏するわけではないため，楽譜上の正規化された位置は，元々演奏者が弾こうとした小節・拍内の量子化された位置を意味する．

市販のシーケンスソフトウェアに搭載されているクォンタイズは，ユーザがグリッド間隔（分解能）を指定し（たとえば，8 分 3 連音符，16 分音符等の一定間隔のグリッドを指定し），発音時刻を最も近いグリッドへ整列させるものである．したがって，拍構造が既知で固定されている演奏に対しては有効だが，ジャムセッションのように，8 分 3 連音符や 16 分音符が頻繁に入れ替わるような演奏を正しくクォンタイズするためには，演奏の部位ごとに人間が手作業で指定する必要があり，我々のフレーズ分割を自動化する目的には使えない．

従来，コネクショニストモデルによるクォンタイズの研究[2)]，音響信号に対するクォンタイズの研究[3)]，ビート・リズムの認識に関するビートトラッキングの研究[4)~8)] 等がなされてきた．文献 2) の研究は，隣接する発音時刻間隔（あるいはその和）の比が整数比になれば安定するエネルギー関数を定義し，繰返し計算で発音時刻のずれを修正していたが，エネルギー関数が固定されていたため，多様な演奏へ対処することは困難であった．また，文献 3) の研究は自動採譜のための拍の推定のためのものであり，文献 4)~8) の研究は拍位置を予測することに主眼が置かれていた．いずれも，問題設定が本研究（テンポ一定の伴奏に合わせて弾かれた演奏のクォンタイズ☆）とは異なっており，直接それらの手法は適用できない．

一方，文献 9) では，本研究に近い問題設定が扱われている．この研究は，連続音声認識の方法論が，音符推定・テンポ推定・拍子推定にも有効であることを示した点が優れており，音符長の遷移と変動を隠れマルコフモデル（HMM）でモデル化することで，市販のシーケンスソフトウェアのクォンタイズを凌ぐ性能を得ることに成功している．しかし，単音の音符列に対する検討しか行っていなかった．

本研究では，今後さらに発展が期待される文献 9) のアプローチを参考にし，発音時刻に焦点を絞った発音時刻の遷移とゆらぎのモデルを新たに提案し，和音を含む演奏に対しての評価を行った．以下，2 章ではクォンタイズの問題を逆問題としてとらえる立場から，隠れマルコフモデルを用いた問題の定式化と実際のクォンタイズの方法について述べる．次に 3 章では，クォンタイズの正解がラベリングされている演奏データを用いて学習することで，提案モデルがシーケンスソフトウェアのクォンタイズを超える性能を持つことを示す．最後に，4 章でまとめと今後の課題を述べる．

## 2. 学習に基づくクォンタイズ

人間の演奏者が同じ演奏を繰り返し弾いた場合でも，MIDI のレベルでまったく同じ演奏となることは稀であり，演奏動作の微妙な差や，演奏の表情づけの差等により発音時刻がゆらぐ．これを，元々演奏者が弾こうとした発音時刻（小節・拍内の正規化された位置）系列から，ゆらぎのある実際に演奏された発音時刻系列を求める順問題とすると，演奏された発音時刻系列から演奏者が元々弾こうとした正規化された発音時刻系列を求めるクォンタイズは，その逆問題☆☆となる（**図 1**）．ここでは，前者のゆらぎの生じる過程を「発音時刻ゆらぎの順モデル」としてモデル化する．そして，この順モデルから求めた「発音時刻ゆらぎの逆モデル」を用いて逆問題を解き，クォンタイズを実現する．

### 2.1 発音時刻の遷移とゆらぎのモデル

発音時刻ゆらぎの順モデルは，元々弾こうとした正規化された発音時刻系列を $\boldsymbol{\theta}$，演奏された発音時刻系列を $\boldsymbol{y}$ としたとき，$P(\boldsymbol{y}|\boldsymbol{\theta})$ で表される．このとき，逆モデルは，Bayes の定理より式 (1) で表される．

$$P(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{y})} \tag{1}$$

$P(\boldsymbol{\theta})$ は，$\boldsymbol{\theta}$ に対する事前分布であり，演奏者がど

☆ このような問題設定でも，単純な閾値処理は有効でなく（ゆらぎによるずれを誤ってクォンタイズする等），難しい課題である．

☆☆ このように，実際の演奏から，その拍構造，楽譜構造を推定する問題を逆問題としてとらえる考え方は，文献 4)~6) で紹介された．文献 9) でも，同様に逆問題ととらえて定式化しており，本研究でもそれを参考にしている．

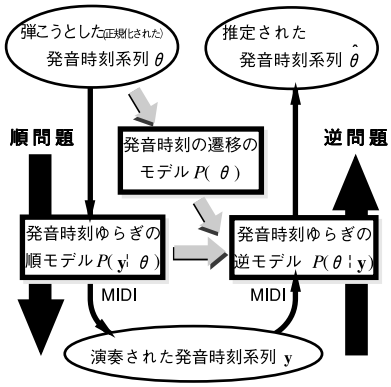



**図 1** クォンタイズにおける順モデルと逆モデル
Fig. 1 Forward model and inverse model in the quantization problem.

のような発音時刻系列を弾きやすいかを表す．逆問題の解 $\hat{\boldsymbol{\theta}}$ は，式 (1) を最大化する $\boldsymbol{\theta}$ である（式 (2)）．

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\theta}{\operatorname{argmax}}\ P(\boldsymbol{\theta}|\boldsymbol{y}) \\ &= \underset{\theta}{\operatorname{argmax}}\ P(\boldsymbol{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \qquad (2)\end{aligned}$$

### 2.2 隠れマルコフモデルによる定式化

本研究では，発音時刻ゆらぎの順モデル $P(\boldsymbol{y}|\boldsymbol{\theta})$ と発音時刻の遷移のモデル $P(\boldsymbol{\theta})$ を組み合わせたモデルを隠れマルコフモデル（HMM）で定式化する．隠れマルコフモデルはマルコフ的な隠れ状態遷移モデル $P(\boldsymbol{\theta})$ と各状態における出力確率分布 $P(\boldsymbol{y}|\boldsymbol{\theta})$ を組み合わせたモデルである．我々が観測することができるのは，出力確率分布から得られる出力のみで，どの状態にいるかを観測することはできない．

