情報処理学会論文誌 Vol.61 No.4 777–788 (Apr. 2020)

INSTRUDIVE：楽器編成の自動認識に基づく楽曲探索システム

高橋 卓見[1,2,a)] 深山 覚[1,b)] 後藤 真孝[1,c)]

受付日 2019年7月5日, 採録日 2020年1月16日

概要：本論文では，楽曲を特徴付ける要素である楽器編成に基づいて楽曲探索を行うことができるシステム INSTRUDIVE を提案する．従来，自動認識した楽器編成を用いて楽曲を探索するための効果的な方法は明らかにされておらず，商用音楽配信サービスでも利用されていなかった．本研究では，大規模な楽曲群を扱う前の第 1 段階として，様々なジャンルの 122 曲で構成される小規模な研究用公開楽曲データセットを対象に，既存の自動楽器認識手法より高性能な手法を提案したうえで，実際に楽器編成に基づいて楽曲探索が可能な方法を示すことを目的とする．提案する楽器認識手法は，畳み込みニューラルネットワーク（CNN）モデルに基づいている．その畳み込み層は音色的な特徴を表現するために周波数軸方向のみに畳み込んだ後に，その時間変化を表現するために時間軸方向のみに畳み込む点が特徴的であり，画像処理で一般的な正方形カーネルを用いた畳み込みとは異なる．最新手法を含む 3 種類の既存楽器認識手法との比較評価の結果，提案手法が最も高い認識性能を示した．次に，その認識結果を楽曲探索に役立たせることができる具体的な方法として，楽器編成に基づく新機能を有する楽曲探索インタフェースを提案する．そのインタフェースを実装し，24 名の被験者に対して予備評価実験をした結果，実験者が監視していない状況でも，被験者は各楽曲の楽器編成を可視化した円グラフでその違いを把握しながら，楽曲探索が可能なことを確認した．

キーワード：音楽情報検索，音楽インタフェース，音楽可視化，楽器認識，畳み込みニューラルネットワーク

INSTRUDIVE: A Music Exploration System Based on Automatic Recognition of Instrumentation

Takumi Takahashi[1,2,a)] Satoru Fukayama[1,b)] Masataka Goto[1,c)]

Received: July 5, 2019, Accepted: January 16, 2020

Abstract: In this paper, we propose a music exploration system INSTRUDIVE focusing on instrumentation that is a key factor in determining musical sound characteristics. Conventionally, an effective method for exploring musical pieces by using automatically recognized musical instruments has not been clarified, and it has not been used in commercial music distribution services. The goal of this research is to propose an automatic instrument recognition method that outperforms existing methods and then show a method that enables music exploration based on instrumentation. Our instrument recognition method is based on a convolutional neural network (CNN) model whose layers first convolve input along the frequency axis to express timbre characteristics, and then convolve along the time axis to express their temporal characteristics. It thus uses a convolution kernel that is different from a typical square kernel used in image processing. In our evaluation, we confirmed that the proposed method was superior to three existing methods. We then propose a music exploration interface with a new function based on the instrumentation. As a result of implementing the interface and conducting a preliminary experiment on 24 subjects, even when the experimenter was not monitoring, we confirmed that subjects were able to explore musical pieces by using a pie chart that visualized the instrumentation of each song.

Keywords: music information retrieval, music interface, music visualization, instrument recognition, convolutional neural network

1. はじめに

楽曲によって様々な音色の楽器が使い分けられており，楽器編成は楽曲の特徴を決める重要な要素である．たとえば，ボーカル，エレキギター，エレキベース，ドラムという楽器編成は，ポップスやロック，メタルといった音楽が連想されやすいが，クラシックやエレクトロダンスミュージックのような音楽は連想されにくい．一方で，楽器編成によってジャンルを1つに特定することはできない．たとえば，ポップス，ロック，ファンク，フュージョンは，同じような楽器編成で演奏されることがある．そのため，楽器編成に注目して楽曲を探すことができれば，ジャンルとは異なる基準で，ある楽曲と同じ，少し異なる，またはまったく異なる楽器編成を持つ楽曲を探索することが可能となる．

ここで，本論文で述べる「楽曲を探索する」とは，確たる目的を持たずに好みの楽曲を探すことを意味する．たとえば，好きなアーティストの新曲を調べるような場合というよりは，なんとなく楽曲を流してみて気に入るものを見つける，また，ある好みの楽曲からそれに似た楽曲を探す，といった楽曲の探し方を表す．

楽曲の探索のために，様々な手法が研究されている．第1に，人手で付けられたジャンルやタグなどのメタ情報を用いた楽曲分類を活用する手法 [1], [2] がある．これは言葉によって楽曲を表現するため，その言葉に共通の認識を持つ聴取者同士では音楽の表現に役立つ一方，認識に齟齬がある場合，聴取者は言葉を理解できず，また楽曲分類の細分化に限りがあるといったデメリットがある．第2に，ある聴取者に対して嗜好の類似した他の聴取者の情報を用いて楽曲を推薦する協調フィルタリング [3] を用いる手法 [4], [5] がある．しかし，これは必ずしも楽曲の音響信号に含まれる情報を利用しないため，履歴のデータが豊富である有名な楽曲が推薦されやすい傾向がある．第3に，音響信号処理によって音楽的特徴を抽出する楽曲分類として，楽曲のジャンルを推定する自動ジャンル分類 [6]，楽曲に付されたタグを推定する自動タグ付け [7] などを用いる手法がある．こういった手法は自動的に楽曲を扱うため大量の楽曲に対しても有効である一方，実用化のためには高い精度の自動認識を実現する必要がある．また，自動楽器認識技術を用いた類似楽曲検索は文献 [8] で議論されていたが，楽曲探索のためのインタフェースは十分には検討されていなかった．

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305–8568, Japan
[2] 筑波大学
University of Tsukuba, Tsukuba, Ibaraki 305–8577, Japan
[a)] takahashi.takumi@aist.go.jp
[b)] s.fukayama@aist.go.jp
[c)] m.goto@aist.go.jp

そこで，本研究では音響信号から推定される楽器編成に基づいて楽曲を探索できるシステム INSTRUDIVE を提案する．楽曲探索のインタフェースを構築する際に，自動認識した楽器編成を用いる効果的な方法が従来は明らかになっておらず，まずは学術的に基礎的な検討が重要な状況であることから，本研究では商用音楽配信サービスのような大規模な楽曲群を扱うのではなく，その前の第1段階として，小規模ではあるものの，様々なジャンルの122曲で構成される研究用公開楽曲データセット MedleyDB [9] を対象とする．そして，既存の自動楽器認識手法より高性能な手法を提案したうえで，実際に楽器編成に基づいて楽曲探索が可能な方法を示すことを目的とする．本研究は，以前に我々が行った楽曲探索システムの研究 [10] の発展版であり，発表後に行った追加実験およびインタフェースの被験者実験の結果をふまえて，改めて論旨を組み立てたものである．

