2003－MUS－51　(8)
2003／8／5

自己組織化マップによる教師なしクラスタリングを利用したドラム演奏の自動採譜

吉井 和佳[†]　北原 鉄朗[†]　櫻庭 洋平[†]　奥乃 博[†]

[†] 京都大学大学院 情報学研究科 知能情報学専攻

yoshii@kuis.kyoto-u.ac.jp　kitahara@kuis.kyoto-u.ac.jp
sakuraba@kuis.kyoto-u.ac.jp　okuno@i.kyoto-u.ac.jp

あらまし 本研究では，自動採譜技術確立のために，複数打楽器演奏を対象とする．打楽器演奏を扱う上での問題として，大きい個体差，データベース不足が挙げられる．本稿では，フィルタ処理を行った後，体鳴楽器の音源同定と，膜鳴楽器の音源同定を別々に行うことにし，特に，自己組織化マップによる教師なしクラスタリングを利用した膜鳴楽器の音源同定手法について報告する．まず，音響信号から発音時刻を検出し，各発音時刻に対する特徴量ベクトルを SOM への入力ベクトル群としてマップを作成し，得られたマップをヒストグラムを用いて分割することでクラスタリングを行う．MIDI 音源及び市販 CD に対する評価実験の結果，平均 90%の膜鳴楽器の音源同定を達成し，提案手法の有効性を確かめることができた．

Automatic Transcription of Drum Performance using Unsupervised Clustering by Self-Organizing Map

Kazuyoshi Yoshii[†]　Tetsuro Kitahara[†]　Yohei Sakuraba[†]　Hiroshi G. Okuno[†]

[†] Dept. of Intelligence Science and Technology,
Graduate School of Infomatics, Kyoto University

Abstract The problems with automatic transcription of plural-duram performance are a large variation of individual percussive instrument and a lack of large corpus of such sounds. Therefore, the percussive instrument identification is divided into two subprocessing, idiophone and membraphone identifications by separating input sounds by spectral filters. This paper reports the details of membraphone identification based on unsupervised clustering with self-orgnizing map (SOM). Features including beats are extracted and used as input vectors of SOM, and feature map is separated by using histogram. The performance of identification is evaluated with MIDI and commercial CDs and about 90% of membraphone identification is attained.

1. はじめに

AV 機器の発達やデジタル信号処理技術の発展，計算機性能の向上により，楽器音や音楽演奏を対象とした研究は 1970 年以降広く行われるようになった．近年，音楽情報処理という研究領域が確立し，活発な研究が行われ，様々な分野で，その重要性が認識されてきている．例えば，計算機による自動採譜技術が実現できれば，楽譜書き起こしの際に人手にかかる負担を大きく減らすことができる．また，インターネットの急激な発達により，WEB 上には無数のマルチメディアデータが氾濫している．このようなマルチメディアデータを扱う上では音楽情報処理技術は不可欠である．なぜなら，蓄積された音楽データに対して，楽器音の自動音源同定により MPEG-7 などのタグ付けを行うことができれば，検索が容易になり，ディジタルアーカイブの利便性を高めることができるからである．

こうした目標の実現のためには，計算機への入力として，楽音（調波構造を持ち，音高が明確な音），非楽音を共に含んでいる音響信号を扱う必要がある．非楽音の代表的なものには打楽器音がある．打楽器音を対象とした音楽情報処理研究は，楽音を対象としたものよりもずっと少なく，余り取り組まれてこなかった．しかし，楽曲のリズムやビートなどの情報は，打楽器

パートに非常に密接に関係しており，打楽器を除いての音楽理解をいうものはあり得ない．そこで，第一段階として，打楽器の発音機構は他の楽器とは大きく異なるので，打楽器音のみで構成された音響信号を入力として音源同定を行う手法を開発する必要がある．我々は，打楽器演奏を入力音響信号とした自動採譜を行うという観点に立って，打楽器音に対する音源同定処理を提案する．打楽器演奏を扱う上での問題は主に 3 つあり，(1) 打楽器音は楽音に比べ，非常に個体差が大きい問題　(2) 学習データ不足の問題　(3) スペクトルの重なりの問題 である．

(1)，(2) の問題は，打楽器を扱う難しさに対応する．打楽器の場合，胴のサイズや膜の材料などが異なると，スペクトル形状が大きく変化する．また，打楽器音の共通データベースが少数であり，十分な学習データを用意することも困難である．このことから，従来通りの教師あり学習を採用することは難しい．(3) の問題は，演奏を扱う難しさに対応する．演奏とはすなわち混合音であり，各楽器のスペクトルが混合し，干渉しあう．その影響を最小限にする手法が必要である．

