第 6 回 東北地区音響学研究会 (2023 年 11 月 25 日) 資料番号 6-11

楽器ごとの周波数特性を考慮した特徴抽出と再帰型 NN によるドラム演奏楽音の認識の試み *

○鴨澤秀郁, 田中元志 （秋田大）

1 はじめに

通信技術の発達や音楽消費の多様化に伴い，ユーザへの楽曲推薦・検索機能などの音楽認識に関する研究の重要性が高くなってきている [1]。音楽認識には，音楽ジャンルなどを含む大域的な特徴の抽出が必要であり，そのための要素技術のひとつに打楽器（ドラム）音の認識があげられる。楽音からドラム構成楽器音を認識できれば，ドラム譜を作成できるだけでなく，テンポ検出への応用等も期待できる。

ドラム採譜に関する検討では，類似したパターンが繰り返し演奏されやすいという特徴を利用し，時系列間の関係を考慮できる再帰的ニューラルネットワーク (RNN) の利用が主流となってきている [2, 3]。それらの方法では，各楽器の発音確率（アクティベーション）を目的変数に，楽音から抽出した音響特徴量を説明変数として RNN に学習させた後，推定アクティベーション波形のピーク検出で発音を認識する。特徴量には，ログメルスペクトログラム（LMS）や定 Q 変換（CQT）スペクトログラムがよく用いられる。一方，筆者らはドラム楽器音の周波数解析を行い，楽器ごとの特徴に対応する帯域の総パワーから認識を検討した [4]。バスドラムやスネアドラムに対して従来手法と同等以上の精度で認識できた。この特徴抽出方法と，RNN を用いたアクティベーション推定法を組合せることで，従来手法と比較して認識率の向上が期待できる。

本検討では，ドラム楽器音の周波数特性に対応させたフィルタバンクを用いて算出した特徴量と，再帰型 NN によるアクティベーション推定を組合せ，ドラム演奏楽音の認識を試みる。ここでは，ドラム楽器の中で主要とされるバスドラム (KD), スネアドラム (SD), ハイハット (HH) を対象とする。KD は，低周波帯の総パワーから認識し，SD と HH は，RNN で推定したアクティベーションから認識する。公開データセットを用いた認識実験により，提案手法を評価する。

2 楽器毎の周波数特性を考慮した特徴抽出

ドラムセットの基本構成である，膜鳴楽器のバスドラム，スネアドラムと，体鳴楽器（シンバル類）音を対象に周波数解析を行った結果，

Table 1 Filterbank for feature extraction [Hz].

	Frequency range	Subband width	Shift
A	20–505	25	20
B	500–1,005	55	50
C	1,000–2,005	105	100
D	2,000–15,000	500	400
E	15,000–20,000	5,000	0

- KD は 45–75 Hz に基音のパワーが集中
- SD は倍音構造をもち，160–270 Hz に基音
- 体鳴楽器のパワーは高域（15 kHz 以上）に集中

などの主要な特徴が観測された [4]。この結果をもとに，Table 1 に示す，各楽器の周波数特性に対応させた周波数バンドとサブバンドを有する計 77 バンドのフィルタバンクを作成した。バンド A, B, C は膜鳴楽器の基音，音高要素，部分音の帯域にそれぞれ対応し，バンド D, E は体鳴楽器の音高要素，噪音成分の帯域にそれぞれ対応させている。

ドラム演奏楽音を時間-周波数解析（ハニング窓，窓長 100 ms，シフト幅 5 ms，2^{16} 点 FFT）し，パワースペクトログラムを求めた。その後，パワースペクトログラムに対し，Table 1 に示したフィルタバンクを適用した。特徴抽出の処理の例を Fig. 1 に示す。各楽器に対応する特徴が強調されている。

3 ドラム演奏楽音の認識方法

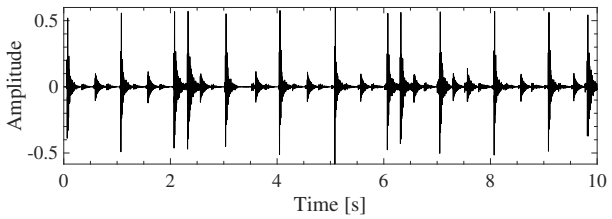
3.1 バスドラムの認識

ドラム音の周波数解析の結果から，75 Hz 以下の帯域の周波数成分は，KD が占有していることが示されている [4]。そこで，前節で述べた方法で楽音から抽出したスペクトルにおいて，75 Hz 以下 (KD 帯域) に対応するビンの総パワーを求め，その閾値判定で KD を認識する。ここで，検出閾値 TH_{KD} は，

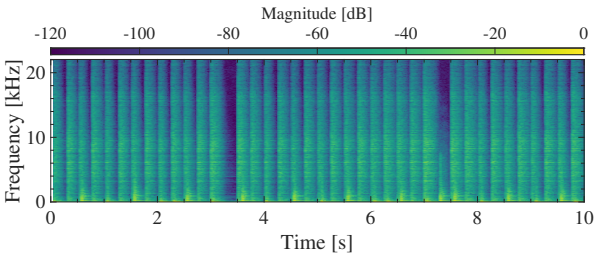
$$TH_{\mathrm{KD}} = A \cdot \sqrt{\frac{1}{N}\sum_{n=1}^{N} P(n)^2} \qquad (1)$$