2. INSTRUDIVE の持つべき要件

INSTRUDIVE の目的は，楽器編成に基づいて任意の楽曲の探索を実現することである．そのためには，任意の楽曲の 1) **楽器編成の自動認識**の仕組みと，認識した楽器編成を用いた 2) **楽曲探索インタフェース**が必要である．

1) 楽器編成の自動認識に関しては，楽曲の音響信号を解析し，楽器編成を認識する手法が必要である．認識する楽器の種類には，楽曲を分類するのに適し，かつ研究用楽曲データセット MedleyDB [9] において比較的多く使用されていることが確認できた10種類（Acoustic Guitar, Clean Electric Guitar, Distorted Electric Guitar, Drums, Electric Bass, FX, Piano, Synthesizer, Violin, Voice）を用いる．

2) 楽曲探索インタフェースに関しては，ユーザが楽器編成に基づいて楽曲を選び，再生できる仕組みが必要である．そのためには，多くの楽曲を一度に俯瞰し，それぞれの楽器編成の違いや類似性を視覚的に把握したうえで，楽曲を再生できるようにするとよいと考えられる．

以上の要件を満たす INSTRUDIVE の概要を**図 1** に示す．1) 楽器編成の自動認識では，楽曲の単位時間（1秒間）ごとの音響信号に前処理を行い，畳み込みニューラルネットワーク（CNN）に入力することで，単位時間ごとの楽器編成を認識し，1曲分の認識結果を統合することで，楽曲全体の楽器編成を推定する．2) 楽曲探索インタフェースでは，各楽曲の楽器編成の推定結果を，円グラフを用いることでアイコンとして表す．円グラフの各色は楽器の種類を，面積比はその楽曲中での演奏時間の比率を表す．得られた円グラフ群を，それらの類似性に基づいてインタフェース上に配置することで，楽器編成に基づいた楽曲探索を実現する．また，楽曲の再生中に，その楽曲全体の楽器編成をよく表す部分を把握できるようにするため，楽器編成の時

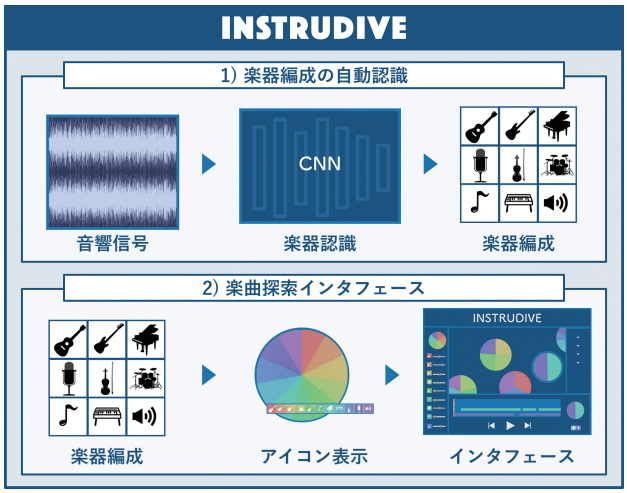

図 1 INSTRUDIVE の概要
Fig. 1 Overview of INSTRUDIVE.

間的な変化を表示する．そのほか，楽器編成に基づいて検索する機能，次に再生する楽曲をあらかじめ指定する機能，類似した楽器編成を連続再生する機能など，効果的な楽曲探索を実現するための一連の機能を備えている．

3. 関連研究

1) 楽器編成の自動認識は，自動楽器認識の研究として，2) 楽曲探索インタフェースにおける楽曲のアイコン表示は，音楽情報可視化の研究として位置付けられる．本章ではそれらの関連研究について述べる．

3.1 自動楽器認識

自動楽器認識の研究タスクは，対象とする音響信号と分類問題の種類に基づいて以下の 5 種類に分けることができる．ここで，混合音とは複数の楽器音が混ざった音響信号を意味する．

A. 単一の楽器音の音響信号に対して，単一の楽器を認識する問題 [11], [12], [13], [14], [15], [16], [17], [18]
B1. 音源モジュールなどを用いて合成された混合音に対して，メインとなる単一の楽器を認識する問題 [14], [19]
B2. 市販の楽曲およびそれに類する混合音に対して，メインとなる単一の楽器を認識する問題 [20], [21], [22]
C1. 音源モジュールなどを用いて合成された混合音に対して，すべての楽器を認識する問題 [8][*1], [16], [23], [24], [25]
C2. 市販の楽曲およびそれに類する混合音に対して，すべての楽器を認識する問題 [26], [27], [28], [29]

最も難しい問題はタスク C2 であり，応用用途が幅広く重要性が高いと考えられ，本研究ではこの種類の自動楽器認識に取り組む．このようなタスクを重視したデータセット（MedleyDB [9]，MusicNet [30]，OpenMIC-2018 [31]）が発表されており，より手軽に取り組むことができるようになってきた．

楽器認識問題において複数の手法の性能を比較した先行研究 [21], [28] では，CNN 以外の機械学習手法を用いる既存手法に対し，CNN を用いる手法の方が高い認識精度を有することが示されていたため，本研究では CNN を用いる手法を改良することとした．

[*1] 文献 [8] は録音された音響信号もあわせて使用されているが，楽器の種類が制限されているためここに分類した．

3.2 音楽情報可視化

音響信号処理を利用した楽曲ごとの音楽情報可視化は大きく 2 つに分けられる．1 つは，楽曲の集合を可視化する研究である．複数の楽曲の類似性を位置関係，色，模様などによって表現する手法 [32], [33], [34], [35], [36], [37], [38] が提案されている．これらの手法の多くは，楽曲の集合から好みの楽曲を探索することを目的としている．もう 1 つは，各楽曲中の音楽的要素を可視化する研究である．楽曲の繰返し構造を可視化する手法 [39], [40], [41], [42]，楽曲のメロディやコード進行などの音楽的要素を可視化する手法 [43]，楽曲の盛り上がりや感情の変化を可視化する手法 [44] が提案されている．これらの手法の主な目的は，繰返し構造などをもとに楽曲の全体像を視覚的に把握すること，もしくは音楽鑑賞を豊かにすることである．

本研究は，これら両方の可視化に関連し，楽曲探索および音楽鑑賞のいずれにも寄与する研究として位置付けられる．さらに，これまで注目されてこなかった，楽器編成に着目した楽曲の音楽情報可視化を行う研究として位置付けられる．

4. 自動楽器認識

本研究で提案する自動楽器認識では CNN を用いる．楽曲の音響信号から楽器編成を認識するため，前処理として単位時間（1 秒間）ごとの楽曲の音響信号を多次元ベクトルに変換し，それを CNN に入力する．