打楽器音の音源同定に関する先行研究としては，Herrera らの研究がある[1)2)3)]．これは，標準的なドラムセットを構成する 9 種類の打楽器☆ を対象として，単独発音に対する音源同定を行う上での，特徴量の次元圧縮法や識別手法について比較検討している．採用されている識別手法は教師あり学習を必要とし，(1)，(2) の問題への対処を考える上では適当ではない．また，打楽器演奏を扱ったものとしては，後藤らの研究がある[4)]．これは，テンプレートマッチングを改良した手法により，ドラム演奏を入力として，高精度の音源同定を実現している．この実験では学習データと評価データは同じ音源で作成されており，音源が異なった場合，テンプレートをそのつど用意することができないところに問題がある．

本研究では，ドラム演奏の自動採譜を実現する上で，今回は特に，膜鳴楽器の音源同定手法について検討する．同一曲内では，各打楽器の特徴量は定常的であると仮定し，抽出した特徴量に対し，教師なしクラスタリングを利用した．教師なし識別法であるために，学習データは必要ない．また，個体差に富む打楽器でも，同一曲内では個体差を考慮しなくてよい．よって，(1)，(2) の問題に対処できる．また，我々はクラス数 （＝楽器の種類数）を既知とした場合のクラスタリング手法が有効に働くことを報告した[5)] が，実際にはクラス数が既知である場合は少ない．そこで，教師なしクラスタリングの実現のために，Kohonen の提唱した自己組織化マップ (SOM : Self-Organizing Map)[6)] を用いた．クラスタリング手法には，寺島らの手法[7)] を採用した．また，すべての音響信号には事前にローパスフィルタ処理を行うことで，周波数帯域を分離し，(3) の問題に対処する．

以下，**2．**及び **3．**で自己組織化マップを用いた教師なしクラスタリングについて説明する．**4．**で本研究のために作成したシステム構成について説明する．**5．**で打楽器演奏を対象とした音源同定実験を行い，**6．**で結論と今後の課題を述べる．

2. 自己組織化マップ

2.1 SOM の採用

近年最もよく用いられるクラスタリング手法の 1 つであり，以前我々も使用した[5)]*k*-means 法では，次のような問題があった[9)]．

(1) クラスタ数をあらかじめ与える必要がある．与えた k が適切でない場合には，データがほとんど存在しない部分に中心が位置するなどの問題が起こることがある．

(2) データの存在しない領域が存在し，データを明らかに分けることができないような場合も，*k*-means 法ではそれを我々に明示できない．

(3) データの存在領域が明らかに分離しているにもかかわらず，それぞれの領域でのデータ数が大きくクラスタがある場合，我々が通常求めたいと思うクラスタと，得られるクラスタがかなり異なることがある．

通常，クラスタ数（＝ 楽器の種類数）が事前に分かることは少ない．クラスタ数を陽に決定するのが困難なことから，従来は，個数決定問題を上位問題として，*k*-means 法を繰り返すなどして対処されてきた．しかし，不適当な k を与えると，(1)，(2) の問題が生じやすい．また，ドラム演奏においては，各楽器の発音数に偏りがあることが予想され，(3) の問題も生じやすいと考えられる．SOM を用いることでこれらの問題は解決する．

2.2 SOM の構造

本研究での SOM は，1 次元上に配列した N 個の素子群 (以下，マップと呼ぶ) を用いる．SOM は任意の次元の入力ベクトル空間 $X \in R^p$ を離散空間 Ω (マップ) に写像，

$$\Phi : X \rightarrow \Omega$$

するものである．Ω は，通常 1 次元か 2 次元であるが，本研究では，視覚的な表現とクラスタリングを容易にするために 1 次元としている．Ω 上には等間隔に素子

☆ Crash Cymbal，Ride Cymbal，HiHat Open，HiHat Close，Bass Drum，Snare Drum，High Tom，Middle Tom，Low Tom

が配置され，各素子 i $(i=1,\cdots,n)$ は，入力ベクトル空間と同じ次元数の重みベクトル m_i を備える．

入力ベクトル x が素子 $\omega \in \Omega$ に写像，

$$\Phi : x \to \omega$$

されるとき，m_ω は x と類似したベクトルを持っている．そのため，素子の並びは入力ベクトル空間 $X \in R^p$ でのデータの近隣関係をよく保存している．

SOM は，3 次元を超える多次元ベクトルを分かりやすく表示するために，そのトポロジカルな関係を保存しながら，1 次元上に写像できるところに利点がある．いわば，非線形の主成分分析と言うべきものである．このとき，副次的な機能として，1 次元上にデータ密度ヒストグラムを作成することでクラスタリングが可能になる（後述）．

2.3 SOM の学習アルゴリズム

SOM の学習アルゴリズムは，時刻 t における入力ベクトル $x(t)$ と各素子の重みベクトル $m_i(t)$ との類似性から更新すべき重みを決定する**類似性マッチング**フェーズと，その重みベクトル m_i を x に近づけるように学習する**更新**フェーズから構成される．この過程を以下に数式で表す．

