Vol.2015-MUS-107 No.48
2015/5/24

特定周波数帯域におけるパワーの立ち上がりに着目したドラムスの自動採譜の検討

栗脇 隆宏[1,a)]　西野 隆典[1,b)]　成瀬 央[1,c)]

概要：本研究では，様々な楽器音が混在した音楽信号を対象としてスネアドラムとバスドラムの自動採譜に取り組む．提案手法では，対象楽曲に使用されているスネアドラムとバスドラムの特徴を表す周波数成分をそれぞれ 1 箇所ずつ検出し，検出したピーク周辺の周波数帯域におけるスペクトログラムのパワーの立ち上がりに着目した認識対象楽器のスペクトルのモデルを作成する．その後各時刻においてモデルとスペクトルとの距離を計算し，閾値処理を行うことでドラムスの発音時刻検出を行う．RWC 音楽データベース中のポピュラー音楽 5 曲を対象とした評価実験より，F 値の平均がスネアドラムでは 0.95，バスドラムでは 0.93 という結果が得られた．

1. はじめに

自動採譜技術とは，音楽データから自動で楽譜を作成する技術である．楽譜の作成は，ある程度の経験や知識が必要となり，その技能を持つ人にとっても時間と労力がかかる作業である．自動採譜技術が実現されれば，誰でも手軽に時間をかけず楽譜を作成できるようになり演奏活動や作曲活動の支援につながる．

自動採譜の研究は数多く行われているが [1]，研究の対象として調波構造を持つ楽器を扱ったもの [2][3] が多く，打楽器などの調波構造を持たない楽器を対象とした研究は少ない．しかし，打楽器はポップスやロックなどといった現代のポピュラー音楽において重要な役割を果たしている楽器であり，その自動採譜技術は必要とされているといえる．

本研究では，調波構造を持つ楽器と調波構造を持たない楽器が混在した音楽信号を対象として，打楽器の自動採譜に取り組む．本報告ではスネアドラムとバスドラムを認識の対象として扱う．打楽器の自動採譜を行う上での主な問題点は，打楽器の音色は個体差が大きくそれらをすべてカバーする音テンプレートを事前に用意できないこと，混合音中から正しく打楽器の音を認識するのが困難であることの 2 点である [4]．

打楽器を対象とした自動採譜の先行研究として，吉井らが提案するテンプレート適応を利用したドラムスの音源同定 [4][5] が挙げられる．吉井らのシステムでは，各ドラム音ごとの基本テンプレートを対象楽曲に使用されているドラム音に適応させることや，距離尺度を改良したテンプレートマッチング手法を用いることで上記の問題点に対応している．この手法では，バスドラムとスネアドラムの平均認識率は 85 %という結果であった．しかし，テンプレートマッチング手法は，しばしば，少数のパターンに過剰に適応してしまい汎化能力に問題が生じるといった問題点がある．そこで本研究では，より安定した認識結果を得るための手法として，認識対象楽器の特徴を表す成分が多く含まれる周波数帯域を推定した後，帯域を限定したスペクトログラムにおける認識対象楽器のスペクトルモデルを作成し，その後各時刻においてモデルとスペクトルとの距離を計算して閾値処理を行うことでドラムスの発音時刻検出を行う手法を提案する．

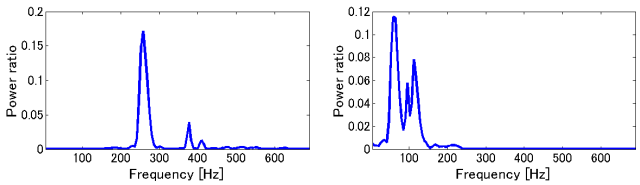
2. ドラムス自動採譜手法

提案手法によるドラムス自動採譜システムの構成について述べる．

2.1 提案手法

スネアドラムとバスドラムのパワースペクトルの例を図 1 に示す．これらの図を見るとわかるように，スネアドラムやバスドラムといった膜鳴楽器のパワースペクトルは，ある周波数にピークを持つという特徴がある．予備実験において数種類のスネアドラムとバスドラムのパワースペクトルのピークを調査した結果，スネアドラムは共通して 200 ～ 300 Hz 付近にピークが存在し，バスドラムは 50 ～

[1] 三重大学大学院工学研究科
Graduate School of Engineering, Mie University
[a)] kuriwaki@pa.info.mie-u.ac.jp
[b)] nishino@pa.info.mie-u.ac.jp
[c)] naruse@pa.info.mie-u.ac.jp



