社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－MUS－66 (12)
2006／8／8

Instrogram：発音時刻検出とF0推定の不要な楽器音認識手法

北原 鉄朗† 後藤 真孝‡ 駒谷 和範† 尾形 哲也† 奥乃 博†

†京都大学大学院情報学研究科知能情報学専攻
‡産業技術総合研究所

{kitahara,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp m.goto@aist.go.jp

あらまし 本稿では，Instrogramと呼ばれる楽器存在確率の時間・周波数表現に基づく新たな楽器音認識手法を提案する．従来の多くの楽器音認識では単音を処理単位としていたため，各単音の発音時刻や基本周波数（F0）を正確に推定する必要があった．しかし，混合音におけるそれらの推定は難しく，推定誤りによって楽器音認識精度は大きく低下していた．本手法では，これらの推定をせずに，あらゆるF0に関して楽器存在確率の時系列を求め，Instrogramと呼ばれるスペクトログラムのような視覚表現として可視化する．実験の結果，Instrogramが実際の楽器構成を表していることを確認した．さらに，Instrogram間の類似度計算に基づいた類似楽曲検索も実現した．

Instrogram: Musical Instrument Recognition Method Without Requiring Onset Detection Nor F0 Estimation

Tetsuro Kitahara† Masataka Goto‡
Kazunori Komatani† Tetsuya Ogata† Hiroshi G. Okuno†

†Dept. of Intelligence Science and Technology,
Graduate School of Informatics, Kyoto University
‡National Institute of Advanced Industrial Science and Technology (AIST)

Abstract This paper describes a new musical instrument recognition method based on the time-frequency representation of instrumentation called an *instrogram*. Because the conventional instrument recognition is performed for each note, accurate estimation of the onset time and fundamental frequency (F0) is required. These estimation is, however, difficult in polyphonic music, and thus their errors deteriorated the recognition performance severely. Without these estimation, our method calculates the temporal trajectory of *instrument existence probabilities* for every possible F0 and visualizes them as a spectrogram-like graphical representation, called an *instrogram*. Experimental results show that instrograms represent actual instrumentation. We have also achieved music information retrieval by calculating the similarity between instrograms.

1. はじめに

本研究では，楽器構成に基づいて楽曲を検索する技術の確立を目指している．「ピアノソナタ」「弦楽四重奏」などのような分類が用いられることからも分かるように，「どのような楽器で演奏されているか」は，その楽曲を特徴づける重要な要素といえる．楽器構成に基づく楽曲検索には，2つの方式が考えられる．1つはユーザが楽器を指定する方式である．「弦楽四重奏を聴きたい」などがこれに該当する．もう1つは，いわゆるQuery-by-Example方式である．この方式では，ユーザがある楽曲を指定すると，システムはその楽曲と楽器構成の近い楽曲を探し出す．これは，たとえばBGMのプレイリスト自動生成などに有用な技術である．

このような検索を実現する上でキーとなる技術は，音響信号からの楽器の認識である．楽器音認識は，1990年代までは単一音を扱ったもの[1]がメインであったが，近年になって多重奏を扱ったものが増えつつある．柏野らは，音楽情景分析の処理モデルOPTIMAを構築し，ベイジアンネットに基づく単音や和音，楽器名の認識を実現した[2),3)]．その後，柏野らは適応型混合テンプレート法に基づく楽器音の同定を実現した[4)]．木下

らは，周波数成分の重なり適応処理によって周波数成分が重なったときの性能低下を軽減する方法を提案した[5]．Egginkらは，周波数成分の重なりの問題に対してミッシングフィーチャー理論による対処を試みた[6]．Vincentらは，採譜と楽器音認識を単一の最適化問題として定式化した[7]．Essidらは，個々の楽器ではなく楽器構成をクラスとした識別処理による方法を試みた[8]．また我々は，上述の周波数成分の重なりの問題に対して，重なりによる影響度の大きさに基づいた特徴量の重みづけを実現し，音楽的文脈に基づいて音楽的に不自然な誤認識を削減する方法も提案した[9]．

これらの研究のほとんどに共通していることは，楽器同定処理をフレーム単位か単音（1つの音符に相当する一単位の音）単位で行っていることである．前者[6]の場合，スペクトルの時間変化などの特徴量を扱うのは難しいという問題があった．後者[2)~5),9)]の場合，各単音の発音時刻と基本周波数（F0）を推定する必要があるため，これらの推定誤差があると悪影響を受けるという欠点があった．実際，文献4), 9)で報告されていた実験では，これらは正解を与えていた．

本稿では，発音時刻検出やF0推定によらない新たな楽器音認識手法を提案する．本手法では，各時刻・各F0において各楽器の音が存在する確率（**楽器存在確率**）をスペクトログラムのような視覚表現として可視化する．この可視化表現をInstrogramと呼ぶ．楽器存在確率は，**不特定楽器存在確率**と**条件付き楽器存在確率**の積で表され，前者をPreFEst[10]で，後者を隠れマルコフモデル（HMM）で求める．前者の計算は，従来法における各単音の発音時刻やF0の推定に，後者は楽器の同定に相当する．従来法が前者の結果を使って後者の処理を行うために，後者の精度が前者の精度に大きく依存するのに対し，本手法では，両者の計算を別々に行うため，一方の誤りが他方に影響しない．