4.1 モデルの特徴

提案する CNN モデルの特徴は主に 2 つある．1 つは，入力データに振幅スペクトログラムを用いることである．複数楽器の混合音を対象とする楽器認識の場合，メルスペクトログラムのようにスペクトル包絡を利用した次元削減を行うと，詳細な倍音成分の情報が失われてしまうため，次元削減の行われない振幅スペクトログラムを採用した．もう 1 つは，畳み込み層のフィルタサイズである．スペクトログラムは画像と違い，縦軸が周波数，横軸が時間と，異なるデータを表すため，先行研究 [21] のように両軸方向に大きさを持つフィルタは非効率だと考え，各フィルタが畳み込む内容を分離した．また，周波数と時間それぞれ 2 層ずつのフィルタがそれぞれ異なる種類の音響的性質を解析するようにフィルタの大きさに変化を持たせた．

表 1 本研究で提案する CNN モデル．Conv, Pool, Dropout, Dense はそれぞれ畳み込み層，マックスプーリング層，ドロップアウト層，全結合層を表す．Output Size は各層から出力されるテンソルの次元数である

Table 1 Architecture of proposed CNN model. Conv, Pool, Dropout, and Dense are convolutional layer, pooling layer, dropout layer, and dense layer, respectively. Output Size denotes tensor dimensions output from each layer.

Layer	Output Size
Input layer	$1{,}024 \times 87 \times 1$
Conv (4×1)	$1{,}024 \times 87 \times 32$
Pool (5×3)	$204 \times 29 \times 32$
Conv (16×1)	$204 \times 29 \times 64$
Pool (4×3)	$51 \times 9 \times 64$
Conv (1×4)	$51 \times 9 \times 64$
Pool (3×3)	$17 \times 3 \times 64$
Conv (1×16)	$17 \times 3 \times 128$
Pool (2×2)	$8 \times 1 \times 128$
Dropout (0.5)	1,024
Dense	1,024
Dense	121
Dense	11

4.2 モデル構造

提案する CNN モデルの構造を**表 1** に示す．ただし，すべての全結合層（Dense）および畳み込み層（Conv）には，バッチノーマライゼーション [45] が行われ，活性化関数は出力層には Sigmoid，それ以外の層には ReLU を用いた．また，重みの初期化には He ら [46] の初期化を利用した．すべてのプーリング層（Pool）にはマックスプーリングが用いられた．学習は 1,000 エポック行い，バッチサイズは 128 とした．

本モデルでは，4.1 節で述べたように，周波数軸と時間軸のそれぞれの方向に畳み込む特徴を持っている．そこで，音色的な特徴を表現するために周波数軸方向のみに 2 層分畳み込んだ後に，その時間変化を表現するために時間軸方向のみに 2 層分畳み込んだ．この計 4 層の各層の畳み込みフィルタの形状（大きさ）の組合せには様々な可能性があるため，**表 2** に示す 8 種類を性能比較する予備実験をした結果，最も高い認識精度だった 2 行目のフィルタ形状の組合せを提案モデルとして採用した．表 1 の Conv の行の () 内に対応している．なお，楽器音では周波数成分の比率で表現される音色が時間変化することが重要だと考えて周波数軸方向の後に時間軸方向に畳み込んだため，逆順のモデル（時間軸方向に畳み込んだ後に周波数軸方向に畳み込むモデル）は検討の対象外とした．

4.3 前処理

使用する楽曲は，サンプリング周波数が 44,100 Hz，ビット深度が 16 のステレオ音源とし，これをモノラルに変換したあと 1 秒間ずつに切り分け，各 1 秒間の音響信号に対して，窓幅は 2,048 サンプル，ホップサイズは 512 サンプルとして短時間フーリエ変換（STFT）を行うことで，振幅スペクトログラムを得る．それぞれの振幅スペクトログラムに対して，平均を 0，分散を 1 とする標準化を行うことで，結果として $1{,}024 \times 87$ 次元ベクトル（周波数方向に 1,024，時間方向に 87）を得る．

表 2 予備実験で性能比較した畳み込みフィルタ形状の 8 種類の組合せ．4 つの畳み込み層を入力層側から昇順に Conv 1 から Conv 4 と記した．表中 2 行目の組合せを提案モデルに採用した

Table 2 Eight combinations of convolutional filter shapes used for our preliminary experiments for performance comparison. Four convolutional layers are numbered in ascending order (Conv 1 to 4) from the input layer. The combination of the second line in this table is used for the Proposed model.

Conv 1	Conv 2	Conv 3	Conv 4
(4×1)	(4×1)	(1×4)	(1×4)
$(\mathbf{4 \times 1})$	$(\mathbf{16 \times 1})$	$(\mathbf{1 \times 4})$	$(\mathbf{1 \times 16})$
(4×1)	(32×1)	(1×4)	(1×32)
(4×1)	(64×1)	(1×4)	(1×16)
(4×1)	(64×1)	(1×4)	(1×64)
(4×1)	(128×1)	(1×4)	(1×16)
(16×1)	(4×1)	(1×16)	(1×4)
(32×1)	(4×1)	(1×32)	(1×4)

5. 自動楽器認識手法の評価

提案モデルの性能を評価するため，既存の自動楽器認識手法との比較を行った．

5.1 データセット

データセットには，様々なジャンルの楽曲が含まれる MedleyDB [9] を用いた．MedleyDB には，122 曲のマルチトラック楽曲およびそれに対応する楽器アクティベーションが含まれる．楽器アクティベーションは，各楽曲の個々の楽器の録音データ（ステムと呼ばれる）のミックス時の音量をもとに，46.4 ms の時間フレームごとに各楽器のエネルギーを 0 から 1 の値で表した時系列データである．MedleyDB に含まれるすべての楽曲の長さは合計約 26,220 秒，1 曲あたりの長さは平均約 215 秒（最小約 17 秒，最大約 1,061 秒）である．

ウェブ上に公開されているソースコード [28] を用いて楽器ラベルの生成とデータ分割を行った．このソースコードでは，70 個の楽器カテゴリのうち，20 曲以上で使われている楽器カテゴリを採用することで，10 個の楽器カテゴリ（Acoustic Guitar，Clean Electric Guitar，Distorted Electric Guitar，Drums，Electric Bass，FX，Piano，Synthesizer，Violin，Voice）が得られる．楽器カテゴリ名は

必要に応じて，システム利用者にとって理解しやすいと思われる表記に改めた．また，ここに含まれないすべての楽器のカテゴリを Others とすることで，本研究では合計 11 個の楽器カテゴリを用いた．

データセットを 5 分割し，そのうち 4 つを学習に，1 つを評価に用いた．このとき，同一の楽曲中の異なる部分が学習と評価に用いられてしまうことがないようにするため，分割は楽曲単位で行った．また，5 分割したデータセットそれぞれに含まれる楽器ラベルの割合を均等にするために，文献 [47] のアルゴリズムを適用した．

5.2 比較手法

提案モデルを 3 つの既存手法と比較した．1 つは問題の難しさの度合いを見極めるためのベースライン手法として機械学習 Support Vector Machine（SVM）を用いた手法，残りの 2 つは最先端の手法との比較のために近年の深層学習 CNN を用いた楽器認識手法を選択した．

