Instrogram：楽器存在確率に基づく音楽視覚表現法*

◎北原 鉄朗（京大），後藤 真孝（産総研），
△駒谷 和範，△尾形 哲也，△奥乃 博（京大）

1 はじめに

自動採譜や音楽情報検索などにおいて重要なタスクである楽器音の認識は，これまで主に単一音を対象に研究されてきた [1] が，近年ようやく多重奏を対象とした研究が増えつつある．多重奏に対する楽器音認識の典型的な手法 [2–5] では，単音（1 つの音符に相当する一単位の音）ごとに楽器を認識するので，まず各単音の調波構造を切り出す必要がある．そのためには，各単音のオンセット時刻と基本周波数（F0）を正確に推定しなくてはならない．多重奏におけるこれらの推定は難しく，ロバストな楽器音認識実現において課題となっていた．

本稿では，オンセット検出と F0 推定の不要な新たな楽器音認識手法を提案する．本手法では，楽器音認識を単音ごとに行うのではなく，**楽器存在確率**を時間・周波数平面上に可視化する．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[6] 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オンセット時刻や 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，スペクトログラムに似た楽器存在確率の視覚表現である．解析対象となる楽器ごとに 1 つの画像が存在し，各画像は，横軸が時刻，縦軸が周波数を表し，各時刻，各周波数でその楽器が演奏されている確率を表す．Fig. 1 に例を示す．これは，ピアノ，バイオリン，フルートによる「蛍の光」の三重奏を，ピアノ，バイオリン，クラリネット，フルートを対象に Instrogram を作成したものである．ここで，時間分解能は 10ms，周波数分解能は 20cent とした．周波数分解能が高すぎて見にくい場合は Fig. 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．Fig. 1 あるいは Fig. 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．

3 Instrogram の作成手法

対象楽器を $\Omega = \{\omega_1, \cdots, \omega_m\}$ とすると，Instrogram は，各 $\omega \in \Omega$ に対して $p(\omega; t, f)$ を可視化したものである．ここで，$p(\omega; t, f)$ は時刻 t において f を F0 とする楽器 ω の音が存在する確率を表し，$\sum_{\omega \in \Omega \cup \{\mathrm{silence}\}} p(\omega; t, f) = 1$ 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，$p(\omega; t, f)$ は次式

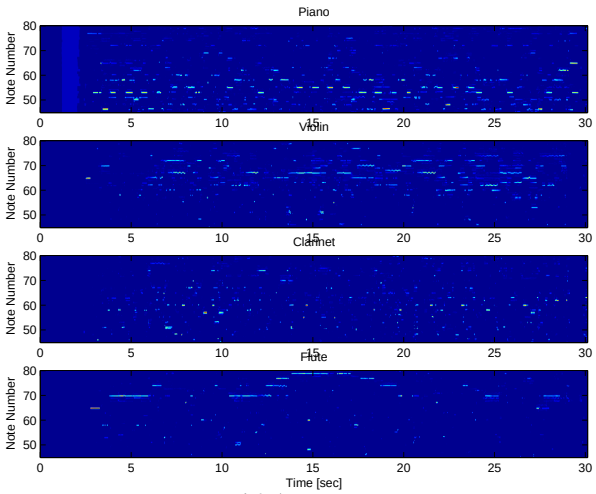

Fig. 1 Instrogram の例（ピアノ，バイオリン，フルートによる「蛍の光」の三重奏）

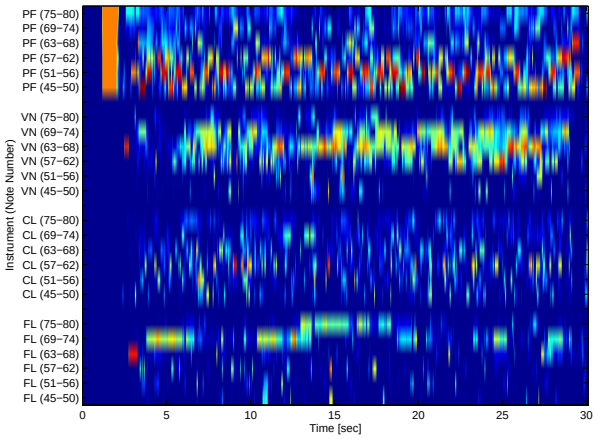

Fig. 2 Fig. 1 の簡略版（低周波数分解能版）

のように 2 つの確率の積で表すことができる：

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

ここで，$p(\mathrm{X}; t, f)$ は**不特定楽器存在確率**と呼ばれ，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，$p(\omega | \mathrm{X}; t, f)$ は**条件付き楽器存在確率**と呼ばれ，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 である確率を表す．以下，各々の計算法について述べる．

3.1 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[6] 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

*Instrogram: Music Visual Representation based on Instrument Existence Probability.
by Tetsuro Kitahara (Kyoto University), Masataka Goto (AIST), Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno (Kyoto University)

Table 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

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

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

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

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 [6].