2. Instrogram

Instrogramは，楽器構成を表す視覚表現である．対象楽器の各々に1枚ずつ画像が存在し，各画像がその楽器が存在する確率を表す．具体的には，各画像の横軸が時刻，縦軸がF0を表し，各点(t, f)の色の強さによって，時刻tにおいて周波数fをF0とする当該楽器音が存在する確率を表す．**図1**に，「蛍の光」（三重奏）をピアノ，バイオリン，フルートで演奏した音響信号に対して，ピアノ，バイオリン，クラリネット，フルートを対象に求めたInstrogramを示す．ここで，時間分解能は10ms，周波数分解能は100centとした．周波数分解能が高すぎて見にくい場合は**図2**のように，周波数軸をいくつかの区間に分割して区間内の値をマージすることで周波数分解能を粗くすることもできる．図1あるいは図2より，この楽曲は高音部はフルート，中音部はバイオリン，低音部はピアノによる演奏であることがわかる．

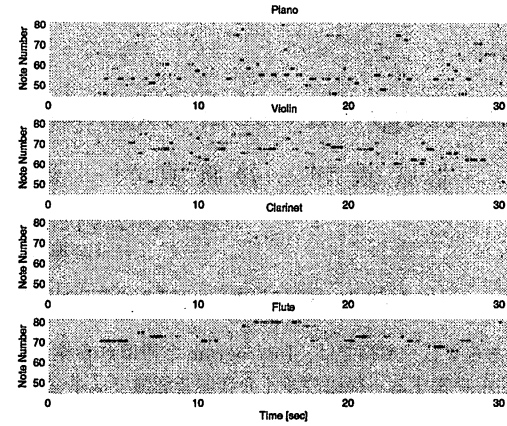


図1 Instrogramの例（ピアノ，バイオリン，フルートによる「蛍の光」（三重奏）．

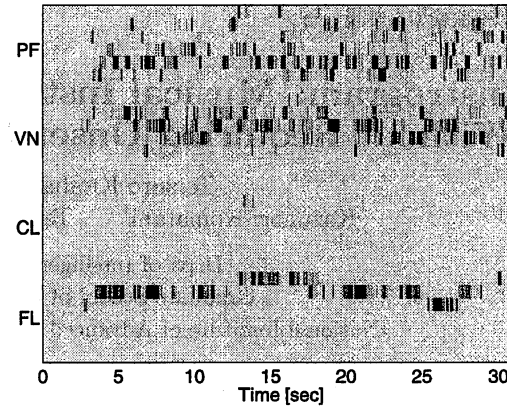


図2 図1の簡略版（低周波数分解能版）

3. Instrogramの作成方法

対象楽器を$\Omega = \{\omega_1, \cdots, \omega_m\}$とすると，ここで求めるべきものは，各$\omega_i \in \Omega$に対する$p(\omega_i; t, f)$である．$p(\omega_i; t, f)$は時刻$t$において$f$をF0とする楽器$\omega_i$の音が存在する確率を表す．いま，同時刻においてF0が同じ音が2つ以上鳴ることはないと仮定する．すなわち，$\forall \omega_i, \omega_j \in \Omega$: $i \neq j \Longrightarrow p(\omega_i \cap \omega_j; t, f) = 0$である．この仮定は，決してすべての楽曲が満たすわけではないが，F0が同じ複数楽器音を各々に分離するのは極めて難しいため，妥当な仮定であると考える．また，$\sum_{\omega_i \in \Omega \cup \{\text{silence}\}} p(\omega_i; t, f) = 1$を満た

す．何らかの楽器の音が存在するという全対象楽器の和事象を $\mathrm{X}(=\omega_1 \cup \cdots \cup \omega_m)$ と書くこととすると，$\omega_i \cap \mathrm{X} = \omega_i \cap (\omega_1 \cup \cdots \cup \omega_i \cup \cdots \omega_m) = \omega_i$ であるので，$p(\omega_i; t, f)$ は次の2つの確率の積で表すことができる：

$$p(\omega_i; t, f) = p(\mathrm{X}; t, f)\ p(\omega_i|\mathrm{X}; t, f).$$

ここで，$p(\mathrm{X}; t, f)$ は**不特定楽器存在確率**といい，時刻 t において f を F0 とする何らかの楽器の音が存在する確率を表し，$p(\omega_i|\mathrm{X}; t, f)$ は**条件付き楽器存在確率**といい，時刻 t において f を F0 とする何らかの楽器の音が存在するとすると，その楽器が ω_i である確率を表す．

3.1 楽器存在確率計算アルゴリズムの概要

楽器存在確率の計算アルゴリズムの概要を**図3**に示す．まず，入力音響信号に対して短時間フーリエ変換を用いてスペクトログラムを作成する．現在の実装ではシフト幅は 10ms，窓幅は 8192 点，窓はハミング窓を用いている．その後，不特定楽器存在確率と条件付き楽器存在確率を計算する．不特定楽器存在確率は，フレーム毎のパワースペクトルに PreFEst[10] を適用することで計算する．一方，条件付き楽器存在確率は，あらゆる周波数に対して，その周波数を F0 とする調波構造の時系列を抽出し，これを隠れマルコフモデル（HMM）でモデル化して計算する．最後に，不特定楽器存在確率と条件付き楽器存在確率の積をとる．

3.2 不特定楽器存在確率の計算

不特定楽器存在確率 $p(\mathrm{X}; t, f)$ は，PreFEst[11] を用いて求める．PreFEst は元々はメロディとベースの F0 を推定する手法であるが，ここでの目的は F0 推定ではなく $p(\mathrm{X}; t, f)$ の計算なので，F0 確率密度関数の計算までの処理（PreFEst-core）のみ用いる．