で定めた。$P(n)$ は第 n サンプル時刻における KD 帯域のパワー，N は総サンプル数である。A は定数で

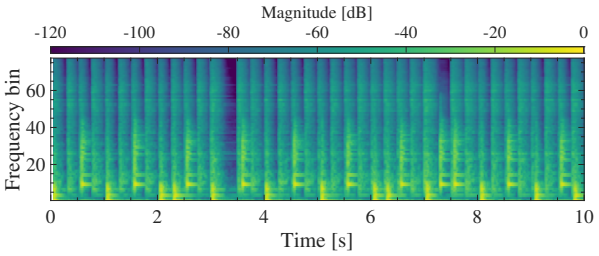
* A study on drum recognition from musical sound by using recursive neural network and feature extraction considering frequency characteristics of each instrument, by KAMOZAWA, Hidefumi and TANAKA, Motoshi （Akita University）.



(a)



(b)



(c)

Fig. 1　Example of feature extraction. (a) waveform of drum sound, (b) STFT spectrogram, and (c) spectrogram after applying the filterbank.

あり，本検討では実験的に A =1.75 とした。

例として，Fig. 1(a) の楽音から求めた KD 帯域のパワーと，それを用いた KD の発音認識結果を Fig. 2 に示す。図上部の垂直線は正解ラベル，図下部の破線は検出閾値，点は検出されたピークを示している。すべての発音が正しく検出されている。

3.2　スネアドラムとハイハットの認識

SD と HH の信号周波数は広帯域に分布し，分離が困難であるため，機械学習によるアクティベーション推定を利用する。再帰的 NN の一つである長短期記憶 (LSTM; Long short-term memory) を用いた。隠れユニット数が 128 の LSTM セルと，2 つの全結合層が密に接続された回帰モデルを作成した。各層にはドロップアウト確率 20%の Dropout 層を挿入した。Fig. 1(c) に示した，次元圧縮されたスペクトログラムを学習データとして，オンセット時刻のときに 1，それ以外の時刻では 0 の値をとるように学習させた。

LSTM を用いて，SD と HH のアクティベーションを推定した結果の例を Fig. 3 に示す。過剰なピークを抑えるために，タップ数 5 のメディアンフィルタを

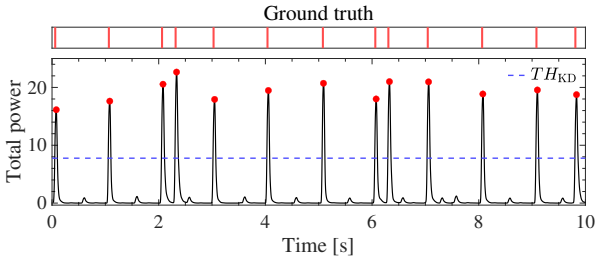
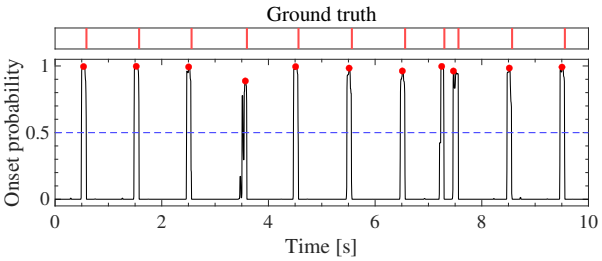
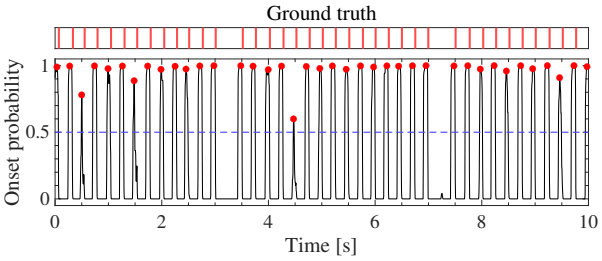


Fig. 2　Example of onset detection of kick drum.



(a)



(b)

Fig. 3　Examples of onset recognition by using LSTM. (a) SD activation and (b) HH activation.