(1) 特徴ベクトルを入力とする SVM を用いた手法 [16]（Bag-of-features 手法と呼ぶ）．1 秒ごとの音響信号から抽出した 136 次元の特徴ベクトル [48] を Radial Basis function（RBF）カーネルの SVM に入力して学習を行った．

(2) メルスペクトログラムを入力とする CNN を用いた手法 [21]（Han 手法と呼ぶ）．1 秒ごとの音響信号のサンプリング周波数を 44,100 Hz から 22,050 Hz にダウンサンプリングし，メルスペクトログラムに変換した後，平均を 0，分散を 1 とする標準化を行い，**表 3** に示す CNN モデルに入力した．すべての全結合層（Dense）および畳み込み層（Conv）には，バッチノーマライゼーションが行われ，各層の活性化関数には LReLU（$\alpha = 0.33$）を，出力層の活性化関数には Sigmoid を用いた．学習は 1,000 エポック行い，バッチサイズは 128 とした．

(3) 音響信号の波形を入力とする CNN を用いた手法 [28]（Li 手法と呼ぶ）．1 秒ごとの音響信号に対して，平均を 0，分散を 1 とする標準化を行い，**表 4** に示す CNN モデルに入力した．すべての全結合層（Dense）および畳み込み層（Conv）には，バッチノーマライゼーションが行われ，各層の活性化関数には ReLU を，出力層の活性化関数には Sigmoid を用いた．学習は 1,000 エポック行い，バッチサイズは 128 とした．

5.3 評価指標

1 秒ごとの認識結果（予測ラベルと呼ぶ）は 11 個の楽器カテゴリに対応する 11 次元ベクトルであり，各要素は 0 から 1 の値であるため，予測ラベルに該当するか否かのバイナリ値で表されることがふさわしい．よって，評価のために予測ラベルの各要素の閾値を 0.5 と定め，それ以上の値を 1（True/その楽器が演奏されている），それ未満の値を 0（False/その楽器が演奏されていない）とした．楽器認識では，3 つの楽器が同時に演奏していれば，11 次元ベクトル中，その楽器に対応する 3 カ所が True となる．これは，True の要素が複数存在するマルチラベル分類問題であるため，提案モデルの性能評価では，マルチラベル分類の性能比較で一般的に用いられる以下の 3 つの指標を用いた．

- F-micro は，すべての楽器カテゴリのすべての 1 秒ごとの認識結果を同列に扱って再現率と適合率を求め，それらの調和平均である F-measure を計算した結果である．楽器カテゴリごとの出現頻度の偏りの影響を受け，高頻度の楽器カテゴリの認識精度が反映されやす

表 3 Han ら [21] の CNN モデル．表中の記法は表 1 と共通である．ただし，Global pool はグローバルプーリング層である
Table 3 CNN model of Han et al. [21]. Wordings are same as ones in Table 1. Global pool represents global pooling layer.

Layer	Output Size
Input layer	$128 \times 43 \times 1$
Conv (3×3)	$130 \times 45 \times 32$
Conv (3×3)	$132 \times 47 \times 32$
Pool (2×2)	$44 \times 15 \times 32$
Dropout (0.25)	$44 \times 15 \times 32$
Conv (3×3)	$46 \times 17 \times 64$
Conv (3×3)	$48 \times 19 \times 64$
Pool (2×2)	$16 \times 6 \times 64$
Dropout (0.25)	$16 \times 6 \times 64$
Conv (3×3)	$18 \times 8 \times 128$
Conv (3×3)	$20 \times 10 \times 128$
Pool (2×2)	$6 \times 3 \times 128$
Dropout (0.25)	$6 \times 3 \times 128$
Conv (3×3)	$8 \times 5 \times 256$
Conv (3×3)	$10 \times 7 \times 256$
Global pool	$1 \times 1 \times 256$
Dense	1,024
Dropout (0.5)	1,024
Dense	11

表 4 Li ら [28] の CNN モデル．表中の記法は表 1 と共通である
Table 4 CNN model of Li et al. [28]. Wordings are same as ones in Table 1.

Layer	Output Size
Input later	$44{,}100 \times 1$
Conv (3101)	$41{,}000 \times 256$
Pool (40)	$2{,}049 \times 256$
Conv (300)	$1{,}750 \times 384$
Pool (30)	87×384
Conv (20)	68×384
Pool (8)	16×384
Dropout (0.5)	16×384
Dense	400
Dense	11

表 5 122 曲中の各楽器カテゴリの出現頻度の比率（全楽曲の正解ラベルにおける，1 秒間ごとの楽器カテゴリの個数の比率）
Table 5 Ratio of appearance of instrument categories in 122 musical pieces (ratio of appearance of one-second ground-truth instrument categories in all musical pieces).

Instrument Label	Frequency
A. Guitar	0.497
Clean E. Guitar	0.319
Distorted E. Guitar	0.719
Drums	0.707
E. Bass	0.649
FX	0.073
Piano	0.701
Synthesizer	0.136
Violin	0.504
Vocal	0.826
Others	0.776

い欠点を持つ．

- F-macro は，F-measure を楽器カテゴリごとに求めた後に，それら 11 個の値を平均した値である．まず，各楽器カテゴリについて，すべての 1 秒ごとの認識結果から再現率と適合率を求め，それらの調和平均として F-measure を計算する．次に，そうして求めた楽器カテゴリ数分の 11 個の F-measure の平均値を求める．
- AUC（Area Under the Curve）は，ROC（Receiver Operating Characteristic）曲線下部の面積を楽器カテゴリごとに求めて平均した値である．

楽曲中で使われる楽器カテゴリの出現頻度に偏りがなければ F-micro と F-macro は近い特性となるが，実際には**表 5** に示す偏りがあるため，これらの指標が重要となる．本研究の目的では，楽器認識の結果を楽器編成として活用することで楽曲探索に応用することを意図しているため，頻度が高く学習データが豊富な楽器カテゴリのみで認識精度が高い状態は望ましくなく，すべての楽器カテゴリで認識精度が高い方が望ましい．そこで我々は，楽器カテゴリの出現頻度に影響されにくい F-macro および AUC の結果をより重視し，F-micro は参考情報として用いることとした．

5.4 結果と考察

実験結果を**図 2** に示す．提案手法（Proposed）は F-micro，F-macro，AUC それぞれにおいて最も高い精度を示した．2 番目に高い精度を示した Han 手法と比較すると，予測ラベルの楽器カテゴリ別の評価の影響が大きい F-macro において約 8 ポイント高い値を示していたことから，提案手法はデータ数の少ない楽器カテゴリに対してもより高い精度で認識できたことが分かる．F-micro より，我々が重視する F-macro において顕著に改善されていることは望ましい結果だといえる．

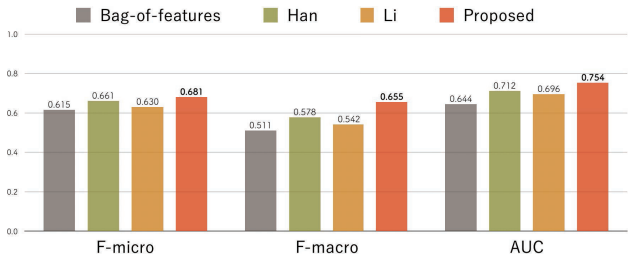

図 2 楽器認識の実験結果
Fig. 2 Experimental results of instrument recognition.