PreFEst-core では，観測されたパワースペクトルを，ある典型的な調波構造のスペクトルをモデル化した音モデルの加重混合と考える．F0 が F の音モデルを $p(x|F)$ とすると，その加重混合モデルは

$$p(x; \theta^{(t)}) = \int_{\mathrm{Fl}}^{\mathrm{Fh}} w^{(t)}(F) p(x|F) dF,$$

$$\theta^{(t)} = \{w^{(t)}(F) | \mathrm{Fl} \leq F \leq \mathrm{Fh}\}$$

と表される．ここで，Fh と Fl は許容される F0 の上限と下限，$w^{(t)}(F)$ は $\int_{\mathrm{Fl}}^{\mathrm{Fh}} w^{(t)}(F) dF = 1$ を満たす音モデルの重みである．もし，観測されたパワースペクトルが $p(x; \theta^{(t)})$ から生成されたかのようにモデルパラメータ $\theta^{(t)}$ を推定できれば，パワースペクトルが個々の音モデルへ分解されたとみなすことができ，重み $w^{(t)}(F)$ は F を F0 とする音モデルの相対的な優勢さを表していると考えることができる．そこで，この重み $w^{(t)}(f)$ を不特定楽器存在確率 $p(\mathrm{X}; t, f)$ とみなす．この重みは EM アルゴリズムで推定できる[11]．

表1 28次元特徴ベクトルの詳細

スペクトルの時間平均に関する特徴	
1	周波数重心
2	全倍音のパワー値の合計に対する基音成分のパワー値の割合
3–10	全倍音のパワー値の合計に対する i 次までの倍音のパワー値の割合 ($i = 2, 3, \cdots, 9$)
11	奇数次倍音と偶数次倍音のパワー比
12–20	持続時間が，最長の倍音のそれの p%以上ある倍音の個数 ($p = 10, 20, \cdots, 90$)
パワーの時間変化に関する特徴	
21	パワー包絡の近似直線の傾き
22–24	時刻 t から時刻 $t + iT/3$ までのパワー包絡の微分係数の中央値 ($i = 1, \cdots, 3$)
変調に関する特徴	
25, 26	振幅変調の振幅と振動数
27, 28	周波数変調の振幅と振動数

3.3 条件付き楽器存在確率の計算

下限周波数 Fl から上限周波数 Fh まで Δf[cent] ごとに以下の処理を行う．

3.3.1 調波構造抽出

当該周波数 f を F0 とする調波構造の時系列 $H(t, f)$ を抽出する．これは，

$$H(t, f) = \{(F_i(t, f), A_i(t, f)) \mid i = 1, \cdots, h\}$$

と表現される．ここで，$F_i(t, f)$ と $A_i(t, f)$ はそれぞれ時刻 t において f を F0 とする音の i 次倍音の周波数と振幅である．$F_i(t, f)$ は基本的には $i \cdot f$ と等しいが，実際にはビブラートなどの原因により等しくはならない．また，h は 10 とした．

3.3.2 特徴抽出

調波構造 $H(t, f)$ から，長さ T（現在の実装では 500ms）の断片 $H_t(\tau, f)$ $(t \leq \tau < t + T)$ を抽出し，**表1**に示す 28 次元特徴ベクトル $\boldsymbol{x}(t, f)$ を求める．これを音響信号の始めから終わりまで Δt（現在の実装では 10ms）ごとに繰り返すことで特徴ベクトルの時系列を得る．その後，主成分分析を用いて 28 次元を 12 次元（累積寄与率：95%）に圧縮する．

3.3.3 確率計算

特徴ベクトルの時系列 $\{\boldsymbol{x}(t, f) | 0 \leq t \leq t_{\mathrm{end}}\}$ を $m+1$ 個の L-to-R 型 HMM $M_1, \cdots, M_{m+1}$ に基づいて解析する．各 HMM M_i は 15 状態からなり，各楽器 ω_i の音または無音をモデル化する．この HMM がマルコフ連鎖としてつながっており，$\{\boldsymbol{x}(t, f)\}$ がこのマルコフ連鎖から生成されたとみなして，時刻 t において $\boldsymbol{x}(t, f)$ が各 HMM M_i から生成された尤度を求める．この尤度が，求めるべき条件付き楽器存在確率 $p(\omega_i|\mathrm{X}; t, f)$ である．なお，特徴量の音の重なりによる影響を考慮するため，学習データを混合音から創り出す混合音テン