類似性マッチング

$$|x(t)-m_C(t)| = \min_i |x(t)-m_i(t)|$$

ただし，$|x-m|$ は x と m のユークリッド距離とする．このフェーズでは，入力ベクトルに最も近い重みベクトルを持つ素子 C を選択する．この素子 C のことを勝利素子と呼ぶ．

更新

if $i \in N_C(t)$

$$m_i(t+1) = m_i(t) + \eta(t)\{x(t) - m_i(t)\}$$

else

$$m_i(t+1) = m_i(t)$$

ここで，$N_C(t)$ は勝利素子 C の近隣素子を要素とする部分集合である．$N_C(t)$ は時刻 t の経過とともにその範囲が小さくなっていくようにする．また，$\eta(t)$ は学習係数であり，$0 \le \eta(t) \le 1$ の範囲をとる．単調減少関数に設定し，学習が収束するようにする．このフェーズでは，勝利素子 C の近隣 $N_C(t)$ の範囲内にある素子の重みベクトルを更新する．

SOM の学習の様子を図 1 に示す．まず，重みの初期値 $m_i(0)$ をランダムに定める．このとき，入力ベクトルの要素の取りうる値の範囲を考慮して設定するのがよい．入力ベクトル x の集合から x をランダムに 1 つ選んで逐次 SOM に入力し，上記のような 2 フェーズ

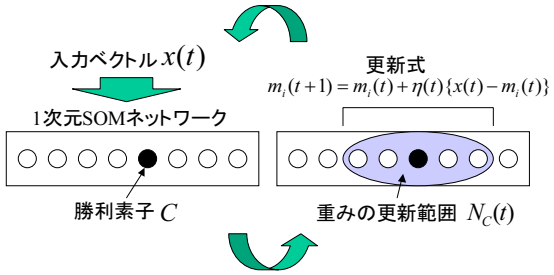

図 1 SOM の学習

の学習を繰り返すことで，素子の更新範囲，学習係数ともに収束し，SOM も収束に向かう．学習時には，勝利素子 C だけではなく，近隣の素子まで重みを更新する．よって，十分な学習が終了すると，入力ベクトル空間上で近いベクトルは，Ω 上で近い素子に対応することになる．

3. 教師なしクラスタリングを用いた音源同定

本研究では，SOM による教師なしクラスタリングを音源同定に応用する．これには，寺島らの手法[7] を参考にした．具体的な手続きは，次の 3 段階から構成される．

(1) SOM によるマップ作成
(2) マップ解析と教師なしクラスタリング
(3) 各クラスへのラベル付け

3.1 入力音響信号の条件

本研究では，入力音響信号は, 標準的な構成のドラムセットによる Tempo120 より遅い 8 ビートのドラム演奏であるとする．ドラムセットは，表 1 の 9 種類の打楽器で構成されている. ここで, 膜鳴楽器と体鳴楽器という分類は, Erich von Hornbostel と Curt Sachs の体系的楽器分類に基づく[8]. 前者は 1kHz 以下の中低域の周波数成分が多く, 後者は 4kHz 以上の高周波数域に広くスペクトルが分布している．

使用楽器および演奏法について, 以下の仮定をおく：

(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．

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), Ride Cymbal (RI) Hihat Close (HC), Hihat Open (HO)

3.2 マップ作成

まず，SOM を用いてマップ作成を行う．ドラム演奏を含んだ入力音響信号から，各発音時刻に対して特徴ベクトルを抽出する（後述）．これを，入力ベクトル x の集合とする．1 次元上に等間隔に配置した素子群に，入力ベクトルを逐次入力し，**2.** で説明したアルゴリズムで学習を行うことで，マップを作成する．

3.3 教師なしクラスタリングのアルゴリズム

3.3.1 評価値の算出

得られたマップから，次の手順で評価値 V_i, L_i を計算し，各評価値についてのヒストグラムを作成する．

V_i の算出　入力ベクトルのうち，素子 i に対応するベクトルの個数を求める．すわなち，素子 i が勝利素子となるような入力ベクトルの個数である．

S_i の算出　左右の隣接素子間との重みベクトルの類似度を求める．これは，

$$S_i = |m_i - m_{i-1}| + |m_i - m_{i+1}|$$

で計算できる．ただし，両端の素子に関しては，片側の素子間との重みベクトルの類似度を計算し，S_i とする．

L_i の算出　上記 2 つの値を用いて，

$$L_i = V_i / S_i$$

とする．

このようにして求めた V_i, S_i は，クラスタの集積度を表す有効な指標である．入力ベクトル空間内で密集しているベクトル群は，マップ上でも距離の近い素子群に写像される．よって，V_i のヒストグラムを描いたとき，山になった部分がクラスタの集積度が高いことを示している．また，隣接素子間の距離が近ければ近いほど，その素子に対応する入力ベクトルの類似性が強い．よって，S_i のヒストグラムを描いたとき，谷になった部分がクラスタの集積度が高いことを示す．