#### 2.2.1 演奏のモデル化

演奏をモデル化する前提条件として，以下のものを考える．

(1) **何をモデル化するのか**

本研究では，演奏の発音時刻すなわち音の立ち上がり時刻の分布をモデル化する．文献 9) では，音符の長さをモデル化していたが，ギターのアルペジオや開放弦を含む演奏や，ピアノで右手と左手を同時に弾いていているような演奏では，音符の長さのみで，演奏をモデル化することは困難であり，発音時刻でモデル化する方が有効であると考える．

発音時刻を用いて演奏をモデル化する際，発音時刻の前後関係が保存されるようにすることで，音の順序が反転することがない妥当なクォンタイズを実現できる．

(2) **演奏のモデル化の単位**

本研究では，演奏をモデル化する単位として 1 拍の長さを採用した．1 拍単位でモデル化を行う理由は，市販のクォンタイズを使った場合の主要な問題点である 8 分 3 連音符と 16 分音符の識別が 1 拍の長さで可能となるためである（8 分 3 連音符は 1 拍を 3 等分した位置，16 分音符は 1 拍を 4 等分した位置に出現する）．実際に曲をクォンタイズする場合には，1 拍のモデルを，次々とつなげあわせて曲全体を表現する．1 拍という短い長さでモデル化することは，計算時間の節約と，学習の際により多くの演奏データを用意しやすいという利点を持つ．

(3) **演奏の量子化単位**

以下の 2 種類の異なる量子化単位を持つ時刻を定義する．

- **拍内のイベント時刻** $k$
  実際に演奏された発音時刻をイベント時刻と呼ぶ．その量子化単位は，シーケンスソフトウェアで採用されることが多い 1 拍の 1/480 を単位とする．拍内のイベント時刻 $k$ は，その拍内での時刻を表し，伴奏の拍線（拍の境目）を基準に 0 を決め，0 から 479 の整数値をとる．
- **正規化後の拍内位置** $i$
  演奏者が元々弾こうとした正規化された発音時刻を表す．量子化単位は，8 分 3 連音符と 16 分音符の両方に対応できるように 1 拍の 1/12 を単位とし，$i$ は 1 から 12 の整数値をとる．

本研究で述べるクォンタイズとは，演奏を通しての観測されたイベント時刻の系列から正規化後の拍内位置の系列を求めることである．

#### 2.2.2 1 拍の HMM モデル

隠れマルコフモデルの隠れ状態を，正規化後の拍内位置に対応づけ，モデルの出力を観測されたイベント時刻，すなわち，実際の発音時刻に対応づける．ここでは，拍を 12 等分した位置を，拍の頭から順に状態 1 から状態 12 までに対応させる．遷移は仮想的な状態である Start から始まって，End で終了する．和音等，同じ発音時刻で複数の音が出力される演奏の場合は，同じ状態を自己ループする遷移となる．以下に，遷移の例をあげる．

- 拍内に 8 分 3 連音符が 3 つ並んでいる場合
  Start → 1 → 5 → 9 → End
- 拍内に 16 分音符が 4 つ並んでいる場合
  Start → 1 → 4 → 7 → 10 → End
- 拍の頭で 2 音同時に鳴る和音の場合
  Start → 1 → 1 → End
- 拍内に音がない場合
  Start → End

HMM（離散 HMM）の各パラメータを以下のように意味づけた．

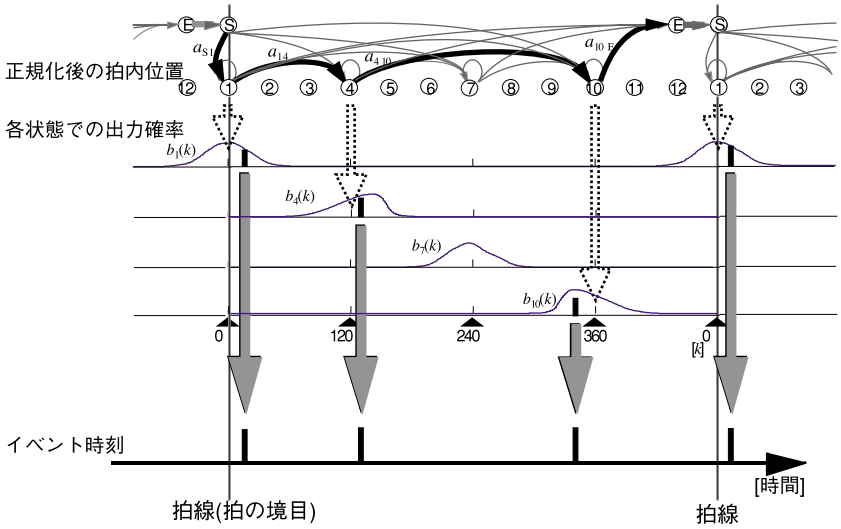



**図 2** 1 拍の隠れマルコフモデルの概略図
Fig. 2 Overview of the quarter-note Hidden Markov Model.

・隠れ状態 $i$：発音時刻の正規化後の拍内位置（$i$ は 1 から 12 の整数）
・出力 $k$：拍内のイベント時刻（$k$ は 0 から 479 の整数）
・状態遷移確率 $a_{ij}$：拍内の $i$ の位置で発音した後 $j$ の位置で発音する確率．遷移は，つねに拍内の前から後ろに向かって進むため，$a_{iS} = 0$，$a_{Ei} = 0$ となり，$i > j$ のときは $a_{ij} = 0$ となる．
・出力確率 $b_i(k)$：発音時刻の正規化後の拍内位置が $i$ のときに，拍内のイベント時刻が $k$ となる確率．状態 Start と End は出力を出さないため，対応する出力分布は存在しない．状態遷移はつねに S から始まるため，HMM 初期状態分布 $\pi_i$ は，$\pi_S = 1$，$\pi_i(i = 1, 2, ..., 12, \mathrm{E}) = 0$ となる．