Han 手法と比べて提案手法が高い精度を示した要因として，Han 手法のモデルが次元数の小さいメルスペクトログラムの入力を用いており，高い周波数帯域の解像度が低い分，混じり合った楽器の音色を認識するのに必要な倍音の情報が欠落しているためだと考えられる．また，提案手法がフィルタ形状を工夫している点も性能向上に寄与したと考えられる．

1 章で述べたように，本研究では，既存の自動楽器認識手法よりも高性能な手法を提案する目的に加え，それによって得られる楽器編成に基づいて，実際に楽曲探索が可能な方法を示すことも目的としている．そこで次章以降では，本研究の重要な貢献の 1 つとして，楽器認識がなければ実現できない具体的な楽曲探索インタフェースを提案するとともに，それを実装して予備評価実験を実施した結果を報告する．

6. 楽曲探索インタフェース

本章では楽器編成に基づいて楽曲探索を行うためのインタフェースについて述べる．このインタフェースでは，自動認識した楽器編成に基づいて楽曲を円グラフとして可視化することで，楽曲の中身を視覚的に把握しながら探すことができる．

インタフェースは大きく 4 つの機能（楽器編成マップ，ビジュアルプレイヤー，検索機能，プレイリスト）で構成され（**図 3**），プログラミング言語 Python を用いて実装した．

6.1 楽器編成のアイコン表現

円グラフを用いて楽器編成を可視化した例を**図 4** に示す．この図は自動楽器認識における正解データを用いて作成された．1 つの円グラフは 1 つの楽曲を表すアイコンとして用いられ，色は楽器の種類に，面積比率は各楽器の演奏時間の比率に対応する．図より，異なる楽器編成の楽曲がそれぞれ異なる外見で表されていることが分かる．

この円グラフを作成するためには，楽曲の長さに対する各楽器の演奏時間の比率を求め，その楽曲における全楽器の演奏時間合計に対する各楽器の演奏時間の比率を求

図 3 INSTRUDIVE のインタフェース画面
Fig. 3 Screenshot of INSTRUDIVE interface.

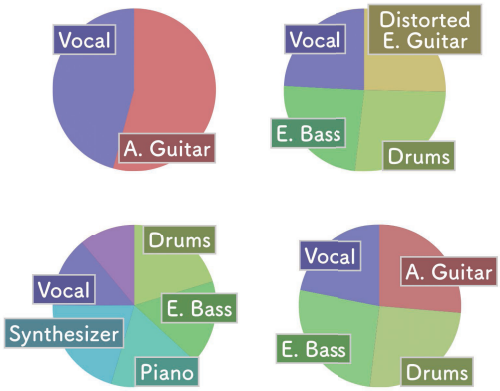

図 4 円グラフを用いた楽曲の楽器編成の可視化の例
Fig. 4 Multi-colored pie charts depicting instrumentation.

める必要がある．本論文では，前者を各楽器の絶対存在率（Absolute Appearance Ratio），後者を各楽器の相対存在率（Relative Appearance Ratio）と呼ぶ．楽曲 p の長さを T_p 秒，楽器 i の演奏時間を t_i としたとき，絶対存在率 AAR_{pi} は

$$AAR_{pi} = \frac{t_i}{T_p} \tag{1}$$

で定義され，相対存在率 RAR_{pi} は

$$RAR_{pi} = \frac{AAR_{pi}}{\sum_i AAR_{pi}} \tag{2}$$

で定義される．色相環上で等間隔になるように各楽器の色を割り当て，円グラフに占める各色の面積比率に相対存在率を用いることで，楽曲の円グラフが作成できる．本論文ではインタフェース上での見やすさを考慮して，彩度を 50%，明度を 78%とし，色相は楽器カテゴリ名のアルファベット順に 0 度から 33 度ずつ増加していくように色の割当てを行った．

6.2 楽器編成マップ

楽器編成マップ（図 3・中央）には，MedleyDB のすべての楽曲（122 曲）がそれぞれ円グラフとして 2 次元平面上に表示される．その表示方法について説明する．

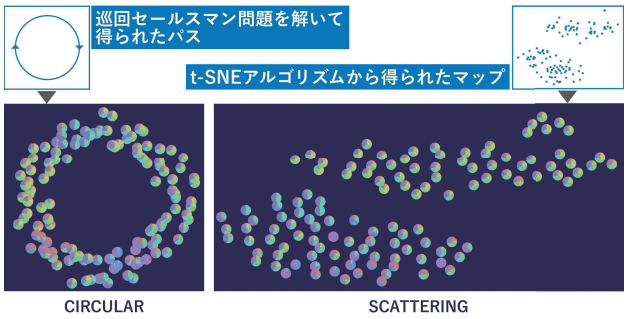

図 5 楽器編成マップにおける楽曲のアイコンの配置の例
Fig. 5 Two modes of arrangement of pie charts in instrumentation map.

各楽曲の楽器編成は「楽器編成ベクトル」として表現できる．このベクトルは，楽器カテゴリ数を次元数とし，各楽器カテゴリの絶対存在率を要素に持つ．扱われるすべての楽曲の楽器編成ベクトルをそれぞれ空間上の座標と見なして巡回セールスマン問題（すべての点を 1 度ずつ巡り初めの点に戻る巡回路の最短経路を求める問題）を解くと，解として得られるパスにおいて隣接する楽器編成ベクトルどうしは類似する．したがって，楽曲の円グラフをその順序で円形に並べることで，似た楽器編成を持つ楽曲が近くなるように円形に配置することができる（**図 5**・左）．この配置方法を CIRCULAR モードと呼ぶ．ただし，円グラフどうしの重なり合いを防ぐため，互いに一定の距離が保たれるように動径方向に拡散させた．また，多次元ベクトルの可視化に適した次元削減手法である t-SNE アルゴリズムを用いて楽器編成ベクトルの次元を 2 次元に削減することで，似た楽器編成を持つ楽曲が近くに配置されるように平面上に分散させて表示することができる（図 5・右）．この配置方法を SCATTERING モードと呼ぶ．この配置方法でも，円グラフどうしの重なり合いを防ぐため，互いに一定の距離が保たれるように拡散させた．CIRCULAR モードと SCATTERING モードは自由に切り替えることが可能である．