(a) スネアドラム　(b) バスドラム

図 1　パワースペクトルの例

150 Hz 付近にピークが存在することが確認された．

本研究ではこの特徴に着目してドラムスの発音時刻検出を行う．このピーク位置の周波数を以下では特徴周波数と記述する．この特徴周波数は，ドラムスの胴の直径や打面の膜を張る強さなどの物理的構造の違いによって変化するが，通常，スネアドラムやバスドラムは同一楽曲内では同じものが使用されることが多いため，本研究では，スネアドラムやバスドラムの特徴周波数は，同一楽曲の中では一定であると仮定する．提案するドラムス自動採譜手法では，音楽信号から認識対象楽器の特徴周波数を検出し，その周辺の周波数帯域に限定したスペクトログラムを利用して，認識対象楽器の特徴周波数付近におけるスペクトルモデルを作成する．その後，各時刻においてモデルとスペクトルとの距離を計算し，閾値処理を行うことでドラムスの発音時刻検出を行う．提案するドラムス自動採譜手法は大まかに以下の部分から構成されている．

(1) 時間周波数解析
(2) 特徴周波数の検出
(3) スペクトルモデルの作成
(4) 発音時刻検出

次節以降，各処理の詳細を述べる．

2.2　時間周波数解析

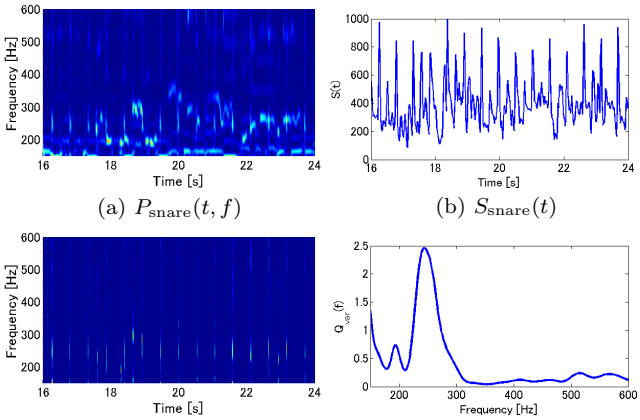
時間周波数解析には，以下の式で定義される Gabor Wavelet 解析を用いた．

$$WT(b,a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\overline{\psi(\frac{t-b}{a})}\, dt \tag{1}$$

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{t^2}{2\sigma^2}} e^{i\omega t} \tag{2}$$

ここで，$x(t)$ は入力信号，a，b はそれぞれ $\psi(t)$ の伸縮変形，平行移動を行う変数である．ω は角周波数であり，a を周波数の逆数とする場合 $\omega = 2\pi$ である．σ は時間周波数の解像度比率を決定するパラメータであり，本研究では $\sigma = 4.0$ とした．

時間分解能は共通で 20 ms とし，解析する周波数帯域は各楽器の特性を考慮して，スネアドラムに関しては 150 Hz から 10 cent 刻みで 2 オクターブの範囲（150 ～ 600 Hz），バスドラムに関しては 25 Hz から 10 cent 刻みで 3 オクターブの範囲 (25 ～ 200 Hz) とした．cent とは，平均律の半音を 100 段階に分けたピッチの単位である．ここで，



(a) $P_{\mathrm{snare}}(t,f)$　(b) $S_{\mathrm{snare}}(t)$

(c) $Q(t,f)$　(d) $Q_{\mathrm{var}}(f)$

図 2　特徴周波数検出処理

時刻 t，周波数 f におけるパワーをそれぞれ $P_{\mathrm{snare}}(t,f)$，$P_{\mathrm{bass}}(t,f)$ で表す．ただし，$P_{\mathrm{snare}}(t,f)$ では 150 Hz を基準として 10 cent 刻みに 2 オクターブの範囲をとるので，$f = 0$ が 150 Hz に対応し，$f = 240$ が 600 Hz に対応する．$P_{\mathrm{bass}}(t,f)$ では，25 Hz を基準として 10 cent 刻みに 3 オクターブの範囲をとるので $f = 0$ が 25 Hz に対応し，$f = 360$ が 200 Hz に対応する．

2.3　特徴周波数の検出

$P_{\mathrm{ins}}(t,f)(\mathrm{ins} = \mathrm{snare\ or\ bass})$ に対して，パワーの合計値が閾値を超える時刻のスペクトル断片のみを取り出すことで打楽器成分を強調し，その後各周波数における時間方向の分散を求めると認識対象楽器による成分を多く含む周波数では分散が大きくなるため，このことを利用して特徴周波数を検出する．具体的な処理を以下に示す．