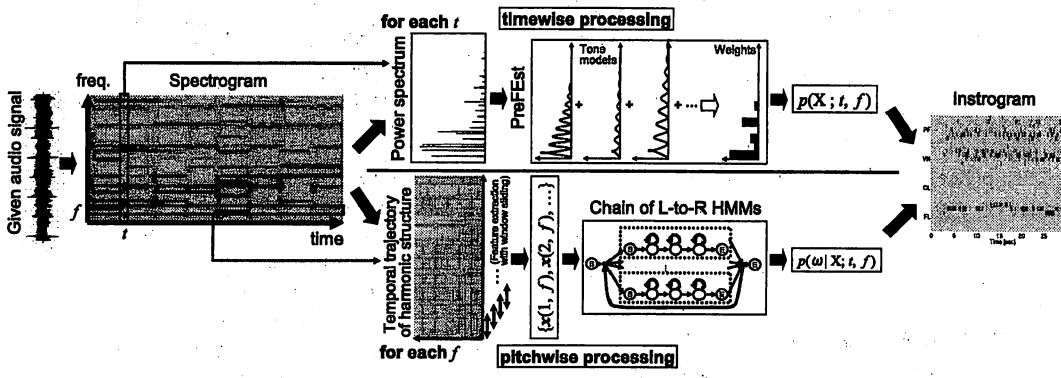


図 3　Instrogram 作成手法の処理の流れ

プレート[9]を用いる．

3.4 Instrogram の簡略化

Instrogram ではあらゆる F0 について楽器存在確率を求めるが，用途によってはこういった詳細な情報は必要ない場合がある．特に検索に用いるのであれば，もっと粗い音域（たとえば，高音域・中音域・低音域）に対して楽器存在確率を計算すれば十分である場合も考えられる．そこで，全周波数区間を N 個の区間 $I_1, \cdots, I_N$ に分割し，k 番目の周波数区間 I_k の楽器存在確率 $p(\omega_i; t, I_k)$ を $p(\omega_i; t; \bigcup_{f \in I_k} f)$ と定義する．これは，周波数軸が実際には離散であることを考慮すると，以下の漸化式で求めることができる：

$$
\begin{aligned}
&p(\omega_i; t, f_1 \cup \cdots \cup f_i \cup f_{i+1}) \\
&= p(\omega_i; t, f_1 \cup \cdots \cup f_i) \;+\; p(\omega_i; t, f_{i+1}) \\
&\quad - \; p(\omega_i; t, f_1 \cup \cdots \cup f_i)\, p(\omega_i; t, f_{i+1}).
\end{aligned}
$$

ここで，$I_k = \{f_1, \cdots, f_i, f_{i+1}, \cdots, f_{n_k}\}$ である．

3.5 イベントベースの表現への変換

Instrogram では，どの楽器による演奏かを決定せずに楽器構成を確率として表現するが，用途によってはそういった決定が必要となる場合がある．そこで，Instrogram から「どの楽器がいつからいつまで演奏されたか」というイベントベースの表現に変換する方法を述べる．この変換は，各状態が各楽器及び無音に対応するマルコフ連鎖（図 4）に基づいて行う．まず，各時刻・各音域において楽器存在確率 $p(\omega_i; t, I_k)$ が最大になる楽器名を求める．そして，音域ごとに，この楽器名の時系列が図 4 のマルコフ連鎖に基づいて生成されたとみなし，最尤パスを Viterbi アルゴリズムで求める．このとき，無音を表す状態から楽器 ω_i を表す状態へ遷移した時刻は楽器 ω_i が演奏を開始した時刻とみなすことができ，逆の遷移をした時刻は演奏を止めた時刻とみなすことができるので，このような状態遷移時刻を求めることにより，各楽器がいつ演奏をはじめていつ終えたかを推定する．

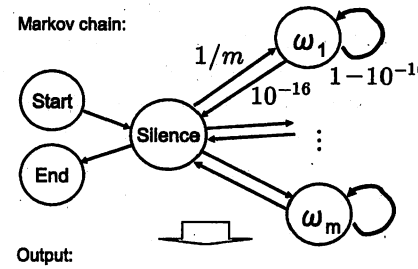


図 4　イベントベースの表現への変換で用いられるマルコフ連鎖

4. Instrogram を用いた類似楽曲検索

Instrogram で採用されたシンボル化せずに楽器構成を表現するアプローチの 1 つのメリットは，類似度を連続値として与えることができるということである．従来の楽器音認識では，認識結果は楽器名あるいはそれを基としたシンボル表現として得られたため，楽器構成の比較は，基本的に一致するかしないかに限られていた．我々は，Instrorgam 間の類似度を連続値として定義し，ユーザに指定された楽曲に楽器構成が近い楽曲を検索する方法を実現する．このような検索は，BGM のためのプレイリスト自動生成や内容に基づく音楽推薦の核となりうる重要な技術である．ここでは，類似度の代わりに，Instrogram 間の距離（非類似度）を DP マッチング（DTW）[12]で求める．

(1) 各時刻 t に対して，ベクトル $\boldsymbol{p}_t$ を全楽器存在確率の連結として求める：
$\boldsymbol{p}_t = (p(\omega_1; t, I_1), p(\omega_1; t, I_2), \cdots, p(\omega_m; t, I_N))'$.
ここで，$'$ は転置を表す．

(2) 2 つのベクトル $\boldsymbol{p}, \boldsymbol{q}$ 間の距離をコサイン距離として定義する：
$\mathrm{dist}(\boldsymbol{p}, \boldsymbol{q}) = 1 - (\boldsymbol{p}, \boldsymbol{q}) / ||\boldsymbol{p}|| \cdot ||\boldsymbol{q}||$.
ただし，$(\boldsymbol{p}, \boldsymbol{q}) = \boldsymbol{p}' R \boldsymbol{q}$, $||\boldsymbol{p}|| = \sqrt{(\boldsymbol{p}, \boldsymbol{p})}$. $R =$

(r_{ij}) は正定値行列で，要素間の関係を定義するものである．たとえば，ある2曲において同じ楽器が異なる音域で演奏されている場合，あるいは同じ楽器族だけど異なる楽器で演奏されている場合，この2曲には高い類似度を与えたい場合があるであろう．そのような場合，対応する r_{ij} に0より大きい値を与えることで，こういった要素間の関係を考慮することができる．R が単位行列のとき，$(\boldsymbol{p}, \boldsymbol{q})$ および $||\boldsymbol{p}||$ は通常の内積およびノルムとなる．

