教師なしクラスタリングと認識誤りパターンを利用した打楽器音の音源同定

吉井 和佳† 北原 鉄朗‡ 櫻庭 洋平‡ 奥乃 博‡
†京都大学工学部情報学科 ‡京都大学情報学研究科知能情報学専攻

1. はじめに

打楽器を含む音楽演奏の自動採譜や, 楽曲検索用の自動タグ付けにおいては, 打楽器の発音機構が他の楽器とは大きく異なるため, 打楽器だけの演奏を対象とした音源同定技術が必要である. Herrera らは, 特徴量抽出・選択に基づき, ドラム単音を対象とした音源同定で 9 割弱の識別率を達成している[1]. しかし, 実際の音楽演奏では, 連続音や複数音の影響により, 性能は低下する. 後藤らは, 予め個々の打楽器音のパワー分布を登録することにより, MIDI 音源を対象とした打楽器演奏の音源分離・識別システムを構築している[2]. しかし, 実際の楽器演奏では, 楽器の個体差などにより, スペクトルが大きく変化するので, 事前学習だけでは対応できない.

本稿では, 楽音を含まない実打楽器演奏での音源同定として, 予め打楽器音の事前登録が不要な教師なしクラスタリングを膜鳴楽器識別に利用し, さらに, 体鳴楽器識別において, 認識誤りを生じやすい演奏パターンを知識として利用する手法を提案し, その評価をする.

2. 打楽器演奏に対する音源同定の問題点

2.1 想定する打楽器群とその演奏方法

本稿では，通常のドラムセットとドラム演奏を想定する. すなわち，**表 1** の 8 種類の打楽器で構成され，演奏法については以下の仮定をおくことができる：

(1) 体鳴楽器が同時に 2 種類発音されることはない. また, 膜鳴楽器も同様のものとする.
(2) 体鳴楽器と膜鳴楽器が同時に発音されてもよい.
(3) 発音間隔は膜鳴楽器で 125ms 以上, 体鳴楽器で 250ms 以上離れている (125ms は Tempo120 で 16 分音符に相当).
(4) 未知楽器はない.

2.2 本研究でのアプローチ

この問題を扱う上での問題とアプローチを述べる.

(a) 膜鳴楽器と体鳴楽器のスペクトルの重なりを抑制するために, 膜鳴楽器認識用に低域フィルタを, 体鳴楽器認識用に高域フィルタをかけて周波数帯域を制限する (**図 1**).

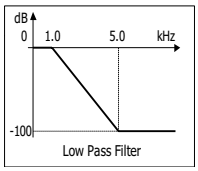
(b) 個体差の大きい膜名楽器は十分な事前学習ができないため, 同一曲内では音色変化が少ないことに着目し, 教師なしクラスタリングを行う.

(c) 体鳴楽器識別時に音の重なりで特徴が変動し, 誤りが避けにくいため, 認識誤りにパターンがあることに着目し, 4 種類の認識誤りパターンに対して認識誤り補正法を行う.

表 1 本稿で扱う打楽器群

膜鳴楽器	Bass Drum (BD), Snare Drum (SD), Low Tom (LT), Middle Tom (MT), High Tom (HT)
体鳴楽器	Crash Cymbal (CR), Hihat Close (HC), Hihat Open (HO)

Percussive Instruments Identification Using Unsupervised Clustering and Recognition Error Patterns
by Kazuyoshi Yoshii, Tetsuro Kitahara, Yohei Sakuraba, and Hiroshi G. Okuno (Kyoto Univ.)



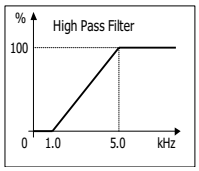


図 1 低域フィルタと高域フィルタ

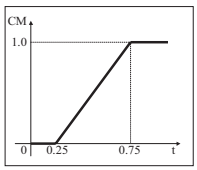


図 2 確信度の時間推移

3. 打楽器の音源同定手法

打楽器の認識は, 膜鳴楽器識別と体鳴楽器識別で別々に行う. 膜鳴楽器の認識では, 入力音響信号に図 1 左の低域フィルタを適用した後, (1) 発音時刻検出, (2) 教師なしクラスタリングを行う. 体鳴楽器の認識では, 入力音響信号に図 1 右の高域フィルタを適用した後, (1) 発音時刻検出, (2) k-NN 法による識別, (3) 認識誤りパターンを利用した補正を行う.