これら 2 つの性質から，L_i を新たに定義することで，ヒストグラム上で，さらにはっきりした山と谷を得ることができると考えられる．本研究では，特に V_i, L_i をクラスタリングに利用する．得られたヒストグラムの例を図 2 に示す．

3.3.2 ヒストグラムの分割

ヒストグラムの分割数は，クラス数であるので，最終的なラベル付けに大きく影響する．楽器が短い間隔で連続して発音した場合などは，抽出した特徴量が変形し，得られるヒストグラムも変形するので，常に正しいクラス数でヒストグラムの分割ができるとは限らない．よって，ヒストグラムの分割数に曖昧性を持たせることを考える．

具体的には，評価値 V_i, L_i の 2 種類のヒストグラムを用いて分割を行うことで，2 通りのクラスタリング

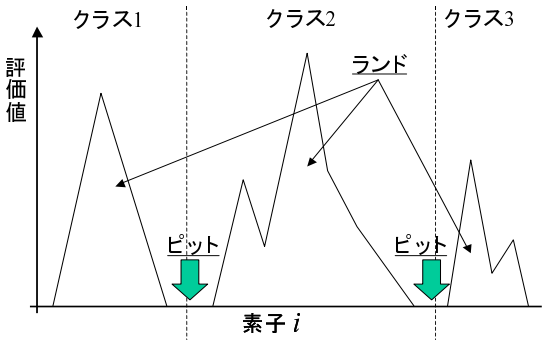

図 2　ヒストグラムと分割

の解釈を与える．以下のようなアルゴリズムを用いる．

(1) ヒストグラム内で，評価値が 0 の部分を**ピット**，それ以外の盛り上がっている部分を**ランド**と呼ぶことにする（図 2）．

(2) 各ランドごとにヒストグラムを分割する．すなわち，ランド数はクラス数に等しい．

(3) 各ランド内の素子に対応する入力ベクトルをまとめて 1 つのクラスとする．

予備実験から，V_i のヒストグラムを分割した方がクラス数が多くなる傾向にあることが分かっている．

3.4 ラベル付けのアルゴリズム

教師なしクラスタリング後，各クラスへのラベル付け処理には, 以下の 2 つの事前知識を用いる．

- 一般的なドラム演奏においては, BD と SD がタム類よりも多く発音される．
- タム類は LT, MT, HT の順に音が高くなるように知覚される．これは, 周波数重心が順に高くなることに対応すると考える．

これらの知識を用いることで, 以下の処理により楽器名同定を行う:

(1) 教師なしクラスタリングにより得られたクラスのうち, 要素数が最大のクラスと 2 番目のクラスを選ぶ．

(2) 各クラス内要素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) で選ばれた以外のクラスに対しても同様に周波数重心の平均値を計算する．このクラス数が 3 の場合は，周波数重心が小さいものから順に LT, MT, HT とラベル付けする．クラス数が 2 以下の場合には，ラベル付けの方法にいくつか候補があるが，いずれも最終的な出力とする．

打楽器音を扱う場合，十分な学習データが用意できないなどの問題のため，得られた各クラスの要素と，事前学習したものとを比較して，ラベル付けを行うのは難しい．そこで，周波数重心による順序付けによっ

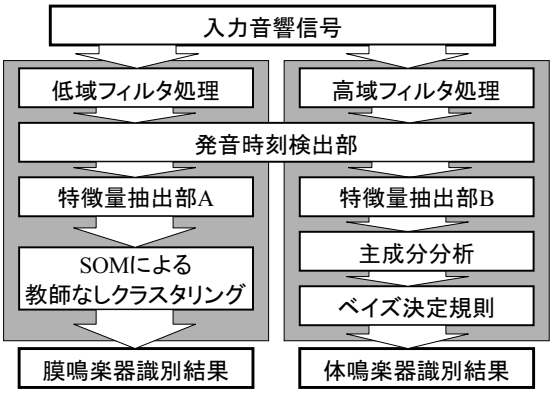

図 3　システム構成

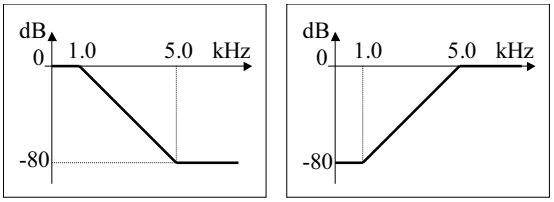

図 4　低域通過フィルタと高域通過フィルタ

て，ラベル付けの候補を絞るだけにとどめる．また，打楽器は多様性に富み，将来的にドラム以外の打楽器を扱う必要が出てきた場合，事前に学習データは持っていることは現実的ではない．そのため，事前学習したものとの比較という操作はできる限りなくす方が望ましいと考える．