(3) 上で定義した距離尺度を利用して，$\{\boldsymbol{p}_t\}, \{\boldsymbol{q}_t\}$ 間の距離（非類似度）をDTWで計算する．

これまでの音楽情報検索研究[13),14)]においても音色類似度（timbral similarity）と呼ばれるものが用いられてきた．これは，メル周波数ケプストラム係数（MFCC）などのスペクトル包絡に関する特徴量を混合音から直接抽出したものを基に計算されていた．このような特徴量は比較的簡単な処理で求めることができ，一定の成功を収めているが，必ずしも楽器の音色を鮮明に反映するものではなく，後述の実験から示唆されるように，和音のボイシングを始めとする編曲などの他の音楽的要素からも影響を受ける．それに対して，Instrogramは楽器構成を直接表すものであるため，楽器構成の類似度をより的確に計算できる．さらに，Instrogramには次のようなメリットもある．

直感性 音楽的な意味が直感的でわかりやすい．

制御性 R を適切に与えることで，同じ楽器の異なる音域での演奏や同じ楽器族の異なる楽器による演奏に対して，これらの違いを無視したり軽視したりできる．

5. MPEG-7アノテーションへの応用

MPEG-7は，マルチメディアコンテンツに対してその内容をメタデータとして付与することで検索などを高度化するための標準規格である．本節では，Instrogramを基にした音楽アノテーションについて議論する．

InstrogramをMPEG-7のタグとして表す最も簡単な方法は，楽器存在確率をそのままベクトルの時系列として記述することである．前章で述べた類似楽曲検索のように，楽器構成を決定せずに処理を行う場合は，この方法が適切であろう．図5に示した例では，ピアノ（16行目）に対する8次元の楽器存在確率ベクトルの時系列が10ms刻み（6行目）で表されている．各次元は各音域に対応し，ここでは65.5Hzから1048Hzを1/2オクターブごとに分割して得られる8つの音域が用いられている（3〜4行目）．

一方，3.5節の処理で得られるようなイベントベースの表現をMPEG-7タグとして記述することもできる．ユーザが楽器名をクエリーとして指定するようなタスクでは，こちらの記述方式のほうが便利であろう．図6の例では，92〜130Hzの音域において（10行目），ピアノ（12行目）が開始から6.850秒の時点で演奏を始め（4行目），それが0.200秒間続いた（6行目）ことを示している．

```
 1:<AudioDescriptor
 2:    xsi:type="AudioInstrogramType"
 3:    loEdge="65.5" hiEdge="1048"
 4:    octaveResolution="1/2">
 5:  <SeriesOfVector totalNumOfSamples="5982"
 6:      vectorSize="8" hopSize="PT10N1000F">
 7:    <Raw mpeg7:dim="5982 8">
 8:      0.0 0.0 0.0 0.0 0.718 0.017 0.051 0.0
 9:      0.0 0.0 0.0 0.0 0.724 0.000 0.085 0.0
10:      0.0 0.0 0.0 0.0 0.702 0.013 0.089 0.0
11:      0.0 0.0 0.0 0.0 0.661 0.017 0.063 0.0
12:                  ......
13:    </Raw>
14:  </SeriesOfVector>
15:  <SoundModel
16:     SoundModelRef="IDInstrument:Piano"/>
17:</AudioDescriptor>
```

図5 楽器存在確率をそのままMPEG-7タグとして記述した例

```
 1:<MultimediaContent xsi:type="AudioType">
 2:  <Audio xsi:type="AudioSegmentType">
 3:    <MediaTime>
 4:      <MediaTimePoint>T00:00:06:850N1000
 5:                          </MediaTimePoint>
 6:      <MediaDuration>PT0S200N1000
 7:                          </MediaDuration>
 8:    </MediaTime>
 9:    <AudioDescriptor xsi:type="SoundSource"
10:       loEdge="92" hiEdge="130">
11:     <SoundModel
12:      SoundModelRef="IDInstrument:Piano"/>
13:    </AudioDescriptor>
14:  </Audio>
                  ......
```

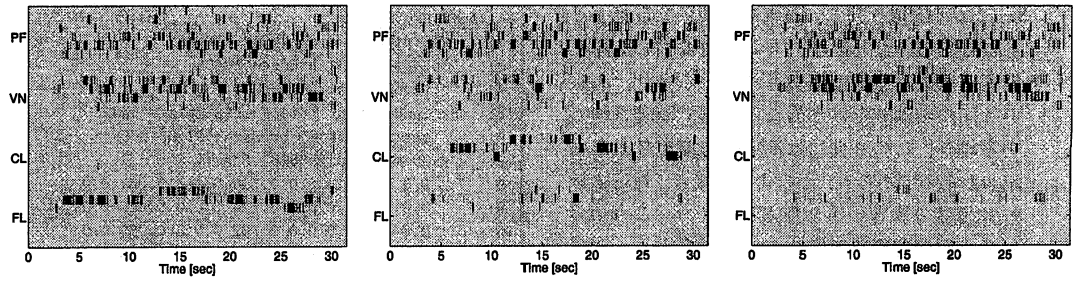
図6 Instrogramから得られるイベントベースの表現をMPEG-7タグとして表した例

表2 実験で用いた実演奏の楽曲およびその楽器構成

Classical	(i) No. 12, 14, 21, 38	Strings
	(ii) No. 19, 40	Piano+Strings
	(iii) No. 43	Piano+Flute
Jazz	(iv) No. 1, 2, 3	Piano solo