**図 2** に，本研究における HMM の概略図を載せる．図では，簡単のため 12 個の状態のうち，1，4，7，10 以外の状態に関する遷移と出力確率は省略している．

### 2.2.3 複数の HMM によるモデル化

以上のような隠れマルコフモデルのすべての遷移を考えた場合，**図 3**(a) のようにアークの本数が非常に多く複雑となり，遷移確率の統計的な推定精度が低くなると考えられる．しかし，演奏者が，1 拍の内部を 16 分音符で演奏しようとする場合と，8 分 3 連音符で演奏しようとする場合とを分けて考えた場合，両者はより単純になる．つまり，16 分で演奏しようとする場合には，状態 1，4，7，10 へ遷移する確率が高くなり，8 分 3 連で演奏しようとする場合には，状態 1，5，9 へ遷移する確率が高くなる．したがって，1 拍を 1 つの HMM によってモデル化するよりも，複数

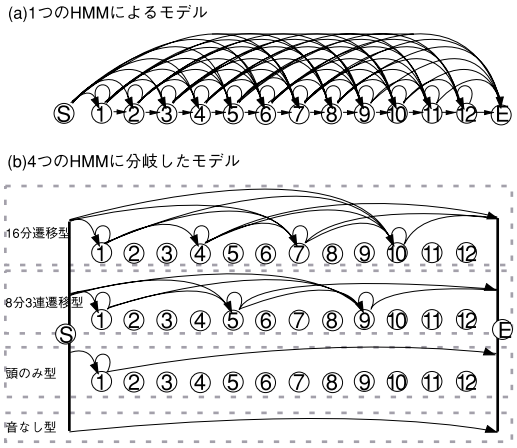



**図 3** 1 つの HMM によるモデルと 4 種類の HMM に分岐したモデル
Fig. 3 A model consisting of a single HMM and a model consisting of four different HMMs.

の HMM を切り替えて用いるほうが，より詳細なモデル化ができ，クォンタイズの性能がさらに向上すると期待される．そこで，本研究では，図 3 のように，1 つの HMM によるモデルとともに，16 分音符で弾かれている「16 分遷移型」，8 分 3 連音符で弾かれている「8 分 3 連遷移型」，拍の頭だけ弾かれている「頭のみ型」，何も弾かれていない「音なし型」の 4 種類の HMM に分岐したモデル☆を用い，実験によりどちらのモデルが性能が高いか比較する．4 つの HMM によ

☆ このような複数の HMM によるモデル化は文献 9) でも扱われており，有効性が確認されている．

るモデルは，1 拍の演奏が 4 種類のモデルのいずれかにあてはまるという制約を与えたものだといえる．通常，1 拍の内部に 16 分系の音符と 8 分 3 連系の音符が混在することは考えにくいため，このような制約は妥当であると考えられる．また，本研究で実験に使った演奏には，32 分音符等 8 分 3 連や 16 分よりも短い音価の音符は含まれていなかった．

### 2.3 最適発音時刻系列の推定

1 曲を通して，最適な状態遷移系列を推定するために，各拍ごとのモデルの Start と End をつなぎ合わせた HMM を用いて，事後確率 $P(\boldsymbol{\theta}|\boldsymbol{y})$ を最大にする状態遷移系列を探索する．そのための方法としては Viterbi アルゴリズムを用いる．

$T$ 個の音からなる演奏を考え，演奏を通しての観測されたイベント時刻の系列を $\boldsymbol{y} = (y_1, y_2, ..., y_T)$ とする．これに対する最適状態遷移系列 $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_T)$ を求めるために，次の値を定義する．ただし，$y_t$ は 1 拍の 1/480 を単位とする．また，$\theta_t$ は各音に対応する状態で，1 から 12 の値をとる．

$$\delta_t(i) = \max_{\theta_1,\theta_2,...,\theta_{t-1}} P[\theta_1, \theta_2, ..., \theta_{t-1} ; \theta_t = i; y_1, y_2, ..., y_t|\lambda] \quad (3)$$

$\delta_t(i)$ は，$t$ 個目のイベントまでを観測した時点で，状態遷移が状態 $i$ で終わっている状態遷移系列中のベストスコア（最も高い確率を与える状態遷移系列の確率）である．また，$\lambda$ はモデルのパラメータを表す．ベストスコアの値は，次のような漸化式を満たす．ただし，次式において A mod B は，A を B で除算した余剰とする．n は，End と Start を挿入した回数で，0 以上の整数である．

$$\delta_{t+1}(j) = \max_i [\delta_t(i) A_{ij}] \cdot b_j(y_{t+1} \bmod 480) \quad (4)$$

ただし，

$$A_{ij} = a_{ij} \qquad (n = 0)$$
$$A_{ij} = a_{iE} \cdot a_{Sj} \cdot a_{SE}{}^n \qquad (n > 0)$$

Viterbi アルゴリズムは，**図 4** のようなトレリス上で，左から右に至る状態遷移系列の探索を行うアルゴリズムである．このトレリスは，横軸にイベントを，縦軸に 1 から 12 の状態を配置しており，各イベントを 1 から 12 のいずれかの状態に対応づけてゆくことで，1 つの状態遷移系列が得られる．拍線（拍の境目）をまたぐ状態遷移には End と Start を挿入する．このようなトレリス上で左から右に向かって式 (4) を用いながら $\delta_t(i)$ を計算してゆくことにより，最も事後確率の高い系列を選択することができる．

本研究ではテンポ一定の伴奏に合わせて弾いた演奏

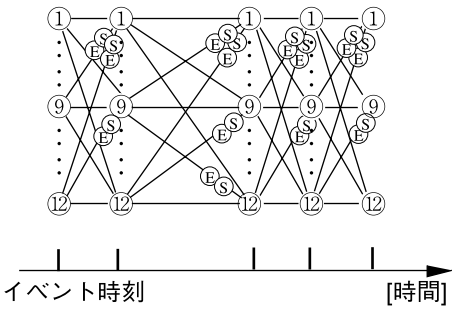



**図 4** 最尤な状態遷移を求めるトレリス
Fig. 4 A trellis for finding the maximum likelihood state transition.

のクォンタイズを扱っているため，伴奏の拍線は既知だが，演奏の発音時刻はゆらいでいるため，演奏の拍線を見つけることは容易ではない．本研究では，演奏のイベント間隔とクォンタイズ後のイベント間隔との差を半拍以内にするという条件で，トレリス上で考えられるすべての 2 つのイベント間のパスで，拍線をまたいだか否かを判定する．実際の演奏でも，発音時刻間隔が半拍以上ずれることはほとんどないため，この判定条件は妥当であるといえる．ここでは，**隠れ状態の拍中の時刻** $q_t$ と**クォンタイズ後のイベント間隔** $r_{t,\,t+1}$ という 2 つの概念を定義したうえで，拍線をまたいだ回数を求める．