3.5　出力結果の曖昧性

本章で説明した手法には，次の 2 つの曖昧性がある．

- 評価値 V_i, L_i によるヒストグラムの違いから生まれるクラス数の曖昧性
- ラベル付けの曖昧性

この結果，音源同定結果には何通りかの仮説が生まれる．この曖昧性は，今後，リズムパターン情報などとの統合で解消できる問題であると考える．

4. システム構成

打楽器の音源同定処理は，膜鳴楽器識別と体鳴楽器識別で別々に行う．本研究のために構築したシステムを図 3 に示す．膜鳴楽器の認識では，入力音響信号に図 4 左の低域通過フィルタを適用した後，(1) 発音時刻検出，(2) 特徴量抽出，(3) SOM を利用した教師なしクラスタリング の順に処理を行い，いくつかの候補を出力する．体鳴楽器の認識では，入力音響信号に図 4 右の高域通過フィルタを適用した後，(1) 発音時刻検出，(2) 特徴量抽出，(3) 主成分分析とベイズ決定規則による識別 の順に処理を行い，体鳴楽器名を出力する．

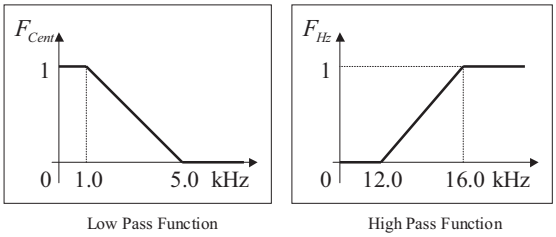

図 5　重み関数 F_{Cent} と F_{Hz}

4.1　発音時刻検出部

発音時刻検出部では，各時刻におけるパワーの立ち上がり度を求め，この値の時間方向の 1 次微分値が大きい値をとる時刻を発音時刻として出力する．本研究では，後藤らの手法を採用した[4]．

ここで，パワー分布形状 $P_k(t,f)$（k = Cent, Hz）を以下のように定義する．まず，周波数軸のスケールは，2 種の対象楽器類で異なったものを使用する．膜鳴楽器の発音時刻検出には，中低域の分布をよく表すように周波数軸を対数尺度の Cent で表し，体鳴楽器の場合には高域の分布をよく表すように周波数軸を Hz で表す．次に，周波数軸を一定の周波数幅 f_c で区切り，各区間内の最大パワーをその区間の代表値として $P_k(t,f)$ に定める．（f_c = 80[Cent], 130[Hz] とした）こうして得られたパワー分布形状 $P_k(t,f)$ から，立ち上がりの度合い $Q_k(t,f)$ 算出法と，発音時刻検出の手順を以下に示す．

(1)　$t = l-1,\ l,\ l+1$ の各時刻において連続して

$$\frac{\partial P_k(t,\ f)}{\partial t} > 0$$

を満たすとき，$t = l$ における $\partial P_k(t,f)/\partial t$ を $Q_k(t,f)$ とする．満たさないときは $Q_k(t,f) = 0$ とする．

(2)　各時刻 t ごとに $Q_k(t,f)$ の重み付き合計値 $S_k(t)$ を次式により定める．

$$S_k(t) = \sum_f F_k(f) Q_k(t,f)$$

ただし，$F_k(f)$ は図 5 のような楽器の特性に応じた重み関数である．

(3)　$S_k(t)$ に対し，Savitzky と Golay の方法による平滑化と微分を用い[10]，極大値を与える時刻を検出する．$S_{Cent}(t)$ から求めた時刻を膜鳴楽器の発音時刻とし，$S_{Hz}(t)$ から求めた時刻を体鳴楽器の発音時刻とする．

4.2　特徴量抽出部

4.2.1　膜鳴楽器の場合

膜鳴楽器識別には，スペクトル上での周波数軸方向のパワー分布の違いが利用できる．そこで，パワー分布の形状を特徴量ベクトル化し，SOM への入力とす

表 2 体鳴楽器の音色を表す 43 個の特徴量

(1)	**スペクトルの定常的特徴** (3 個 : FT1 - FT3) 周波数重心，最大パワー周波数，最大から 5 番目までのパワーを持つ周波数の時間方向の平均
(2)	**2-4 次モーメントに関する特徴** (3 個 : FT4 - FT6) パワーの分散，歪度，尖度の時間方向の平均
(3)	**アタック (atk) 区間に関する特徴** (8 個 : FT7 - FT14) atk エネルギー，atk 時間，ログ atk 時間，ゼロクロス割合，atk エネルギー/atk 時間，atk エネルギー/(最大パワー*atk 時間)，atk 時間重心，atk 時間重心/atk 時間
(4)	**ディケイ (dcy) 区間に関する特徴** (6 個 : FT15 - FT20) 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, 尖度の時間方向の平均
(5)	**パワー分布に関する特徴** (7 個 : FT21-FT27) 5k-7k, 7k-10k, 10-13k, 13k-16k, 16k-20k [Hz] 各セクションのパワーが占める割合，5%, 20%以上のパワーの周波数の占める割合
(6)	**残響成分に関する特徴** (2 個 : FT28 - FT29) 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 : FT30 - FT43) 13 次元+E の MFCC の分散の時間平均

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.