(1) 各時刻 t について，パワーの総和 $S_{\mathrm{snare}}(t)$，$S_{\mathrm{bass}}(t)$ を次のように定義する．

$$S_{\mathrm{snare}}(t) = \sum_{f=0}^{240} P_{\mathrm{snare}}(t,f) \tag{3}$$

$$S_{\mathrm{bass}}(t) = \sum_{f=0}^{360} P_{\mathrm{bass}}(t,f) \tag{4}$$

ここで，例として図 2(a) の $P_{\mathrm{snare}}(t,f)$ に対して $S_{\mathrm{snare}}(t)$ を求めた結果を図 2(b) に示す．

(2) $S_{\mathrm{ins}}(t)(\mathrm{ins} = \mathrm{snare\ or\ bass})$ の，値の大きさの上位 1%の平均値を求め，それを 0.8 倍した値を閾値 θ とする．そして，閾値 θ より上に存在し，極大値をとる時刻を検出する．図 2(b) の例では，$\theta = 587.5$ という値が求まるので，587.5 より上に存在する極大値をとる時刻が検出される．極大値は，3 フレームの移動平均フィルタを 1 回適応した後，連続する 3 フレームにおいて $S_{\mathrm{ins}}(t-1) < S_{\mathrm{ins}}(t)$ かつ $S_{\mathrm{ins}}(t) > S_{\mathrm{ins}}(t+1)$ を満たすこととした．上記の条件を満たす時刻における値のみを残したスペクトログラムを $Q(t,f)$ と定義す

る．図 2(a) の $P_{\mathrm{snare}}(t,f)$ に対して $Q(t,f)$ を求めた結果を図 2(c) に示す．

(3) 特徴周波数を検出するために，$Q(t,f)$ の各周波数における時間方向の分散の値 $Q_{\mathrm{var}}(f)$ を求めて，その結果を利用する．$Q_{\mathrm{var}}(f)$ は式 (5) で求める．ここで，n は入力信号のスペクトログラムにおける時間方向のフレーム数である．図 2(c) の $Q(t,f)$ に対して $Q_{\mathrm{var}}(f)$ を求めた結果を図 2(d) に示す．

$$Q_{\mathrm{var}}(f) = \frac{1}{n}\sum_{t=0}^{n}\left(Q(t,f) - \frac{1}{n}\sum_{t=0}^{n}Q(t,f)\right)^2 \quad (5)$$

(4) 求めた $Q_{\mathrm{var}}(f)$ に対して Savitzky-Golay の方法 [6] による 25 フレームの平滑化微分を用いて，ピークの検出を行う．検出されたピークのうち，最大値をとる周波数を特徴周波数 F として検出する．

2.4 スペクトルモデルの作成

楽曲中で使われている認識対象楽器の特徴周波数周辺のスペクトル形状のモデルを作成する．ここで，スペクトログラム $P(t,f)$ に対して，各時刻において特徴周波数 F の前後 25 フレームのみを取り出したスペクトログラムを $P'(t,f)$ とする．そして，スペクトルの細かな変動をなくすために $P'(t,f)$ に対して，周波数方向に前後 5 フレームの移動平均フィルタを 3 回繰り返し適応する．その結果のスペクトログラムを $\hat{P}(t,f)$ とする．

2.4.1 スペクトル断片の選択

$\hat{P}(t,f)$ から，認識対象楽器が含まれているスペクトル断片を以下の手順で選択していく．

(1) 各時刻 t について，$\hat{S}(t)$ を次のように定義する．

$$\hat{S}(t) = \sum_{f=F-25}^{F+25}\hat{P}(t,f) \quad (6)$$

(2) $\hat{S}(t)$ の，値の大きさの上位 1%の平均値を求め，0.8 倍した値を閾値 $\hat{\theta}$ とする．そして，$\hat{S}(t)$ に対して閾値 $\hat{\theta}$ 以上の極大値をとる時刻を検出していく．極大値の検出条件は，2.3 節 (2) と同様である．この条件を満たした時刻におけるスペクトル断片を $\hat{P}_j(f)(j=0,1,...,N)$ とする．