- 隠れ状態の拍中の時刻 $q_t$

  $q_t$ は，拍の頭を 0 としたときの，1 から 12 までの各状態が対応するイベント時刻を示し次式で定義される．

  $$q_t = \left(\frac{\theta_t - 1}{12}\right) \times 480 \quad (5)$$

- クォンタイズ後のイベント間隔 $r_{t,\,t+1}$

  $r_{t,\,t+1}$ は，2 つのイベント間のパスが実際に選択されたと考えた場合のクォンタイズ後のイベントの間隔であり，次式で定義される．

  $$r_{t,\,t+1}(n) = (q_{t+1} - q_t) + n \times 480 \quad (6)$$

  ここで $n$ は，End と Start を挿入した回数である．

このとき，2 つのイベントが拍線をまたいだ回数 $\hat{n}$ は，実際の演奏のイベント間隔 $y_{t+1} - y_t$ とクォンタイズ後のイベント間隔 $r_{t,\,t+1}$ との差を最小にする $n$ である（式 (7)）．

$$\hat{n} = \operatorname*{argmin}_n \{(y_{t+1} - y_t) - r_{t,\,t+1}(n)\} \quad (7)$$

たとえば，**図 5** のような音符のクォンタイズを考えたとき，これに対応する適切な遷移は，状態 1（$q_t = 0$）から状態 1（$q_{t+1} = 0$）への遷移となる（実際には，あらゆる状態間の遷移を計算する）．このとき，式 (6)

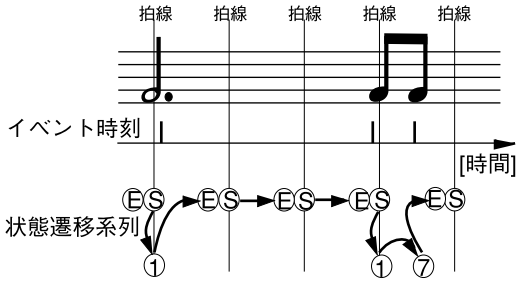



**図 5** 付点 2 分音符の状態遷移の例
Fig. 5 An example of state transition for a dotted half note.

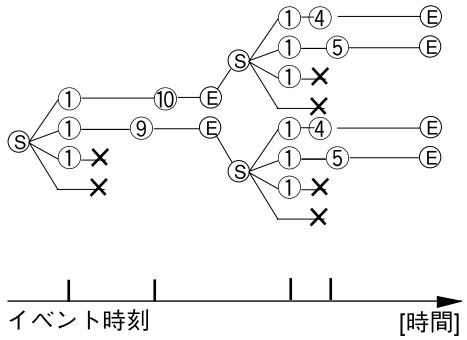



**図 6** 4 つの HMM による探索木の例
Fig. 6 An example of a search tree when using four HMMs.

で $q_{t+1} - q_t$ の値は 0 となり，代入すると式 (7) は

$$\hat{n} = \underset{n}{\operatorname{argmin}} \{(y_{t+1} - y_t) - n \times 480\} \qquad (8)$$

となる．このとき，$1200 \leq y_{t+1} - y_t \leq 1680$ とすると $\hat{n}$ の値は 3 となり，図 5 のように 3 回 End と Start を挿入すればよいと求まる．

4 つの HMM に分岐したモデルの場合には，各拍で 4 つの HMM のどれを用いるかを，**図 6** のような木構造を用いて探索していかなければならないが，プログラムとしては図 4 と同様のトレリス上の探索に帰着させることができる．その場合，トレリスの縦軸にあと 2 行を加え，16 分遷移型の状態 1 と，8 分 3 連遷移型の状態 1 と，頭のみ型の状態 1 とをそれぞれの別々の状態として扱う．また，音なし型の場合には，トレリス上に Start と End を 1 つずつ挿入する．

## 3. モデルパラメータの学習

本学習では，元々弾こうとした正規化された発音時刻系列 $\boldsymbol{\theta}$（正解データ）と，演奏された発音時刻系列 $\boldsymbol{y}$ の両方を学習用データとして用意することにより，発音時刻の遷移のモデル $P(\boldsymbol{\theta})$ すなわち $a_{ij}$ と，発音時刻ゆらぎの順モデル $P(\boldsymbol{y}|\boldsymbol{\theta})$ すなわち $b_j(k)$ を学習する．学習によって得られたモデルパラメータを使ってクォンタイズすることにより，モデルが正しい挙動をしていることを確認する．

### 3.1 学習用データ

学習用のデータセットとして，人工的に生成したデータと，人間の演奏者のデータとの 2 通りを用意した．人工データは，プログラムの動作および，発音時刻のゆらぎの大きさによるクォンタイズの性能変化についての検証を意図したものである．

(1) **人工データ**

人工データの生成では，まずその拍を 8 分 3 連音符の演奏にするか，16 分音符の演奏にするかを乱数で決定し，1 拍に含まれる音数を 1 から 6☆までの乱数で決定する．次に，それぞれの音が，どの位置に割り振られるか乱数で決定する．8 分 3 連音符では 3 通り，16 分音符では，4 通りのうちいずれかの位置となる．以上のようにしてできたランダムデータ $\boldsymbol{\theta}_a$ の発音時刻を，平均 0，標準偏差 $\sigma$ の正規分布の乱数を用いて，ゆらぎをつけたものを人工データ $\boldsymbol{y}_a$ とする．$\sigma$ =10, 20, 30（1 拍=480）とした 3 セットを作った．