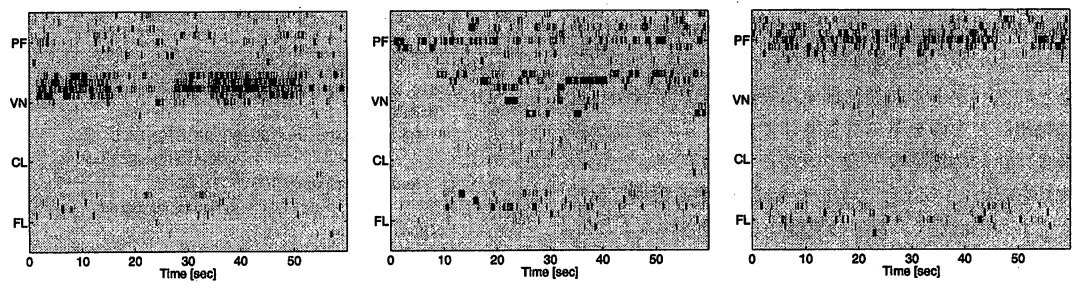
6. 評価実験

音楽音響信号からInstrogramを作成する実験を行った．実験には，楽器音データベースを切り貼りして作成した多重奏の音響信号（以下，切り貼り信号という）と実演奏を収録した音響信号の両方を用いた．切り貼り信号は，「蛍の光」（三重奏）の楽譜に基づいて「RWC研究用音楽データベース：楽器音」（RWC-MDB-I-2001）



(a) FL–VN–PF　(b) VN–CL–PF　(c) VN–VN–PF

図 7 切り貼り信号（三重奏「蛍の光」）から求めた Instrogram．FL–VN–PF は左から高音・中音・低音パートの楽器を表す．紙面の制約から主要 3 パターンのみ掲載し，他は http://winnie.kuis.kyoto-u.ac.jp/~kitahara/instrogram/ に掲載する．



(a) RM-C No. 14 (Str.)　(b) RM-C No. 40 (Str.)　(c) RM-J No. 1 (Pf.+Str.)

図 8 実演奏から求めた Instrogram．RM-C，RM-J はそれぞれ RWC-MDB-C-2001，RWC-MDB-J-2001 を表す．こちらも紙面の制約から 3 曲分のみ掲載し，他は同 URL に掲載する．

に収録されている音響信号（奏法：ノーマル，バリエーション番号：1，強度：中）を切り貼りして作成した．対象楽器は，ピアノ，バイオリン，クラリネット，フルートの 4 つとし，音域的に発音可能な範囲で総当たりの組み合わせとした．実演奏には，「RWC 研究用音楽データベース」のうちクラシック（RWC-MDB-C-2001），ジャズ（RWC-MDB-J-2001）のものから選んだ比較的小編成の 10 曲（**表 2**）を用いた．学習データは，音の重なりによる特徴変動に対する頑健性向上のために多重奏の音響信号から作成した[9]．この音響信号は，RWC-MDB-C-2001 の楽曲番号 13 の SMF に従って RWC-MDB-I-2001 の音響信号（奏法：ノーマル，バリエーション番号：2・3，強度：強・中・弱）を用いて作成した．ただし，実演奏に対する実験では，その音響的多様さに対応するため，すべてのバリエーション番号（1〜3）の音響信号の他，他の楽器音データベース（NTTMSA-P1）も用いた．

Instrogram 作成結果を**図 7**，**図 8** に示す．切り貼り信号（図 7）については，(a) と (b) を比べると，(a) はフルートの楽器存在確率が高くてクラリネットの楽器存在確率が非常に低い（ほとんど 0）のに対し，(b) はクラリネットの楽器存在確率が高くてフルートの楽器存在確率は非常に低かったことがわかる．(c) では，バイオリンとピアノの楽器存在確率が高く，他の楽器存在確率はほぼ 0 であった．実演奏（図 8）については，(a) はバイオリンの楽器存在確率が高く，(c) はピアノの楽器存在確率が高かった．(a), (c) ではこれ以外の楽器存在確率は十分に低かった．(b) はピアノとバイオリンの両方の楽器存在確率が高くなっているが，ピアノの楽器存在確率は演奏開始すぐに高くなるのに対し，バイオリンの楽器存在確率は演奏開始 10 秒後に高音部で高くなり，その後，他の音域でも高くなる，という違いがあった．この結果は，この楽曲における両楽器の実際の演奏を正しく反映している．

こうして得られた Instrogram に基づき，3.5 節で述べたイベントベースの楽器記述の実験を行った．評価には次式を用いる：

$$\frac{1}{m}\sum_{\omega_i\in\Omega}\sum_{k=1}^{N}\frac{\left(\begin{array}{c}\text{音域 } I_k \text{ で楽器 } \omega_i \text{ による演奏と}\\ \text{正しく判定されたフレーム数}\end{array}\right)}{\left(\begin{array}{c}\text{音域 } I_k \text{ で楽器 } \omega_i \text{ による演奏と}\\ \text{判定された全フレーム数}\end{array}\right)}.$$