3.1 前処理としての発音時刻抽出

発音時刻検出には後藤らの手法[2]を用いる. 一定の周波数幅 f_c(= 80(cent), 130(Hz)) で周波数軸を区切り, 各区間内の最大パワーをその区間の代表値とし, パワー分布形状 $P_k(t, f)$ (k = Cent, Hz) とする. 各時刻のパワーの立ち上がり度を算出し, そのピーク時刻を発音時刻とする.

また, 各音符の発音時刻からを代表する代表パワー分布形状 $V_k(f)$ を以下の手順で計算する.

(1) 特徴量抽出区間の最大パワーフレームを求める.
(2) 最大パワーフレーム後 100ms の $P_k(t, f)$ の時間方向の平均値を $V_k(f)$ とする.

3.2 膜鳴楽器の教師なしクラスタリング

$V_{Cent}(f)$ に対し, k-means 法を利用して教師なしクラスタリングを行う. クラス数 (使用楽器数) は, 事前に与えられている. 一般に曲の中では BD と SD が多く叩かれることから, 以下の処理により音源同定ができる：

(1) 教師なしクラスタリングにより得られたクラスタのうち, 要素数が最大のクラスタと 2 番目のクラスタを選ぶ.
(2) 各クラスタの周波数重心の平均値を計算し, 小さい値のクラスタを BD, 大きい値のクラスタを SD とする.
(3) (1) で選ばれなかったクラスタも同様に周波数重心の平均値を計算し, 小さいものから順に LT, MT, HT とする.

3.3 体鳴楽器の識別と認識誤り補正

$V_{Hz}(f)$ を特徴量として k-NN 法 (k=10) により得られる識別結果を, 本章で述べるアルゴリズムで補正する.

表 2 35 個の特徴量の概要

(1)	**スペクトルの定常的特徴**（3 個） 周波数重心, 最大パワー周波数, 最大から NTH 番目までのパワーを持つ周波数などの時間方向の平均 (NTH = 3)
(2)	**N 次モーメントに関する特徴**（3 個） パワーの分散, 歪度, 尖度の時間方向の平均
(3)	**アタック区間に関する特徴**（8 個） 発音から最大パワーフレームまでの時間とその対数, パワーの時間方向の平均値, パワー包絡の面積とその割合, ゼロクロスの割合, 時間方向の重心とアタック時間に対する割合
(4)	**ディケイ区間に関する特徴**（6 個） ゼロクロスの割合, 周波数重心の時間方向の平均値と分散, パワーの分散, 歪度, 尖度などの時間方向の平均
(6)	**残響成分に関する特徴**（2 個） 最大パワーフレーム後 Y(ms) 後までの残響度合い (Y = 100, 200) (エンベロープの面積 / 最大パワー * Y ms)
(7)	**MFCC に関する特徴**（13 個） 特徴量抽出区間内の 13 次元 MFCC の時間平均

* アタックとは最大パワーフレームまでの区間, ディケイとはそれ以降の区間を指す.

3.3.1 認識誤りパターンと補正法の学習

次の 4 つの認識誤りパターンを利用する. [I], [II] は音の重なりに起因した誤り, [III], [IV] は音の特徴の類似性に起因した誤りである. 特に, 前者の誤りは, 単音で学習して演奏中の混合音を認識する上で避けられない認識誤りであり, この補正を行うことは極めて重要である.

[I] HC を SD の同時発音時に CR と誤認識
[II] CR の残響影響下の HC を CR と誤認識
[III] 残響系シンバル類同士の HO を CR と誤認識
[IV] パワー分布の似た HO を HC と誤認識

上記の各認識誤りパターンの補正を行う識別機械として, 決定木 T_i $(i = I, \cdots, IV)$ を以下の手順により構築する.

(1) 認識誤りパターンに該当する演奏データを作成する.
(2) 演奏データから特徴量を抽出し, 特徴量集合 S とする. 特徴量は従来研究[1),3)]を参考に定めた (**表 2**).
(3) クラス A をクラス B に誤認識する認識誤りパターンの補正法として, クラス B に属する特徴量と S を識別する決定木を決定木学習法 C5.0 により構成する.