(2) **実演奏データ**

3 人の演奏者 A，B，C が，テンポ一定の伴奏に合わせて MIDI ギターで即興演奏した記録を，実演奏データ $\boldsymbol{y}_h$ とする．演奏の長さは 12 コーラス（1 コーラス=12 小節）であり，そのうち約半分がソロ，残りが伴奏である．伴奏部分の多くは 2 音の和音の演奏である．各演奏者に 2 回ずつ，一方はテンポ 120，もう一方は演奏者自身が設定したテンポでの演奏とし，合計 6 セットのデータを作成した．

(3) **正解データ**

学習を行うためには，正解データ，すなわち，演奏者が元々弾こうとした正規化された発音時刻を教師データとして与える必要がある．人工データの場合には，ゆらぎを加える前のランダムデータ $\boldsymbol{\theta}_a$ が正解データとなる．

一方，人間の即興演奏データの場合には，正解データが存在しないので，人間が手作業でラベルづけをすることにより正解データ $\boldsymbol{\theta}_h$ を作成した．具体的には，市販のシーケンスソフトウェア（Twelve Tone System, Cakewalk Pro Audio 9）を用いて，ピアノロール表示☆☆した音符の位置と長さを視覚的に確認しながら，1 音ずつ適切な位置にラベルづけしていった．発音時

☆ 実験で使用した実演奏データの 1 拍内の音数の最大は 6 音であった．

☆☆ ピアノロール表示とは，縦軸を音の高さ，横軸を時間とし，音の出るタイミングと鳴り続けている長さを表示するものである．

刻のゆらぎが大きく，いくつかの正解候補が考えられた場合には，すべての候補のクォンタイズ結果を作って聴き比べ，より適切だと判断したものを採用した．

### 3.2 モデルパラメータの推定

正解データ $\boldsymbol{\theta}$ と演奏された発音時刻系列 $\boldsymbol{y}$ から，$a_{ij}$，$b_j(k)$ を推定した．出力確率 $b_j(k)$ の分布の学習は，学習データの音数の不足を補うため，音が入ったイベント時刻を中心とし標準偏差が 5（1 拍=480）のガウス分布を足し合わせることにより求めている．

**図 7** 中の (a) は，人工データを作るときのゆらぎの標準偏差 $\sigma = 20$（1 拍=480）とした場合の人工データから学習された，$b_5(k)$ の分布である．(b) は，演奏者 C の 2 回目の演奏（演奏 C2 と呼ぶ）から求めた $b_5(k)$ の分布である．人工データは，元々正規分布の乱数でゆらぎを与えているので，$b_5(k)$ の分布 (a) は，正規分布となった．一方，実演奏データから求めた，$b_5(k)$ の分布 (b) もほぼ正規的な分布となったが，分布の中心がやや後ろにずれていた．このことは，状態 5 の位置の音符が，元々演奏者が弾こうとした発音時刻よりも遅れて演奏される傾向があることを示している．実際，演奏 C2 を調べたところ，8 分 3 連音符の頭の音が通常より長く弾かれ，次の音の発音時刻が遅れる傾向が認められた．

**図 8** は，演奏 C2 から求めた 4 つの HMM の遷移確率 $a_{ij}$ である．8 分 3 連遷移型，頭のみ型に対し 16 分遷移型の自己ループの確率が低くなっていた．これは 16 分遷移型では和音が出力されにくい傾向があることを示しているが，実際の C2 の演奏でもそのようになっていた．

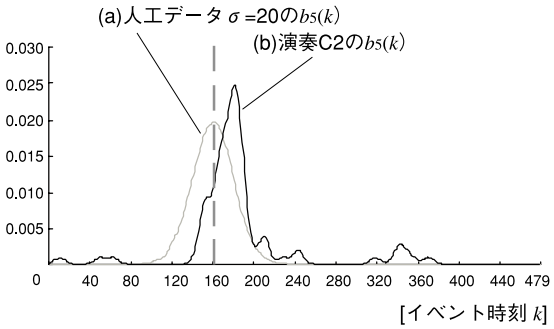



**図 7** 人工，実演奏データから求めた $b_5(k)$
Fig. 7 $b_5(k)$ obtained from artificial data and human-performance data.

### 3.3 クォンタイズ結果

クォンタイズの評価を，音単位と拍単位の 2 通りで行った．音単位の評価は，各音の発音時刻が正解と一致している割合である．拍単位の評価は，拍の種類を，16 分遷移型，8 分 3 連遷移型，頭のみ型，音なし型の 4 種類に分けたとき，各拍の種類が正解の種類と一致している割合である．この 4 種類への分け方は，1 つの HMM でのモデルの場合は，クォンタイズ結果から求め，4 つの HMM に分岐したモデルの場合は，クォンタイズしたときに選択された HMM の型から求めている．

**表 1** に正解データに含まれる 4 種類の拍の割合を示す．演奏者 A，B が 2 回の演奏でそれぞれ似た傾向の演奏をしているのに対し，演奏者 C は，1 回目は 16 分音符を多用した演奏，2 回目は 16 分音符のまったくない演奏であった．

本研究では，一致率を以下のように定義し，クォンタイズの性能を評価する．

$$(\text{音単位の一致率}) = \frac{(\text{正規化後の拍内位置が正解と一致した音数})}{(\text{正解データ中の音数})} \quad (9)$$