実験結果を**図 9**，**図 10** に示す．平均精度は切り貼り音で 81.8%，実演奏で 76.2%であった．切り貼り音は

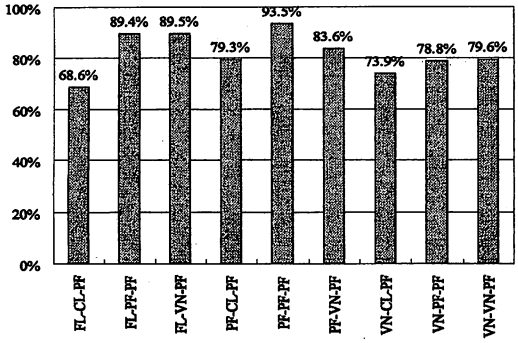


図 9　切り貼り信号に対してイベントベースの楽器記述をした結果

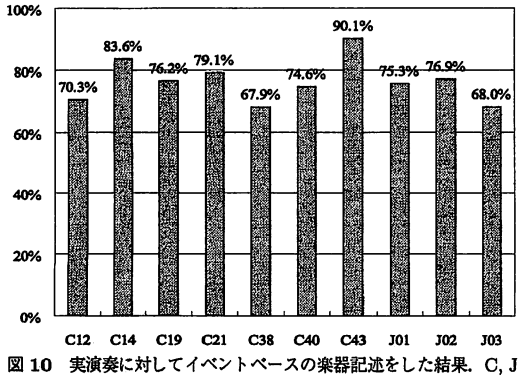


図 10　実演奏に対してイベントベースの楽器記述をした結果．C, J は表 2 中のジャンルを，それに続く数字は楽曲番号を表す．

FL–CL–PF 以外のすべての場合で精度が 70%を超え，実演奏では 10 曲中 8 曲が 70%を超える精度を示した．

次に，Instrogram 間の（非）類似度計算に関する実験を実演奏の 10 曲を用いて行った．R には単位行列を用いた．実験結果を表 3(a) に示す．表から類似度計算結果は次のようにまとめられる．

- グループ内の非類似度はおおむね 7000 以下であった（グループ (ii) を除く）．
- グループ (i)（弦楽）とグループ (iv)（ピアノ）の間の非類似度はおおむね 9000 以上であった．なかには 10000 を超えるものもあった．
- グループ (i) とグループ (ii)（ピアノ+弦楽），グループ (ii) とグループ (iii)，グループ (ii) とグループ (iv) 間の非類似度はおおよそ 8000 程度であった．これらの組み合わせでは 1 つの楽器だけが共通に使われているので，この結果は妥当なものといえる．
- グループ (iii) とグループ (iv) 間の非類似度はおおよそ 7000 程度であった．これらの間の違いはフルートの有無だけであるので，この結果も妥当なものといえる．

比較のため，MFCC を用いて非類似度を計算した結果を表 3(b) に示す．両者の比較により次のことがわかる．

- グループ (i) 内での非類似度とグループ (i) と他のグループ間での非類似度とでは，Instrogram を用いた場合ははっきりと両者に違いがあったのに比べ，MFCC の場合にはあまり差異はなかった．実際，グループ (i) の楽曲の各々に対して類似楽曲ベスト 3 を求めたところ，Instrogram を用いた場合はすべてグループ (i) の楽曲が選ばれたのに対して，MFCC を用いた場合はグループ (i) 以外の楽曲も含まれていた．
- 弦楽器を含まない楽曲（グループ (iii) および (iv)）に対して類似楽曲ベスト 3 を求めると，Instrogram の場合はすべて弦楽器を含まない楽曲だったのに対して，MFCC を用いた場合は弦楽器を含む楽曲（C14，C21 など）も選ばれていた．

また，この結果に基づき，指定された楽曲に楽器構成が類似する楽曲を検索するプロトタイプを構築した．ユーザがある楽曲を指定すると，システムはその楽曲とデータベース中の各曲との間の楽器構成の類似度を 5 章で述べた方法で計算し，類似度が高い順に楽曲を並べて表示する．ユーザがそのなかから 1 曲選ぶと，システムは楽器存在確率をリアルタイムに表示しながら指定された楽曲を再生する．この表示は，各楽器の存在確率が棒グラフとして表され，音楽の再生に合わせて時々刻々と変化するもので，よく音楽プレーヤーに搭載されているディスプレイ上のスペクトル表示を楽器存在確率に置き換えたようなものである．

7. おわりに

本稿では，Instrogram という楽器存在確率の時間・周波数表現に基づく新たな楽器音認識手法を提案した．我々は，楽器存在確率を「各時刻，各周波数において当該楽器の音が存在する確率」と定義し，楽器音認識を「楽器存在確率を対象楽器の各々に対し，全時刻，全周波数に渡って網羅的に求めていく問題」ととらえて定式化した．楽器存在確率を不特定楽器存在確率と条件付き楽器存在確率の 2 つの積で定義し，それぞれを PreFEst および HMM でモデル化した．このように，単音を処理単位としないため，多くの従来手法のように各単音の発音時刻や F0 を推定する必要がなく，これらの推定エラーによる悪影響を回避することを可能にした．また，この楽器存在確率の時間・周波数表現に対して類似度を定義することで，楽器構成の類似度に基づく音楽検索も実現した．MFCC との比較により，Instrogram 間の類似度が，MFCC 間の類似度よりも楽器構成の類似度を正確に表していることを確認した．