楽器編成の提示に基づいて直感的に楽曲探索を行えるようにするため，円グラフをクリックすることでメニューが開き，楽曲の再生もしくはプレイリスト（6.5 節参照）への追加を行えるようにした．

6.3 ビジュアルプレイヤ

ビジュアルプレイヤ（図 3・下）は，楽器編成を色やアイコンで可視化することによって，楽器編成に注目して音楽を聴取することを可能とする．

ビジュアルプレイヤの最下部には，「再生・停止」「次の曲」「前の曲」の 3 つのボタンが配置されており，楽曲の再生状態の制御ができる．その上部には曲名とアーティスト名が表示されており，さらにその上部には再生中の楽曲の楽器編成における時間的変化がタイル状のグラフで可視化

されている．それぞれのタイルは 1 秒間において各楽器が使用されているかどうかを表す．このグラフに表示されるのは 60 秒間分であるが，枠内を左右にスクロールすることができ，さらにクリックすることで好みの時刻に再生位置を移動することができる．たとえば，ボーカルが入る部分をグラフから探し，その時刻まで再生位置を移動するといった使い方ができる．

図 3 右下の円グラフの形は毎秒変化し，その 1 秒間における楽器編成を表す．また，その下にはその楽曲中で主に使用されている楽器のイラストが表示される．

6.4 検索機能

検索機能（図 3・左）を用いると，楽器編成をクエリとして楽曲検索を行うことができる．

楽器のイラストが描かれたボタンをクリックすることでボタンの色は反転し，選択されたことを表す．選択された楽器は上部の円グラフに反映され，ボタン上部のスライダによって比率の調整を行うことができるので，特に重視して検索に反映させたい楽器を細かく設定することができる．

これらのスライダはそれぞれ楽器の絶対存在率を表し，クエリは楽器編成ベクトルである．クエリに似た楽器編成の楽曲を探すには，クエリである楽器編成ベクトル $\vec{q}$ と楽曲 p の楽器編成ベクトル $\vec{d}_p$ のコサイン類似度

$$sim(\vec{q}, \vec{d}_p) = \frac{\vec{q} \cdot \vec{d}_p}{|\vec{q}||\vec{d}_p|} \tag{3}$$

をすべての楽曲 p に対して計算すればよい．Search と書かれたボタンをクリックして検索を実行すると，その結果の上位 10 曲が，次節で述べるプレイリストに加えられる．

6.5 プレイリスト

プレイリスト（図 3・右）には，次に再生される楽曲が上から順に表示される．各楽曲の円グラフ，曲名，アーティスト名，再生ボタンと削除ボタンが表示されており，特定の楽曲を再生またはリストから削除することができる．たとえば，検索機能で選ばれた 10 曲がプレイリストに表示されている状態では，検索したクエリに近い順に楽曲が再生されていく．

7. 楽曲探索インタフェースの予備評価実験

INSTRUDIVE のインタフェースによって，実際に楽器編成に基づく楽曲探索が可能になったことを実証するために，被験者 24 名（19〜25 歳の学生）を対象とする予備評価実験を行った．なお，インタフェースの各機能の有用性をヒューマンコンピュータインタラクションの観点から網羅的に評価することは本論文の範囲を超えるため意図しておらず，この予備評価実験はその前段階として，上記の実証をすることを意図して実施した．そのうえで，ユーザが

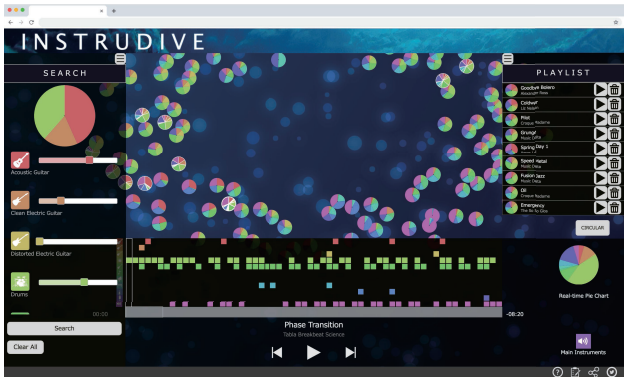

図 6 INSTRUDIVE ウェブアプリケーションの画面
Fig. 6 Screenshot of INSTRUDIVE web application.

楽器編成の可視化および楽器編成を用いた楽曲探索に対してどのように感じるかについて，今後の研究に役立つような基礎的な知見も得られることを狙っている．

オンライン上で実験を行うために，INSTRUDIVE インタフェースの挙動を再現するウェブアプリケーションを，プログラミング言語 JavaScript を用いて再実装した（**図 6**）．

7.1 実験手順

実験者が付き添っていない状況でも，被験者が INSTRUDIVE によって独力で楽器編成に基づく楽曲探索が可能になることを実証するために，実験に際しては実験者が監視することなく，被験者の望む時間に各自の自宅において行われた．被験者は，まず指定されたウェブページにアクセスし，INSTRUDIVE の概要説明を読み，使用方法を説明する動画を観てから INSTRUDIVE を使用した．ただし，必ず 10 分間以上は使用することとし，そのあとは可能な範囲で日常的な音楽聴取と同じ状況にするため，既存の音楽プレイヤと同じように，別の作業をしてもよいので使用をやめたくなるまで使用を続け，使用が終わればアンケートに回答するように指示した．アンケートは，選択肢付きの質問，自由記述の質問で構成されており，回答はウェブ上のアンケートフォームを用いて行われた．

7.2 結果と考察

楽曲探索のために利用した，楽曲の可視化手法と楽器編成を用いた検索機能に関する 4 つの質問項目を以下に示す．

- 質問 a：円グラフの違いを見ることで，楽曲ごとの楽器編成の違いを容易に把握することができた．
- 質問 b：円グラフの表示位置や色の違いをもとに，その楽曲がどのような音楽であるか，実際に聴く前に予想できた．
- 質問 c：楽器編成をクエリとして行う楽曲検索は楽曲を探すのに便利だ．
- 質問 d：INSTRUDIVE が自動的に楽器編成を推定して表示した円グラフは，楽器編成に基づいて楽曲を探

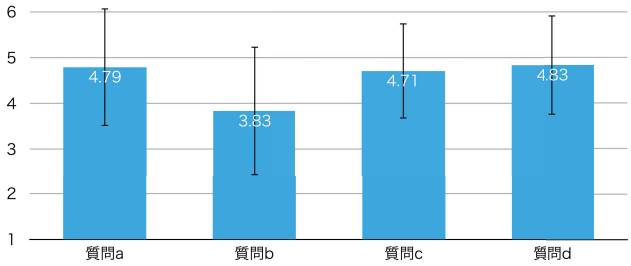

図 7 INSTRUDIVE インタフェースに対するアンケートの回答結果．黒のエラーバーは標準偏差を表す

Fig. 7 Results of questionnaire on INSTRUDIVE interface. Each black line represents range of standard deviation.