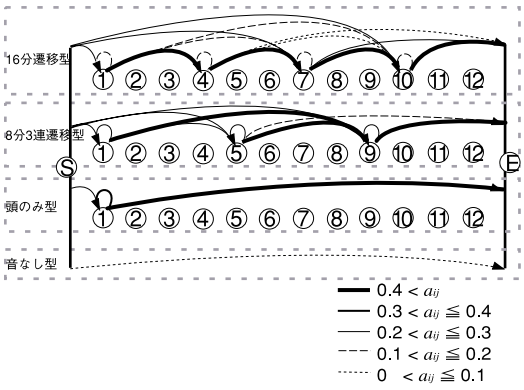



**図 8** 演奏 C2 から求めた $a_{ij}$
Fig. 8 $a_{ij}$ obtained from C2-performance data.

**表 1** 正解データに含まれる 4 種類の拍の割合
Table 1 Percentages of four kinds of beat included in correct data.

| | 人工データ | | | 演奏者 A | | 演奏者 B | | 演奏者 C | |
|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 10$ | $\sigma = 20$ | $\sigma = 30$ | A1 | A2 | B1 | B2 | C1 | C2 |
| 16 分の拍の割合 | 38.8% | 38.8% | 38.8% | 21.4% | 5.5% | 6.5% | 2.7% | 37.2% | 0.0% |
| 8 分 3 連の拍の割合 | 37.8% | 37.8% | 37.8% | 54.2% | 68.7% | 58.2% | 85.8% | 42.1% | 39.7% |
| 頭だけの拍の割合 | 6.9% | 6.9% | 6.9% | 6.8% | 10.1% | 11.9% | 6.4% | 5.9% | 26.7% |
| 音がない拍の割合 | 16.4% | 16.4% | 16.4% | 17.6% | 15.7% | 23.3% | 5.1% | 14.8% | 33.8% |

**表 2** 機械クォンタイズと本手法の比較（音単位の一致率）
Table 2 Comparison between a commercial sequence software and our method (a percentage of correct onsets).

| | 人工データ | | | 演奏者 A | | 演奏者 B | | 演奏者 C | |
|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 10$ | $\sigma = 20$ | $\sigma = 30$ | A1 | A2 | B1 | B2 | C1 | C2 |
| 機械クォンタイズ（8 分 3 連） | 65.6% | 58.2% | 62.0% | 67.6% | 85.6% | 79.4% | 88.6% | 57.0% | 97.7% |
| 機械クォンタイズ（16 分） | 63.9% | 67.9% | 65.9% | 54.5% | 37.3% | 36.8% | 34.7% | 70.7% | 45.5% |
| 機械クォンタイズ（16 分 3 連） | 77.1% | 70.7% | 60.8% | 57.7% | 48.4% | 57.8% | 51.3% | 56.1% | 82.5% |
| 制約つき機械クォンタイズ | 100% | 99.7% | 96.3% | 74.9% | 66.0% | 63.3% | 48.1% | 81.0% | 77.6% |
| 1 つの HMM でのクォンタイズ | 99.6% | 95.9% | 86.5% | 75.9% | 84.8% | 80.0% | 90.5% | 85.1% | 95.0% |
| 4 つの HMM でのクォンタイズ | 99.5% | 95.9% | 86.5% | 82.3% | 89.8% | 81.4% | 92.8% | 85.5% | 95.7% |

※機械クォンタイズに対して，本手法のほうが性能が向上した箇所に下線をつけた．

$$(\text{拍単位の一致率}) = \frac{(\text{拍の種類が正解と一致した拍の数})}{(\text{正解データの拍の数})} \quad (10)$$

### (1) 機械クォンタイズ

まず，閾値処理による機械的なクォンタイズ（機械クォンタイズと呼ぶ）の性能を評価するために，8 分 3 連，16 分，16 分 3 連の 3 種類の分解能で機械クォンタイズを実行し，音単位での一致率を求めた（**表 2**）．なお，これ以外の分解能では一致率はさらに悪かった．その結果，A1 や C1，人工データのように 16 分の拍が多く含まれる演奏では，低い一致率しか得られなかった．ただし，C2 のように 16 分で弾かれた拍がまったくない演奏では，8 分 3 連の分解能での機械クォンタイズが高い一致率を示すこともあった．

### (2) 制約つき機械クォンタイズ

実験に用いた演奏には，8 分 3 連と 16 分の両方の音符が含まれている．したがって，1 種類の分解能の機械クォンタイズで良い結果を得ることは困難である．そこでここでは，1 拍内の音符が，8 分 3 連系か 16 分系のいずれかであるという制約つきのクォンタイズを考える．制約つきクォンタイズでは，1 拍内の発音時刻に対して 8 分 3 連および 16 分の 2 種類の機械クォンタイズを同時に行う．そして，クォンタイズ前の発音時刻の位置とクォオンタイズ後の発音時刻の位置の距離（差の絶対値）の合計が少ない方のクォンタイズ結果を，拍ごとに選択していく．

その結果，人工データに対しては，$\sigma = 10$ では 100%の音単位の一致率を示し，$\sigma = 30$ でも 96%以上という高い一致率を示していた．一方，実演奏データでは，A1 と C1 を除き，8 分 3 連の機械クォンタイズより性能が劣っていた．このことから制約つきクォンタイズは，人工データのように演奏音の発音時刻の分布が正規分布でばらつきが小さい場合には有効であるが，多くの実演奏データ等演奏音のばらつきが大きい場合ではクォンタイズの選択が難しくなり，性能が低下することが分かった．

### (3) 学習に用いた演奏の HMM クォンタイズ

学習結果のモデルパラメータを用いて，学習に使ったのと同じデータについて HMM クォンタイズを実行し評価した（表 2）．以下に結果を述べる．

- HMM クォンタイズと制約つき機械クォンタイズの比較
  両者の一致率を比較したところ，実演奏データでは HMM クォンタイズのほうが高い値を示していた．このことから，発音時刻の状態遷移確率や出力確率に傾向（たとえば状態 5 の位置の音符が，元々演奏者が弾こうとした発音時刻よりも遅れて演奏される傾向）を持つ実演奏データの場合では，HMM によるモデルが有効であることが示された．本研究では，各状態での出力確率 $b_j(k)$ を 1 曲の学習データから得た分布を用いているため，曲の中で局所的にリズムがはしったり，もたったりした部分では，正しくクォンタイズできない場合があった．このような局所的なリズムの変化への対処については，今後検討していく必要がある．
  一方，人工データでは HMM クォンタイズの性能が制約つき機械クォンタイズの性能を下回った．その原因としては，出力確率 $b_j(k)$ を学習する際に足し合わせたガウス分布の標準偏差の値が大きすぎたことが考えられる．標準偏差の値を大きくすると，学習データが少ない場合でも出力確率がなめらかになるという利点があるが，値が大きすぎると分布のすそのが広がってしまい，クォンタイズの性能が悪化するという欠点もある．人工データのように発音時刻の分布のばらつきが小さい場合には，標準偏差の値が小さいほうが高い性能が得られる．実際，人工データ（$\sigma = 10$）を標準偏差を 1 として学習した出力確率 $b_j(k)$ で HMM クォンタイズしたところ，一致率は 100%となった．