2.4.2 調波構造除去

2.4.1 節の処理で得られたスペクトル断片 $\hat{P}_j(f)$ の中には，認識対象楽器以外の調波構造をもつ楽器によるスペクトルの成分が含まれているものも存在する可能性があるため，そのようなスペクトル断片を除去する処理を行う．本研究では，調波構造を持つ楽器によるスペクトルは打楽器によるスペクトルと比べて急峻なピークをもつという特徴に着目して，認識対象楽器以外のスペクトル成分を含むスペクトル断片の除去を行う．具体的には，以下の方法で判定を行う．

(1) スペクトル断片 $\hat{P}_j(f)$ における極大値をすべて検出し，その中で最大値をとる周波数を検出する．

(2) 極大値をとる周波数の上下方向それぞれについて，極大値の 0.6 倍以下の値をとる周波数のうちもっとも極大値をとる周波数に近い周波数を検出する．

(3) 検出した上下 2 箇所の周波数の差が 30 フレーム以下である場合には，そのスペクトルは打楽器以外の成分によるものと判断し，そのスペクトル断片を除外する．

最終的に残ったスペクトル断片を $\hat{P}_k(f)(k=0,1,...,M)$ とする．

2.4.3 スペクトルモデル作成

式 (7) に示すように，最終的に得られたスペクトル断片 $\hat{P}_k(f)$ の周波数ごとに平均を求めることで認識対象楽器のスペクトルモデル $\hat{P}_{\mathrm{model}}(f)$ を作成する．ここで M はスペクトル断片の総数である．

$$\hat{P}_{\mathrm{model}}(f) = \frac{1}{M}\sum_{k=0}^{M}\hat{P}_k(f) \quad (7)$$

2.5 発音時刻検出

2.4.1 節で求めた $\hat{\theta}$ の半分の値を閾値として，$\hat{S}(t)$ に対して閾値以上の極大値をとる時刻を検出する．極大値の検出条件は，2.3 節 (2) と同様である．条件を満たした時刻における $\hat{P}(t,f)$ の値のみを残したスペクトログラムを $\hat{Q}(t,f)$ と定義する．

2.5.1 距離計算

各時刻における，$\hat{Q}(t,f)$ とスペクトルモデル $\hat{P}_{\mathrm{model}}(f)$ との距離 $D(t)$ を以下の式で定義する．

$$D(t) = \sum_{f=F-25}^{F+25}\left(\hat{Q}(t,f) - \hat{P}_{\mathrm{model}}(f)\right)^2 \quad (8)$$

2.5.2 閾値処理

閾値 Θ を以下のように定義する．

$$\Theta = \sum_{f=F-25}^{F+25}\left(\frac{\hat{P}_{\mathrm{model}}(f)}{2}\right)^2 \quad (9)$$

距離 $D(t)$ が Θ 以下である時刻を，認識対象楽器の発音時刻として検出する．

3. 評価実験

提案手法の有効性を評価するため，音楽信号を対象としたスネアドラムとバスドラムの発音時刻検出実験を行った．

3.1 実験条件

実験対象として，ポピュラー音楽データベース RWC-MDB-P-2001[7] に収録されている楽曲のうち 5 曲を用い，各曲の最初から 1 分を切り出して使用した．これらには，ドラム音だけでなく様々な楽器音やボーカルが含まれてい

る．すべての音楽信号は 16 bit，44,100 Hz でサンプリングされたステレオ信号である．実験には，モノラル信号に変換し，スネアドラムの場合は 1,260 Hz，バスドラムの場合は 420 Hz にダウンサンプリングしたものを使用した．

正解は，検出された発音時刻と実際の発音時刻とのずれが 2 フレーム (40 ms) 以下であることとした．実際の発音時刻を定めるために，各楽曲の標準 MIDI ファイルから認識対象楽器のみを抽出し，発音時刻を検出した．実験結果の評価は，再現率，適合率，F 値で行うものとし，それぞれ次式で算出する．

$$\text{再現率} = \frac{\text{正解した発音時刻数}}{\text{実際の発音時刻数}} \tag{10}$$

$$\text{適合率} = \frac{\text{正解した発音時刻数}}{\text{検出された発音時刻数}} \tag{11}$$

$$\text{F 値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}} \tag{12}$$

3.2 実験結果

提案手法によるスネアドラムを対象とした評価実験結果を表 1 に，バスドラムを対象とした評価実験結果を表 2 にそれぞれ示す．

提案手法における 5 曲の F 値の平均は，スネアドラムでは 0.95，バスドラムでは 0.93 という結果であった．従来手法 [4] では，同様の 5 曲に対して同じ条件で評価実験を行い，F 値の平均がスネアドラムでは 0.84，バスドラムでは 0.85 であったと報告されているので，この結果から提案手法は従来手法よりもスネアドラムとバスドラムの発音時刻の検出精度を向上させることができたといえる．