ることを考える．これを，代表パワー分布ベクトルと呼ぶことにする．代表パワー分布ベクトルは次のように定義し，検出された各発音時刻に対して計算することで，SOM への入力ベクトル群を求める．

(1) 発音時刻以降, 60ms 以内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 60ms 間の $P_{Cent}(t,f)$ の時間方向の平均値を $V_{Cent}(f)$ とする．

4.2.2 体鳴楽器の場合

体鳴楽器識別には，従来どおりの教師付き統計的識別法を採用する．そのときに利用する特徴量を表 2 に示す．体鳴楽器識別においては，体鳴楽器のスペクトルが広範囲の周波数帯域に広く分布しているので，代表パワー分布ベクトルを用いても十分な識別能力が得られない．また，残響の影響や，膜鳴楽器のスペクトルの重なりなどで，抽出する特徴量ベクトルが変形する問題がある．この問題への対処法は現在検討中である．

4.3 音源同定部

4.3.1 膜鳴楽器の同定 : 教師なしクラスタリング

特徴量抽出部の処理により，特徴量ベクトル群が得られる．それらのベクトルをクラスタリングし，各クラスにラベル付けを行って最終的な出力とする．SOM の学習アルゴリズムと教師なしクラスタリングの方法については **2.** 及び **3.** で説明をした．この処理により，複数のラベル付け候補が出力として得られる．ここで，学習係数 $\eta(t)$，重みの更新範囲 $N_C(t)$ を，

$$\eta(t) = 0.5\left(1 - \frac{t}{T}\right)$$

$$N_C(t) = \left\lfloor 10\exp\left(-\frac{t^2}{1.5\times 10^5} + 1\right)\right\rfloor$$

のように設定し，SOM の学習を行った．ただし，T は学習の反復回数であり，$T = 3000$ とした．

4.3.2 体鳴楽器の同定 : ベイズ決定規則による識別

体鳴楽器の同定には，ベイズ決定規則による識別を用いる．表 2 の特徴量を抽出し，主成分分析により，11 次元に次元圧縮を行い，パターンベクトルとする．

各クラスのパターンが多次元正規分布に従うと仮定する．このとき，入力ベクトル x がクラス c である確率密度関数 $p(x|c)$ は次式で与えられる．

$$p(x|c) = \frac{1}{(2\pi)^{d/2}|\Sigma_c|^{1/2}}\exp\left(-\frac{1}{2}D^2(x,\mu_c)\right)$$

ここで，d はベクトルの次元数，D はマハラノビス距離である．また，Σ_c, μ_c はそれぞれ，クラス c の共分散行列，平均ベクトルである．

このとき，ベイズ決定規則に基づいた識別関数 g は以下のように導出できる．

$$\begin{aligned} g_c(x) &= \log p(x|c) + \log p(c) \\ &\equiv -D^2(x,\mu_c) - \log|\Sigma_c| \end{aligned}$$

ここで，$p(c)$ はクラス c の起こる事前確率であり，すべてのクラスで等しいと仮定している．このとき，

$$cls = \mathrm{argmax}_c\ g_c(x)$$

となるクラス cls を音源同定結果とする．

5. 評 価 実 験

5.1 実 験 方 法

膜鳴楽器の識別には学習データは必要ない．体鳴楽器の学習データとして，市販のサンプリング CD からデータを収集した (サンプリング周波数 44.1kHz，モノラル) ．電子音も含まれているが，実楽器音と大きくかけ離れたものは除外した．内訳を表 3 に示す．

評価用データは，YAMAHA 社製 MIDI 音源 MU-2000 を用いて作成した．ポップスやロックで使われる，8 ビートの標準的なドラムパターンである．2 種類の音色セット（S1, S2）に対し，それぞれ Tempo120 で 4 パターン（P1〜P4），合計 8 データを作成した．表 4 に各パターンに含まれる楽器数の内訳を示す．また，打楽器による実演奏（RP）に対しても評価実験を行った．

周波数解析には，発音時刻検出部で窓幅 2048 点，窓シフト長 20ms，特徴量抽出部で窓幅 4096 点，窓シフ

表 3 体鳴楽器識別に用いる学習データの内訳

CR	RI	HC	HO	合計
28	32	121	48	229

表 4 各パターンの楽器数の内訳

	P1	P2	P3	P4	RP
BD	13	19	17	16	16
SD	9	9	11	11	12
LT	1	1	0	3	0
MT	2	0	2	6	0
HT	1	1	4	3	0
CR	2	2	5	5	0
RI	0	0	0	7	0
HC	21	25	16	9	22
HO	5	3	5	3	5