用いてアクティベーション波形を平滑化し，閾値を超えた極大値を発音時刻として検出した。閾値は，良好な認識率が得られる値として，0.5 を実験的に用いた。すべての発音を正しく検出できている。

4　ドラム演奏楽音の認識実験

4.1　テストデータ

ドラム採譜に関する検討において最も利用されるデータセットである IDMT-SMT-Drums[5] を使用した。任意のタイミングで演奏された KD, SD, HH が，サンプリング周波数 44.1 kHz, 量子化ビット数 16 bit で，各音源につき 10–30 s 程度モノラル録音されている。データセットは，ドラムセット（音色）ごとにサブセット化されており，その中の Wave-Drum サブセット（計 70 音源）を学習に，Real-Drum サブセット（計 12 音源）をテストに用いた。

4.2　認識結果

音楽情報検索コンテストである MIREX で採用されているオンセット検出タスクの評価方法 [6] に準拠

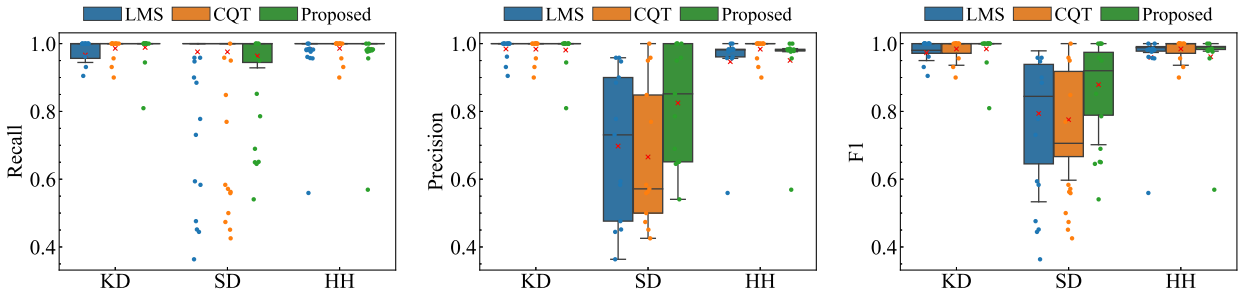


Fig. 4　Recognition results for each test drum sound. (a) Recall, (b) Precision, and (c) F1.

Table 2　Average of recognition rate [%].

		LMS	CQT	Proposed
KD	Recall	96.6	98.6	**98.9**
	Precision	**98.4**	**98.4**	98.1
	F1	97.3	98.4	**98.5**
SD	Recall	**97.6**	**97.6**	96.4
	Precision	70.0	66.6	**82.5**
	F1	79.4	77.6	**87.9**
HH	Recall	98.3	**98.6**	98.3
	Precision	94.7	**98.4**	95.0
	F1	95.9	**98.4**	96.1
Overall F1		90.9	91.5	**94.2**

し，許容誤差 ±50 ms 以内として各楽器ごとに評価指標の Recall, Precision, F1 を求めた。各テストデータごとに算出した評価指標をまとめた結果を Fig. 4 に示す。また，全データの認識結果の平均値を Table 2 にまとめる。比較のため，学習データとしてよく用いられる LMS と CQT スペクトログラムを用いた場合の結果も同時に示す。周波数解析の条件や学習方法，認識方法をすべて統一した。

認識実験の結果，KD と HH は，LMS や CQT スペクトログラムを用いて学習した場合と同等以上に認識できた。一方，SD については，他の特徴量で学習した場合よりも高い F1 値が得られた。楽器ごとの周波数特性を考慮したフィルタバンクによる特徴強調が認識率向上に寄与したことが考えられ，本特徴抽出方法のドラム演奏楽音認識への有用性が示唆される。

本手法における誤認識のほとんどは，アクティベーション波形からのピーク検出時で生じており，発音時刻のずれが許容誤差（50 ms）をこえて推定される場合が多く見受けられた。アクティベーション波形の平滑化による影響が考えられる。また，本検討で用いた閾値はすべて一定値であるため，振幅の小さい成分を見逃す場合があり，さらなる改善が必要である。

5　おわりに

ドラム構成楽器音の周波数特性を考慮した特徴抽出と，再帰的 NN によるアクティベーション推定を組み合わせ，ドラム演奏楽音の認識を試みた。公開データベースを用いた認識実験では，平均で 94.2%の F1 値が得られ，本手法の有効性を示した。演奏楽音からドラム譜を作成するためには，タムやシンバル類の認識も必要であり，これらを含めた検討は今後の課題である。

謝辞

本研究で用いたフィルタバンクのアイデアは，小岩洸喜氏（平成 30 年度本学修了）が行ったドラム音の周波数解析結果を基にした。

参考文献

[1] 亀岡他: “音楽音響信号処理技術の最先端”, 信学誌, 98(6), 467-474.
[2] C. Southall, *et al.*: “Automatic Drum Transcription Using Bi-directional Recurrent Neural Networks”, Int. Soc. Mus. Inf. Retriev. Conf., (New York, USA), 591-597, 2016.
[3] R. Vogl, *et al.*: “Recurrent Neural Networks for Drum Transcription”, Int. Soc. Mus. Inf. Retriev. Conf., (New York, USA), 730-736, 2016.
[4] 小岩, 田中: “楽器毎の周波数特性を考慮したドラム演奏楽音の認識に関する検討”, 第 1 回 東北地区音響学研究会, 1-1, 2019.
[5] C. Dittmar and D. Gärtner: “Real-Time Transcription and Separation of Drum Recordings Based on NMF Decomposition”, Int. Conf. Digit. Audio Effects (Erlangen, Germany), DAFx-14, 2014.
[6] 2021:Audio Onset Detection: https://www.music-ir.org/mirex/wiki/2021:Audio_Onset_Detection (Acceced October 30, 2023).