**表 3** 1 つの HMM によるモデルと 4 つの HMM によるモデルの比較（拍単位の一致率）
Table 3 Comparison between the one-HMM model and the four-HMM model (a percentage of correct beats).

| | 人工データ | | | 演奏者 A | | 演奏者 B | | 演奏者 C | |
|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 10$ | $\sigma = 20$ | $\sigma = 30$ | A1 | A2 | B1 | B2 | C1 | C2 |
| 1 つの HMM でのクォンタイズ | 98.6% | 93.8% | 81.3% | 79.0% | 85.4% | 84.2% | 91.9% | 83.9% | 94.9% |
| 4 つの HMM でのクォンタイズ | 98.6% | 93.8% | 81.3% | 84.9% | 89.2% | 87.5% | 93.5% | 84.7% | 97.3% |

**表 4** 他の演奏のモデルパラメータでのクォンタイズ結果
Table 4 Quantization results by using the model parameters for other performances.

| | 人工データ | | | 演奏者 A | | 演奏者 B | | 演奏者 C | |
|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 10$ | $\sigma = 20$ | $\sigma = 30$ | A1 | A2 | B1 | B2 | C1 | C2 |
| 1）自分ので別の演奏 | — | — | — | 75.9% | 85.3% | 79.4% | 91.6% | 55.5% | 77.9% |
| 2）自分以外の演奏 | — | — | — | 70.2% | 79.0% | 59.2% | 73.5% | 76.9% | 93.1% |
| 3）ランダムデータで | — | — | — | 73.3% | 53.4% | 60.8% | 55.2% | 82.5% | 86.1% |

- 1 つの HMM によるモデルと 4 つの HMM によるモデルの比較
  両者の音単位の一致率を比較すると，実演奏データでは，4 つの HMM によるモデルのほうが，一致率が高かった．これは，状態 1 を，16 分遷移型，8 分 3 連遷移型，頭のみ型の 3 状態に分けることにより，各々の場合の $b_1(k)$ の分布を別々に学習し，より詳細にモデル化されたためであると考えられる．実際，拍単位の評価でも，4 つの HMM によるモデルのほうが，1 つの HMM によるモデルより高い一致率を示していることから，16 分遷移型と 8 分 3 連遷移型の識別性能が向上したことが確認できた（**表 3**）．
  一方，人工データでは，1 つの HMM によるモデルを使った場合と 4 つの HMM によるモデルを使った場合とが同じ一致率となった．これは，そもそも人工データが正規分布の乱数でゆらぎを与えているために，状態 1 を分けた場合の 3 つの $b_1(k)$ も，分ける前の $b_1(k)$ も，いずれも同じ正規分布となるためである．
- クォンタイズ結果の例
  **図 9** は，演奏 C2 のクォンタイズ結果の一部を示したものである．機械クォンタイズ（16 分）では，8 分 3 連音符のクォンタイズに失敗し，機械クォンタイズ（8 分 3 連）では，16 分音符のクォンタイズに失敗した．また，制約つきクォンタイズでも 2 拍目でクォンタイズに失敗した．これは，2 拍目の 8 分 3 連の発音時刻が遅れ，16 分の機械クォンタイズが選択されたためである．一方，HMM クォンタイズでは，1 つの HMM と 4 つの HMM の両方とも，すべてのイベント時刻のクォンタイズに成功した．

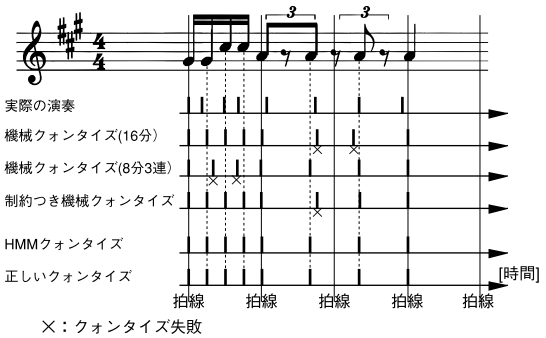



**図 9** クォンタイズ結果の例
Fig. 9 An example of quantization results.

### (4) 他の演奏のモデルパラメータでの HMM クォンタイズ

各実演奏データに対して，それ以外の演奏で学習したモデルパラメータを用いて，HMM クォンタイズを実行し評価した（**表 4**）．評価に用いたモデルパラメータは，1）同じ演奏者の別の演奏で学習したもの，2）自分以外の 2 人の演奏で学習したもの，3）ランダムデータ（$\sigma = 20$）で学習したもの，の 3 通りである．評価は，4 つの HMM によるモデルで行った．
1）は，ある演奏から得られたモデルパラメータが，同じ演奏者のほかの演奏に適用できるか調べるためのものである．1）の結果，演奏者 A，B は，その曲から学習したパラメータを使ってクォンタイズするのと比べて，一致率が最大 6.4%しか劣っていなかった．一方，演奏者 C は，20%近く劣っていた．これは，2 回の演奏をしたときに，演奏者 A，B はそれぞれ似た傾向の演奏を 2 度したのに対し，C は，大きく異なる傾向の演奏をしたためだと考えられる．このことから，A，B のように，演奏傾向が似ている場合には同じモデルで適用できるが，大きく異なる場合には，適用で

きないということが分かった．2），3）は，モデルパラメータにいくつかの演奏の平均的なモデルや，正規的な分布を与えることにより，クォンタイズの性能がどう変化するかを調べたものであるが，いずれ場合も，A，B では 1）より性能が劣っており，また C でも，1）より高くなったものの，その曲から学習したパラメータを使う場合と比べて一致率が低下していた．このことは，本研究で提案したモデルが演奏者の発音時刻の特徴や癖を獲得しており，ある演奏者のモデルを一度作ればその演奏者が似た傾向の演奏をしたときには，同じモデルでクォンタイズが可能であることを示している．