表 5 発音時刻検出結果

	膜鳴楽器		
	S1	S2	RP
再現率	129/129	129/129	28/28
適合率	129/130	129/139	28/28

	体鳴楽器		
	S1	S2	RP
再現率	93/108	94/108	24/27
適合率	92/111	79/116	24/24

ト長 10ms とし，FFTW を利用した[11]．各発音に対する特徴量抽出区間長は，膜鳴楽器の場合 60ms，体鳴楽器の場合 200ms とした．

また，以下の基準を用いて，得られた結果を評価する．

(1) 検出された発音時刻と正解とのずれが，その演奏の Tempo で 32 分音符の長さ以内である場合，正しく発音時刻が検出できたとする．

(2) 膜鳴楽器識別の場合，ラベル付けの仮説が複数得られる．その中で，最も高いラベル正解率をとる仮説を評価に用いる．また，SOM の重みベクトルに与える初期値によって結果に変動があるので，識別精度と出現頻度を総合して評価する必要がある．具体的には，20 回実験を繰り返し，音源同定率の最高，最低，平均を求める．

(3) 体鳴楽器識別の場合，得られた音源同定結果が正しいものであるかを評価する．

5.2 実験結果

発音時刻検出結果を表 5 に示す．膜鳴楽器識別結果と体鳴楽器識別結果は，正しく発音が検出されたもののうち，音源同定が正解したものがいくつあるかで評価している．それぞれ表 6，表 7 に示す．

5.3 考　　察

5.3.1 膜鳴楽器の識別に関する考察

まず，膜鳴楽器の発音時刻検出であるが，表 5 を見れば，非常に高精度に実現できていることが分かる．これは，低域通過フィルタ処理が有効に働き，体鳴楽器のスペクトルが混合する影響を十分小さくできたからであると考えられる．

次に，音源同定結果について考察する．表 6 を見ると，SOM の重みベクトルに与える初期値により音源同定精度が変動することが分かる．このとき，音源同定率の最高は 95%前後，最低は 75%前後であり，平均的には 90%前後の音源同定率が得られている．Herrera らは，統計的手法を用いてドラム単音の音源同定実験を行ったところ，90.7%であったと報告している[1]．混合音であるドラム演奏に対してこのような音源同定手法をそのまま適用すると，識別率低下は避けられないが，本研究の提案手法では，教師なしにもかかわらず，同程度の音源同定率を達成することができた．これは，提案手法の有効性を示すと同時に，同一楽曲内では，各楽器の特徴量は定常的であるという仮定を裏付けるものである．

表 6 膜鳴楽器識別結果

	最高	最低	平均
S1P1	96.2%	76.9%	89.6%
S1P2	96.7%	80.9%	93.2%
S1P3	100.0%	76.5%	84.7%
S1P4	94.9%	76.9%	85.4%
S2P1	96.2%	76.9%	91.2%
S2P2	76.7%	73.3%	86.5%
S2P3	100.0%	82.4%	86.0%
S2P4	97.4%	76.9%	86.7%
RP	100.0%	78.6%	97.9%

表 7 体鳴楽器識別結果

	音源同定率
S1	58.1% (60/93)
S2	45.1% (42/93)
RP	41.7% (10/24)

誤認識が生じた箇所について考察する．BD と SD に関する誤りは比較的生じにくく，タム類に関する誤りは多くなる傾向にある．BD と SD で誤りが生じやすい箇所は，BD と SD が 16 分音符で連続して発音されるような場所である．このような場所では，抽出した特徴量がそれら 2 つの間の中間的なものとなり，誤認識を引き起こしやすい．

また，タム類はフィルインと呼ばれる演奏箇所で密集して多発する傾向があり，そのような場所から抽出した特徴量は崩れ，誤認識につながりやすい．その上，楽器個体差の影響もあり，ドラム演奏において，タム類の識別に教師付き識別法を採用することは難しい．しかし，教師なしクラスタリングでは，近い特徴量ベクトル同士をまとめあげていくため，教師ベクトルと抽出した特徴量ベクトルとの距離に影響されずに識別可能である．SOM に初期値依存性があるとはいえ，100%の音源同定率を得られる可能性があることは着目すべき点である．これは，SOM が多少の特徴量変動をうまく吸収できることを示している．

最後に，本手法の問題点を挙げる．

表 8 体鳴楽器単音の音源同定実験結果

	CR	RI	HC	HO	識別率
CR	23	0	1	4	82.1%
RI	0	22	3	7	68.8%
HC	0	0	104	17	86.0%
HO	2	9	2	35	72.9%

計 80.3% (184/229)