3.3.2 信頼度の導入と補正法適用の判定

特徴量ベクトルの信頼度を次の 2 つの積として定義する.

- CM_1 : 抽出した特徴量ベクトル X と識別対象クラスの特徴量ベクトルの平均 M との類似度 $(= \frac{(X,M)}{\|X\|\|M\|})$.
- CM_2 : 信頼度 0.5 以上の CR 検出後, 特徴量の信頼性は時間経過で上昇 (**図 2**) すると仮定してモデル化した値.

信頼度 CM は, $CM = CM_1 * CM_2$で計算し, 各クラスごとに CM_{Class} (Class = CR, HO, HC) とする. どの時点でどの認識誤り補正法を適用するかは以下の手順に基づく.

if CR と識別 && $CM_{CR} < \theta_1$
 if SD が同時発音 **then** *apply*(決定木 T_I)
 elseif CR 検出後 0.5s 以内 **then** *apply*(決定木 T_{II})
 else *apply*(決定木 T_{III}) **fi**
elseif HC と識別 && $CM_{HC} < \theta_2 < CM_{HO}$
 then *apply*(決定木 T_{IV}) **fi**

表 3 学習データの概要

楽器名	楽器個体	強弱	総数
CR	10	4 種類	40 音
HO, HC	5	4 種類	20 音

表 4 音源同定結果

	膜鳴楽器識別	体鳴楽器識別	
		補正前	補正後
Roland	91.1% (51/56)	77.6%* (45/58)	89.7%* (52/58)
YAMAHA	89.3% (50/56)	52.6% (30/57)	77.2% (44/57)
市販 CD	100.0% (26/26)	65.5% (19/29)	69.0% (20/29)

*印の項目は closed 実験.

4. 実験と考察

体鳴楽器の学習用データを Roland 社 SC-88VL を用いて作成した (**表 3**). 膜鳴楽器の識別には学習データは必要ない. 評価用データは SC-88VL と YAHAMA 社 MU-2000 で典型的な 8 ビートを作成した. 市販のドラム演奏を収録した CD も対象とした. CD にはタム類, CR の発音はなかった. 補正の必要性判定で用いる閾値は$\theta_1 = 0.5$, $\theta_2 = 0.4$ とした.

発音検出は膜鳴楽器で 100%, 体鳴楽器で 9 割程度であった. 発音検出できたものに対する音源同定結果を**表 4** に示す.

- 教師なしクラスタリングにより, 学習データなしに 9 割程度の識別率を達成した. BD, SD が 16 分音符で連続する場所で, $V_{Cent}(f)$ がそれら 2 つの間の中間的なものとなり, 誤認識が生じやすい傾向が見られた. これは今後, 音符の遷移に着目した音楽的制約で解消可能だと考える.
- 体鳴楽器識別に関しては, 3 つの評価データに対し識別率が向上し, 本手法の有効性が示された. 市販 CD に対し向上が大きくないのは, SD と HC の同時発音による認識誤りの補正は有効に働いたが, 演奏データに HC が多いため, 補正するべき箇所が少なく, 間違って不要な HO への補正を行う場合が見られたからである. これは, MIDI 音源 1 種類での補正法の学習が十分でないためだと考えられる. 今後, 学習サンプル数を増やすことで対処できる. また, 音源や楽器個体が異なれば, どの誤りパターンが多く現れるかも異なる. ヒューリスティクスの改善により識別率向上が望める.

5. おわりに

本稿では, 打楽器演奏を対象とする音源同定において, 教師なしクラスタリングと認識誤り補正法導入の有効性を確認した. 今後は, サンプル数を増やして識別に有効な特徴量の検討, 音楽知識の導入も進めていく. 本研究は, 日本学術振興会科研費基盤研究 (B) 第 12480090 号の援助を受けた.

参考文献

1) Perfecto Herrera, etc al.: Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques, ICMAI, LNAI2445, pp.69-80, 2002.
2) 後藤, 村岡: 打楽器音を対象にした音源分離システム, 信学論, J77-D-II, 5, pp.901-911, 1994.
3) 北原, 後藤, 奥乃: 楽器音を対象とした音源同定：音高による音色変化を考慮する識別手法の検討", 情処学会 音情研報告, 2002-MUS-46-1, Vol.2002, No.63, pp.1-8, 2002