## 4. ま と め

本論文は，テンポ一定の伴奏に合わせて弾かれたゆらぎを含む即興演奏の発音時刻から，演奏者が元々弾こうとした正規化された発音時刻を推定する手法を提案した．本手法では，発音時刻の遷移とゆらぎを隠れマルコフモデルを用いてモデル化することにより，和音を含む演奏を確率モデルで表すことを可能とした．そして，正解データを用いた学習をすることにより，1 つの HMM によるモデルが市販のシーケンスソフトウェアのクォンタイズを超える性能を持つことを示した．また，演奏の各拍に表れるパターンが 16 分系か 8 分 3 連系のいずれかしか含まない場合には，そのパターンに対応した 4 種類の HMM を用意することにより，性能が向上することを示した．ここでは，4 種類の HMM によるモデルしか扱っていなかったが，今後 32 分音符等を含む曲のクォンタイズを考えた場合，HMM の数を増やしていく可能性について検討しなくてはならない．提案手法によるクォンタイズが，発音時刻のゆらぎの大きさや演奏の傾向によらず安定して高い性能を示したことは，クォンタイズへの学習の導入が有効であることを表している．

本研究では，学習によって得られたモデルパラメータを，実演奏データのクォンタイズに使ったが，各演奏者の出力確率分布は，楽譜上に量子化された演奏への人間的なノリの付加（ヒューマナイズ）に利用することもでき，ある演奏者の発音時刻の特徴を反映した演奏の生成が期待できる．

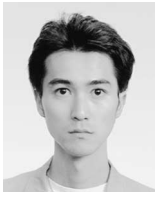
今後，フレーズデータベースの自動作成を実現するため，正解データがなく実演奏データのみしか与えられない場合の，モデルパラメータの推定法について検討していく．また，本研究で扱ったのは，テンポが一定で伴奏の拍位置が明らかな場合での演奏のクォンタイズであるが，可変テンポの場合への対応可能性についても検討していく．

**謝辞** 演奏者として実験に協力していただいた，橋本大輔氏，千田真一氏，斎田康弘氏に感謝いたします．

## 参 考 文 献

1) 浜中雅俊，後藤真孝，大津展之：学習するジャムセッションシステム：演奏者の振る舞いのモデルの獲得，情報処理学会研究報告，2000-MUS-34, Vol.2000, No.19, pp.27–34 (2000).
2) Desain, P. and Honing, H.: The Quantization of Music Time: A Connectionist Approach, *Computer Music Journal*, Vol.13, pp.56–66 (1989).
3) 片寄晴弘，井口征士：自動採譜システム，人工知能学会誌，Vol.5, No.1, pp.59–66 (1990).
4) 後藤真孝：音楽音響信号を対象としたリアルタイムビートトラッキング，人工知能学会研究会資料 AI チャレンジ研究会，SIG-Challenge-9801-2, pp.7–14 (1998).
5) Goto, M. and Muraoka, Y.: An Audio-based Real-time Beat Tracking System and Its Applications, *Proc. ICMC*, pp.17–20 (1998).
6) 後藤真孝：拍節認識（ビートトラッキング），bit 別冊コンピュータと音楽の世界—基礎からフロンティアまで，pp.100–116，共立出版 (1998).
7) Dannenberg, R. and Mont-Reynaud, B.: Following an Improvisation in Real Time, *Proc. ICMC*, pp.241–248 (1987).
8) Allen, P. and Dannenberg, R.: Tracking Musical Beats in Real Time, *Proc. ICMC*, pp.140–143 (1990).
9) 齋藤直樹，中井 満，下平 博，嵯峨山茂樹：隠れマルコフモデルによる音楽演奏からの音符列の推定，情報処理学会研究報告，99-MUS-33, Vol.99, No.106, pp.27–32 (1999).

(平成 13 年 6 月 14 日受付)
(平成 13 年 12 月 18 日採録)



**浜中 雅俊**（学生会員）

1998 年日本大学理工学部精密機械工学科卒業．2000 年筑波大学大学院工学研究科電子・情報工学専攻博士前期課程修了．現在，同専攻博士後期課程在学中．音楽情報処理の研究に興味を持つ．2001 年情報処理学会山下記念研究賞，2001 年 SCI（5th World Multiconference on Systemics, Cybernetics and Informatics）in Art 優秀論文賞各受賞．

**後藤　真孝**（正会員）

1993 年早稲田大学理工学部電子通信学科卒業．1998 年同大学大学院博士後期課程修了．同年，電子技術総合研究所（2001 年に独立行政法人産業技術総合研究所に改組）に入所し，現在に至る．2000 年より科学技術振興事業団さきがけ研究 21 研究員を兼任．博士（工学）．音楽情報処理，音声言語情報処理等に興味を持つ．1992 年 jus 設立 10 周年記念 UNIX 国際シンポジウム論文賞，1993 年 NICOGRAPH'93 CG 教育シンポジウム最優秀賞，1997 年情報処理学会山下記念研究賞，1999 年平成 10 年電気関係学会関西支部連合大会奨励賞，2000 年 WISS2000 論文賞・発表賞，2001 年日本音響学会第 18 回粟屋潔学術奨励賞・第 5 回ポスター賞各受賞．電子情報通信学会，日本音響学会，日本ソフトウェア科学会，日本音楽知覚認知学会，IEEE，ICMA，ISCA 各会員．

**麻生　英樹**

1981 年東京大学工学部計数工学科卒業．1983 年同大学大学院工学系研究科情報工学専攻修士課程修了．同年電子技術総合研究所に入所．現在，独立行政法人産業技術総合研究所情報処理研究部門研究グループ長．知的学習システムの実現に関する研究に従事．電子情報通信学会，人工知能学会，日本神経回路学会，行動計量学会各会員．

**大津　展之**

1969 年東京大学工学部計数工学科卒業．1971 年同大学大学院数理工学専攻修士課程修了．同年電子技術総合研究所入所．以来，パターン認識，画像処理，多変量データ解析，人工知能に関する数理的研究に従事．工学博士．数理情報研究室長，首席研究官，知能情報部長を経て，現在，産業技術総合研究所フェロー．筑波大学連携大学院教授，東京大学大学院情報理工学研究科教授を併任．電子情報通信学会，日本行動計量学会等各会員．