表 3 Instrogram 間の非類似度計算結果．(i)～(iv) は表 2 におけるグループに対応する．

(a) Instrogram（楽器存在確率）を用いた場合

	(i)				(ii)		(iii)	(iv)			3-best-similarity
	C12	C14	C21	C38	C19	C40	C43	J01	J02	J03	pieces
C12	0										C21, C14, C38
C14	6429	0									C21, C12, C38
C21	5756	5734	0								C14, C12, C38
C38	7073	6553	6411	0							C21, C14, C38
C19	7320	8181	7274	7993	0						C21, C12, C38
C40	8650	8353	8430	8290	8430	0					J02, J01, C43
C43	8910	9635	9495	9729	8148	8235	0				J01, J02, J03
J01	9711	10226	10252	10324	8305	8214	6934	0			J02, J03, C43
J02	9856	10125	10033	10610	8228	8139	7216	6397	0		J01, C43, J03
J03	9134	9136	8894	9376	8058	8327	7480	6911	7223	0	J01, J02, C43

(b) MFCC を用いた場合

	(i)				(ii)		(iii)	(iv)			3-best-similarity
	C12	C14	C21	C38	C19	C40	C43	J01	J02	J03	pieces
C12	0										C21, C40, J02
C14	17733	0									C43, C12, J02
C21	17194	18134	0								C12, J01, J02
C38	18500	18426	18061	0							J01, J02, C21
C19	17510	18759	18222	19009	0						J02, C12, J03
C40	17417	19011	18189	19099	18100	0					C12, J02, J01
C43	18338	17459	17728	18098	18746	18456	0				J01, C14, J02
J01	17657	17791	17284	17834	18133	17983	16762	0			J02, C43, J03
J02	17484	17776	17359	18009	17415	17524	17585	15870	0		J01, J03, C21
J03	17799	18063	17591	18135	17814	18038	17792	16828	16987	0	J01, J02, C21

本研究のアプローチは，後藤が提唱した楽譜表現によらない音楽理解の立場[15]からも興味深い．人は，楽譜を思い浮かべなくてもどんな楽器の音かを聞き分けることができるが，従来の単音ベースの楽器音認識は，楽譜に相当する情報の抽出を前提としていた．本手法はこうした立場からの楽器音認識の一実現例と言うことができる．今後，本手法をベースに，より高度な音楽理解・楽器音認識の実現に取り組んでいきたい．

謝辞 本研究の一部は，日本学術振興会科学研究費補助金基盤研究，および 21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」によるものである．また，本研究の実験において，「RWC 研究用音楽データベース」（RWC-MDB-C-2001, RWC-MDB-J-2001, RWC-MDB-I-2001）および楽器音データベース NTTMSA-P1 を使用した．NTTMSA-P1 の使用許可をくださった NTT コミュニケーション基礎研究所に感謝する．最後に，ご討論いただいた渡部晋治氏（NTT）に感謝する．

参 考 文 献

1) Martin, K. D.: *Sound-Source Recognition: A Theory and Computational Model*, PhD Thesis, MIT (1999).
2) 柏野邦夫, 中臺一博, 木下智義, 田中英彦: 音楽情景分析の処理モデル OPTIMA における単音の認識, 信学論, **J79-D-II**, 11, pp. 1751–1761 (1996).
3) 柏野邦夫, 中臺一博, 木下智義, 田中英彦: 音楽情景分析の処理モデル OPTIMA における和音の認識, 信学論, **J79-D-II**, 11, pp. 1762–1770 (1996).
4) 柏野邦夫, 村瀬洋: 適応型混合テンプレートを用いた音源同定, 信学論, **J81-D-II**, 7, pp. 1510–1517 (1998).
5) 木下智義, 坂井修一, 田中英彦: 周波数成分の重なり適応処理を用いた複数楽器の音源同定処理, 信学論, **J83-D-II**, 4, pp. 1073–1081 (2000).
6) Eggink, J. and Brown, G. J.: Application of Missing Feature Theory to the Recognition of Musical Instruments in Polyphonic Audio, *Proc. ISMIR* (2003).
7) Vincent, E. and Rodet, X.: Instrument Identification in Solo and Ensemble Music using Independent Subspace Analysis, *Proc. ISMIR*, pp. 576–581 (2004).
8) Essid, S., Richard, G. and David, B.: Instrument Recognition in Polyphonic Music, *Proc. ICASSP*, **III**, pp. 245–248 (2005).
9) Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H. G.: Instrument Identification in Polyphonic Music: Feature Weighting with Mixed Sounds, Pitch-dependent Timbre Modeling, and Use of Musical Context, *Proc. ISMIR*, pp. 558–563 (2005).
10) Goto, M.: A Real-time Music-scene-description System: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals, *Speech Comm.*, **43**, 4, pp. 311–329 (2004).
11) 後藤真孝, 橋口博樹, 西村拓一, 岡隆一: RWC 研究用音楽データベース：研究目的で利用可能な著作権処理済み楽曲・楽器音データベース, 情処学論, **45**, 3, pp. 728–738 (2004).
12) Myers, C. S. and Rabiner, L. R.: A Comparative Study of Several Dynamic Time-warping Algorithms for Connected Word Recognition, *The Bell Syst. Tech. J.*, **60**, 7, pp. 1389–1409 (1981).
13) Tzanetakis, G. and Cook, P.: Musical Genre Classification of Audio Signals, *IEEE Trans. Speech Audio Process.*, **10**, 5, pp. 293–302 (2002).
14) Aucouturier, J.-J. and Pachet, F.: Music Similarity Measure: What's the Use?, *Proc. ISMIR*, pp. 157–163 (2002).
15) 後藤真孝: リアルタイム音楽情景記述システム：全体構想と音高推定手法の拡張, 情処研報, 2000-MUS-37, pp. 9–16 (2000).