すのに_____．

各質問はそれぞれ 6 件法で，質問 a，b，c の選択肢は［6…とてもよくあてはまる/5/4/3/2/1…まったくあてはまらない］，質問 d の選択肢は［6…とても役に立った/5…役に立つことが多かった/4…どちらかといえば役に立つことが多かった/3…どちらかといえば役に立たないことが多かった/2…役に立たないことが多かった/1…まったく役に立たなかった］である．

各質問に対する回答結果を**図 7** に示し，これらについて 3 つの観点から考察を行う．

7.2.1 楽曲の可視化手法

まず，楽器編成を円グラフで表し，2 次元平面上に表示する可視化手法が，楽曲探索に利用できたかについて考察する．質問 a の結果 4.79 より，楽曲ごとの楽器編成の違いを円グラフによって把握できていたことが確認できる．したがって，楽器編成を円グラフで表す可視化手法は，楽器編成の類似性に基づいて楽曲を探索するのに役立つと考えられる．一方，質問 a の結果 4.79 と比較して，質問 b の結果 3.83 は低い値となっている．このことから，円グラフの違いから楽曲どうしの違いを把握することはできても，位置や外見から各楽曲の内容を想像することまでは相対的に困難であったことが分かる．

これに関連して，複数の被験者から，似た色の楽器が見分けにくかったという感想が得られた．このことから，色による可視化方法に改善の余地があると考えられる．たとえば，色の違いだけでなく，模様や楽器の絵などの情報を加えることで，より判別しやすく，想像しやすい可視化を行うことで，好みの楽曲をより容易に探索することができる可能性がある．

7.2.2 楽器編成を用いた検索機能

次に，楽器編成を用いた検索機能が，楽曲探索において有用であるかについて考察する．質問 c の結果 4.71 より，楽器編成をクエリとする楽曲検索を実際に行った被験者が，その機能を便利だと感じたことが分かる．INSTRUDIVE 以外のインタフェースにはそうした楽曲検索機能は備わっておらず，今回の予備評価実験で被験者は初めて利用していたことから，楽器編成を用いた楽曲探索における楽曲検索の有用性が裏付けられたと考えられる．

7.2.3 自動楽器認識の精度による影響

最後に，4 章で提案した CNN モデルの認識精度による，楽曲探索への影響について考察する．質問 d の結果 4.83 より，それぞれの円グラフは楽器編成に基づく楽曲探索に役立ったことが確認できる．質問 a から，円グラフによって楽器編成の違いが可視化できたことが分かっているため，可視化のために用いられた各楽曲の楽器編成の自動認識は，楽曲探索に役立つほどの精度に達しており，楽曲探索に応用可能であると考えられる．

自由記述の質問では，認識精度に対する印象として，十分および不十分の双方の意見が得られた．ある被験者からは，検索機能や楽器編成表示の精度が想像よりも高かった，という感想が得られた．一方，別の被験者からは，鳴っていない楽器が認識されることが多く，精度が不十分だと感じた，という感想が得られた．これは，4 章の実験結果において，F-macro が提案手法でも 0.655 であったことから，楽曲によって認識精度にばらつきが生じていたのが原因である．

以上の一連の実験結果と考察により，INSTRUDIVE のインタフェースによって，実際に楽器編成に基づく楽曲探索が可能になったことが実証された．今後，より多くの人々にとって有用な楽曲探索を実現するためには，さらなる研究開発によって自動楽器認識の精度向上が必要である．

8. おわりに

本論文では，従来の楽曲探索では十分に扱われていなかった楽器編成に注目し，その円グラフによる新たなアイコン表現や，楽曲集合全体を可視化できる楽器編成マップ，楽曲内での楽器編成の時間変化を可視化できるビジュアルプレイヤーなどの豊富な機能を搭載した楽曲探索システム INSTRUDIVE について述べた．深層学習に基づく楽器推定の性能を向上する手法を提案し，楽器カテゴリごとの出現頻度の偏りも考慮した評価実験をした結果，最新の手法を含む 3 種類の既存楽器認識手法よりも，F-macro と AUC で特に高い性能を持つことを確認した．これは，データ数の少ない楽器カテゴリに対しても高い精度で認識ができたことを意味する．INSTRUDIVE のインタフェースをウェブアプリケーションとして手軽に試せる形で実装し，24 名の被験者が使用する予備評価実験を実施した結果，実験者が立ち会っていない状況下においても，楽器編成に基づく楽曲探索が実現できていたことを確認できた．

本研究で対象としたような多数の楽器音が混ざった混合音中の各楽器音の認識は，その技術的な難易度が高いために，本研究も含む最先端の手法を用いても，性能がけっして高いとはいえない未成熟な段階にある．そのため，1 章でも指摘したように，楽曲探索のインタフェースを構築す

る際に，自動認識した楽器編成を用いるための基礎的検討が不十分な状況であった．しかし，音楽情報処理分野の学術的な発展のためには，楽器音認識の性能向上を追求し，その性能が十分高くなるのを待つだけではなく，その技術が成熟した未来においてどのような利活用が可能になるのかを見越して，具体的な応用例としてのインタフェースを研究開発することは重要である．それによりインタフェース側からの需要を示すことは，過去に十分な研究開発がなされてきたとはいえない楽器認識問題の重要性を顕在化し，さらなる研究開発を促すことにもつながる．本研究はそのような問題意識から，性能向上のための具体的な手法を示す貢献に加えて，それを効果的に用いた新機能を搭載した従来にない楽曲探索インタフェースを提案する貢献をした．インタフェースの各機能を網羅的に評価するのは本論文の範囲を超えた今後の課題となるが，本論文の予備検討実験において，楽器編成がなければ不可能な楽曲探索が実現できていた点は，楽器編成の活用事例が乏しい現時点では意義がある．まだ商用音楽配信サービスのような大規模な楽曲群を扱える段階ではないが，今後の研究では，本論文の提案内容がそのような場面でどう活用できるのかを探求していく予定である．

謝辞 本研究の一部は JST ACCEL（JPMJAC1602）の支援を受けた．

参考文献

[1] Whitman, B. and Lawrence, S.: Inferring descriptions and similarity for music from community metadata, *Proc. 2002 International Computer Music Conference*, pp.591–598 (2002).

[2] Berenzweig, A., Ellis, D.P.W. and Lawrence, S.: Anchor space for classification and similarity measurement of music, *Proc. 2003 International Conference on Multimedia and Expo, ICME '03*, Vol.1, pp.I–29 (2003).

[3] Schafer, J.B., Frankowski, D., Herlocker, J. and Sen, S.: *Collaborative Filtering Recommender Systems*, pp.291–324, Springer Berlin Heidelberg (2007).

[4] Cohen, W.W. and Fan, W.: Web-collaborative filtering: recommending music by crawling the web, *Computer Networks*, Vol.33, No.1, pp.685–698 (2000).