(1) 誤検出した発音時刻を除去する機構がないため，そのような場所から抽出した特徴量ベクトルまで SOM への入力としてしまう．

(2) SOM の初期値依存性により，音源同定率が 70% 近くまで低下してしまうことがある．

(3) クラスタリング後のラベル付けに曖昧性を残しており，その解消法が必要である．

これらの課題を解決していくことで，さらに高精度で高度に自動化された採譜処理が実現できると考える．

5.3.2 膜鳴楽器の識別に関する考察

表 5 をみれば分かるように，体鳴楽器の発音時刻検出率は膜鳴楽器の場合に比べて明らかに低い．体鳴楽器は，広い周波数帯域に渡ってなだらかにスペクトルが分布しており，スペクトル上に鋭いピークを持たない．そのため，発音時刻検出は難しくなる．

次に，音源同定結果について考察する．表 7 を見ると，音源同定率は 50%前後と非常に低い．参考として，表 3 にあるデータから表 2 の特徴量を抽出し，主成分分析で 11 次元に圧縮後，Leave One Out 法で音源同定実験を行った結果を表 8 に示す．表 7 と表 8 を比較しても明らかなように，従来どおりの統計的識別法を用いただけでは，十分な精度が得られないことが確認できた．

誤認識が生じやすい場所は，体鳴楽器が SD やタム類と同時発音するような箇所であった．特に SD は，太鼓の裏に金属のひもが張られており，その金属振動により，高周波までスペクトルが分布する特徴がある．そのため，高域通過フィルタだけでは，SD の影響を十分に減少させることができない．また，CR の残響の影響でも誤認識は増加する傾向が見られ，これらの問題への対処法は今後の課題とする．

6. おわりに

本稿では，ドラム演奏の自動採譜を実現する上で，膜鳴楽器の識別に自己組織化マップによる教師なしクラスタリングを利用する手法を提案した．この手法により，楽器個体差の問題，打楽器音データベースの不足の問題に対処できる．また，膜鳴楽器と体鳴楽器とのスペクトルの重なりの影響を最小限にするため，帯域通過フィルタ処理も導入し，それぞれ別々に音源同定を行うことにした．本手法を実装，実験した結果，膜鳴楽器の音源同定率は平均的に 90%前後を達成できた．教師付き統計的識別法を用いた場合の体鳴楽器の音源同定率は 50%程度にとどまった．

膜鳴楽器識別に関しては，**5.3.1** の最後で述べたような課題が残っている．また，体鳴楽器識別に関しては，残響などが原因で、特徴量が崩れた場合の対処法が必要になる．今後はこれらの課題に対して 1 つづつ対処していくことで，打楽器演奏の自動採譜の精度を向上させていく予定である．

謝辞 本研究は，日本学術振興会科学研究費補助金基盤研究 (A) 第 15200015 号，およびサウンド技術振興財団研究助成による．また，有益なご助言をくださった片寄 晴弘氏（関西学院大学），柏野 邦夫氏（NTT コミュニケーション科学基礎研究所），中臺 一博氏（株式会社ホンダ・リサーチ・インスティチュート・ジャパン）に感謝する．

参 考 文 献

1) Perfecto Herrera, Alexandre Yeterian, Fabien Gouyon：“Automatic Classification of Drum Sounds: A Comparison of Feature Selection Methods and Classification Techniques”，ICMAI, LNAI2445, pp.69–80, 2002.
2) Perfecto Herrera, Amaury Dehamel, Fabien Gouyon：“Automatic Labeling of Unpitched Percussion Sounds”, *Proc.* Audio Engineering Society 114th Convention, pp1–14, 2003.
3) Fabien Gouyon, Perfecto Herrera：“Exploration of techniques for automatic labeling of audio drum tracks’ instruments”，*Proc.* MOSART Workshop on Current Research Directions in Computer Music, 2001.
4) 後藤 真孝，村岡 洋一：“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.
5) 吉井 和佳，北原鉄朗，櫻庭洋平，奥乃博：“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”，情処第 65 回全国大会, 2, pp187–188, 2003.
6) T.Kohohen：“Self-Organizing map”，*Proc.* IEEE, **78**, 9, pp.1464–1480, 1990.
7) 寺島 幹彦，白谷 文行，山本 公明：“自己組織化特徴マップ上のデータ密度ヒストグラムを用いた教師なしクラスタ分類法”，信学論，**J79-D-II**, 7, pp.1280–1290, 1996.
8) ダイヤグラムグループ編，皆川達夫監修：“楽器”，株式会社マール社
9) 田中 雅弘，古河 靖之，谷野 哲三：“自己組織化マップを利用したクラスタリング”，信学論，**J79-D-II**, 2, pp.301–304, 1996.
10) A. Savitzky and M. J. E. Golay：“Smoothing and Differentiation of Data by Simplified Least Squares Procudures”, *Anal. Chem.* **36**, 8, pp.1627–1639, 1964.
11) FFTW：http://www.fftw.org