3.2　条件付き楽器存在確率の計算

条件付き楽器存在確率は，以下のように求める．

3.2.1　短時間フーリエ変換

入力音響信号に対して短時間フーリエ変換を行う．シフト幅は 10ms として，8192 点ハミング窓を用いた．

3.2.2　調波構造抽出

許容されるすべての周波数 f に対して，f を F0 とする調波構造（10 次倍音まで）の時系列 $H(t,f)$ を抽出する．以下の処理はすべて周波数 f ごとに行う．

3.2.3　特徴抽出

調波構造 $H(t,f)$ から，長さ T（現在の実装では 50ms）の断片 $H_t(\tau,f)$ $(t \le \tau < t+T)$ を抽出し，ここから Table 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．

3.2.4　確率計算

楽器$\omega_1, \cdots, \omega_m$および silence の各々に対して 15 状態からなる L-to-R 型 HMM を用意し，特徴ベクトルの時系列$\boldsymbol{x}(t,f)$がこの $m+1$ 個の HMM のマルコフ連鎖から生成されたとみなす．このとき，各 HMM M_i $(i = 1, \cdots, m, \mathrm{silence})$ に対して，$\boldsymbol{x}(t,f)$ が時刻 t において M_iから生成された確率 $p(\boldsymbol{x}(t,f)|M_i;t)$ は，時刻 t において f を F0 とする楽器音がω_iである確率を表す．すなわち，条件付き楽器存在確率 $p(\omega_i|\mathrm{X};t,f)$ は $p(\boldsymbol{x}(t,f)|M_i;t)$ として計算できる．

Table 2　楽器同定結果（フレーム単位の認識率）

PF–PF–PF	97.1%	VN–CL–PF	73.2%
PF–VN–PF	84.7%	FL–PF–PF	82.8%
PF–CL–PF	86.9%	FL–VN–PF	78.5%
VN–PF–PF	81.1%	FL–CL–PF	78.7%
VN–VN–PF	86.7%		

4　実験

RWC 研究用音楽データベース（楽器音）[7] の音響信号（ピアノ (PF)，バイオリン (VN)，クラリネット (CL)，フルート (FL)）をスタンダード MIDI ファイル（SMF）に従って切り貼りした三重奏の音響信号から Instrogram を作成する実験を行った．SMF には，柏野らが用いた「蛍の光」の楽譜 [2] に基づいて作成したものを用いた．HMM には HTK 3.0 を用いた．

Instrogram 作成結果は紙面の制約から省略し，http://winnie.kuis.kyoto-u.ac.jp/~kitahara/instrogram/ に掲載する．FL, CL, PF による三重奏では，FL の楽器存在確率が高い値を示したのに対し，VN, CL, PF による三重奏では，FL の存在確率は非常に低い値（ほぼ 0）となった．VN, VN, PF による三重奏では，VN と PF の楽器存在確率が高い値となり，それ以外の楽器存在確率は低い値となった．同様に，PF のみによる演奏では，PF の楽器存在確率が高い値となり，それ以外の楽器存在確率は低い値となった．

次に，得られた Instrogram を用いてフレーム単位での楽器同定を行った．時刻t, 周波数fごとに $p(\omega;t,f)$ が最大となる楽器名を求め，周波数 f ごとに，この楽器名の時系列が楽器$\omega_1, \cdots, \omega_m$および silence の $m+1$ 状態からなるマルコフ連鎖から生成されたとみなし，Viterbi アルゴリズムで最尤パスを求めた．その結果，Table 2 に示す認識率が得られた．

5　おわりに

本稿では，Instrogram という楽器存在確率の視覚表現法と，それに基づく，オンセット検出・F0 推定の不要な新たな楽器音認識手法を提案した．今後は，楽器存在確率の計算精度の向上に加え，Instrogram を活用した新たな音楽検索について検討していく．

謝辞　本研究の一部は，日本学術振興会科学研究費補助金，21 世紀 COE プログラム，CREST の支援を受けた．

参考文献

[1] K. D. Martin: Sound-Source Recognition: A Theory and Computional Model, PhD Thesis, MIT, 1999.
[2] 柏野 他：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，**J-79-D-II**, 11, pp.1751–1761, 1996.
[3] K. Kashino et al.: A Sound Source Identification System for Ensemble Music based on Template Adaptation and Music Stream Extraction, Speech Comm., **27**, pp.337–349, 1999.
[4] 木下 他：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.
[5] 北原 他：混合音からの特徴量テンプレート作成と音楽的文脈利用による多重奏の音源同定，音講論集（秋），3-10-15，2005.
[6] M. Goto: A Real-time Music-scene-description system: Predominant-F0 Estimation for Detecting Melody and Bass Lines in Real-world Audio Signals, Speech Comm., **43**, pp.311–329, 2004.
[7] 後藤 他：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.