ただし，提案手法は，多くの場合で F 値が 0.95 前後の高い値を得ることが出来ているが，バスドラムの No.6 の結果を見ると，再現率が他と比べて低くなっている．これは，No.6 の楽曲中のバスドラム音が通常の音量と小さな音量の 2 種類を使い分けて演奏されており，小さな音量のバスドラム音を検出できなかったためである．バスドラムに限らず，音量の小さな成分は他の楽器の周波数成分に埋もれてしまうため検出するのは困難であり，この問題の解決は今後の課題である．

表 1 スネアドラム評価実験結果

曲番号	再現率	適合率	F 値
No.6	100.0%(63/63)	92.6%(63/68)	0.96
No.11	94.3%(33/35)	94.3%(33/35)	0.94
No.30	100.0%(70/70)	93.3%(70/75)	0.97
No.50	98.1%(102/104)	88.7%(102/115)	0.93
No.52	98.6%(72/73)	91.1%(72/79)	0.95

表 2 バスドラム評価実験結果

曲番号	再現率	適合率	F 値
No.6	64.9%(72/111)	94.7%(72/76)	0.77
No.11	100.0%(53/53)	100.0%(53/53)	1.00
No.30	100.0%(130/130)	100.0%(130/130)	1.00
No.50	95.6%(65/68)	92.9%(65/70)	0.94
No.52	96.1%(123/128)	91.8%(123/134)	0.94

4. おわりに

音楽信号を対象としたドラムスの自動採譜を実現するために，特定周波数帯域に着目して発音時刻検出を行う手法を提案した．現段階ではスネアドラムとバスドラムを対象とした処理を扱っており，パワーのピークのみを残したスペクトログラムの各周波数における時間方向の分散を利用することで，楽曲中で使用されているスネアドラムとバスドラムの特徴を表す周波数成分の検出を行った．そして，検出した周波数周辺の帯域に限定したスペクトログラムを利用して認識対象楽器のスペクトルモデルを作成し，各時刻においてモデルとスペクトルとの距離を計算して閾値処理を行うことでドラムスの発音時刻検出を行った．

提案手法の有効性を評価するため，音楽信号を対象としたスネアドラムとバスドラムの発音時刻検出実験を行い，従来手法 [4] と F 値の平均を比較したところ，スネアドラムでは 0.84 から 0.95，バスドラムでは 0.85 から 0.93 へと向上することが確かめられた．この結果から，特徴周波数の検出は正しく機能し，提案手法は従来手法よりも発音時刻の検出精度を向上させられるということが確認された．

今後の課題は，認識対象楽器の音圧が小さい場合や，他の楽器による周波数成分の重なりの影響を受ける場合などにも正しく特徴周波数を検出できるような処理を実現することである．また，ハイハットシンバルやタム類など，現在は認識対象としていない打楽器を扱うためのシステム拡張についても検討していく．

参考文献

[1] 亀岡弘和，嵯峨山茂樹，“多重音解析と自動採譜，”情報処理 50(8)，711-716，2009.
[2] 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，vol.J81-D-II， no.7，pp.1510-1517，1998.
[3] 北原鉄朗，後藤真孝，駒谷和範，尾形哲也，奥乃博，“多重奏を対象とした音源同定 : 混合音テンプレートを用いた音の重なりに頑健な特徴量への重み付け及び音楽的文脈の利用，”信学論，vol.J89-D，no.12，pp.2721-2733，2006.
[4] 吉井和佳，後藤真孝，奥乃博，“テンプレート適応を利用した実世界の音楽音響信号に対するドラムスの音源同定，”情処研報 [音楽情報科学]，2003(127)，55-60， 2003.
[5] 吉井和佳，後藤真孝，奥乃博，“単一テンプレート適応法による音楽音響信号を対象としたハイハットシンバルの音源同定，”情処研報 [音楽情報科学]，2004(84)，49-56，2004.
[6] Abraham Savitzky and M. J. E. Golay，“Smoothing and Differentiation of Data by Simplified Least Squares Procedures，”Analytical Chemistry 36 (8), 1627-1639, 1964.
[7] 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，”情報処理学会論文誌 45(3)，728-738，2004.