[5] Sánchez-Moreno, D., González, A.B.G., Vicente, M.D.M., Batista, V.F.L. and García, M.N.M.: A collaborative filtering method for music recommendation using playing coefficients for artists and users, *Expert Syst. Appl.*, Vol.66, pp.234–244 (2016).

[6] Oramas, S., Nieto, O., Barbieri, F. and Serra, X.: Multi-label music genre classification from audio, text and images using deep features, *Proc. 18th International Society for Music Information Retrieval Conference* (*ISMIR 2017*), pp.23–30 (2017).

[7] Kim, T., Lee, J. and Nam, J.: Sample-level CNN architectures for music auto-tagging using raw waveforms, *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*IEEE ICASSP 2018*), pp.366–370 (2018).

[8] Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Instrogram: Probabilistic representation of instrument existence for polyphonic music, *IPSJ Journal*, Vol.2, No.1, pp.279–291 (2007).

[9] Bittner, R.M., Salamon, J., Tierney, M., Mauch, M., Cannam, C. and Bello, J.P.: MedleyDB: A multitrack dataset for annotation-intensive MIR research, *Proc. 15th International Society for Music Information Retrieval Conference* (*ISMIR 2014*), pp.155–160 (2014).

[10] Takahashi, T., Fukayama, S. and Goto, M.: Instrudive: A music visualization system based on automatically recognized instrumentation, *Proc. 19th International Society for Music Information Retrieval Conference* (*ISMIR 2018*), pp.561–568 (2018).

[11] Martin, K.D.: *Sound-Source Recognition: A Theory and Computational Model*, PhD thesis, Massachusetts Institute of Technology (1999).

[12] Eronen, A.J.: Musical instrument recognition using ICA-based transform of features and discriminatively trained hmms, *Seventh International Symposium on Signal Processing and Its Applications* (*ISSPA 2003*), pp.133–136 (2003).

[13] Essid, S., Richard, G. and David, B.: Musical instrument recognition by pairwise classification strategies, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.14, No.4, pp.1401–1412 (2006).

[14] Simmermacher, C., Deng, D. and Cranefield, S.: Feature analysis and classification of classical musical instruments: An empirical study, *Advances in Data Mining, Applications in Medicine, Web Mining, Marketing, Image and Signal Mining, 6th Industrial Conference on Data Mining* (*ICDM 2006*), pp.444–458 (2006).

[15] Joder, C., Essid, S. and Richard, G.: Temporal integration for audio classification with application to musical instrument classification, *IEEE Trans. Audio, Speech & Language Processing*, Vol.17, No.1, pp.174–186 (2009).

[16] Hamel, P., Wood, S. and Eck, D.: Automatic identification of instrument classes in polyphonic and poly-instrument audio, *Proc. 10th International Society for Music Information Retrieval Conference* (*ISMIR 2009*), pp.399–404 (2009).

[17] Yu, G. and Slotine, J.-J.: Audio classification from time-frequency texture, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing* (*IEEE ICASSP 2014*), pp.1677–1680 (2009).

[18] Lostanlen, V. and Cella, C.-E.: Deep convolutional networks on the pitch spiral for music instrument recognition, *Proc. 17th International Society for Music Information Retrieval Conference* (*ISMIR 2016*), pp.612–618 (2016).

[19] Eggink, J. and Brown, G.J.: Instrument recognition in accompanied sonatas and concertos, *Proc. IEEE International Conference on Acoustics, Speech, and Signal* (*IEEE ICASSP 2004*), pp.217–220 (2004).

[20] Bosch, J.J., Janer, J., Fuhrmann, F. and Herrera, P.: A comparison of sound segregation techniques for predominant instrument recognition in musical audio signals, *Proc. 13th International Society for Music Information Retrieval Conference* (*ISMIR 2012*), pp.559–564 (2012).

[21] Han, Y., Kim, J. and Lee, K.: Deep convolutional neural networks for predominant instrument recognition in polyphonic music, *IEEE/ACM Trans. Audio, Speech, and Language Processing*, Vol.25, No.1, pp.208–221 (2017).

[22] Pons, J., Slizovskaia, O., Gong, R., Gómez, E. and Serra, X.: Timbre analysis of music audio signals with con-

volutional neural networks, *Proc. 25th European Signal Processing Conference* (*EUSIPCO 2017*), pp.2744–2748 (2017).
[23] Little, D. and Pardo, B.: Learning musical instruments from mixtures of audio with weak labels, *Proc. 9th International Conference on Music Information Retrieval* (*ISMIR 2008*), pp.127–132 (2008).
[24] Heittola, T., Klapuri, A. and Virtanen, T.: Musical instrument recognition in polyphonic audio using source-filter model for sound separation, *Proc. 10th International Society for Music Information Retrieval Conference* (*ISMIR 2009*), pp.327–332 (2009).
[25] Barbedo, J.G.A. and Tzanetakis, G.: Musical instrument classification using individual partials, *IEEE Trans. Audio, Speech & Language Processing*, Vol.19, No.1, pp.111–122 (2011).
[26] Kobayashi, Y.: Automatic generation of musical instrument detector by using evolutionary learning method, *Proc. 10th International Society for Music Information Retrieval Conference* (*ISMIR 2009*), pp.93–98 (2009).
[27] Fuhrmann, F. and Herrera, P.: Polyphonic instrument recognition for exploring semantic similarities in music, *Proc. International Conference on Digital Audio Effects* (*DAFx 2010*) (2010).
[28] Li, P., Qian, J. and Wang, T.: Automatic instrument recognition in polyphonic music using convolutional neural networks, arXiv preprint arXiv:1511.05520 (2015).
[29] Hung, Y.-N. and Yang, Y.H.: Frame-level instrument recognition by timbre and pitch, *Proc. 19th International Society for Music Information Retrieval Conference* (*ISMIR 2018*), pp.135–142 (2018).
[30] Thickstun, J., Harchaoui, Z. and Kakade, S.M.: Learning features of music from scratch, *Proc. International Conference on Learning Representations* (*ICLR 2017*) (2017).
[31] Humphrey, E., Durand, S. and McFee, B.: OpenMIC-2018: An open data-set for multiple instrument recognition, *Proc. 19th International Society for Music Information Retrieval Conference* (*ISMIR 2018*), pp.438–444 (2018).
[32] Pampalk, E., Dixon, S. and Widmer, G.: Exploring music collections by browsing different views, *Proc. 4th International Conference on Music Information Retrieval* (*ISMIR 2003*) (2003).
[33] Torrens, M., Hertzog, P. and Arcos, J.-L.: Visualizing and exploring personal music libraries. *Proc. 5th International Conference on Music Information Retrieval* (*ISMIR 2004*) (2004).
[34] Pampalk, E. and Goto, M.: MusicRainbow: A new user interface to discover artists using audio-based similarity and web-based labeling, *Proc. 7th International Conference on Music Information Retrieval* (*ISMIR 2006*), pp.367–370 (2006).
[35] Lamere, P. and Eck, D.: Using 3D visualizations to explore and discover music, *Proc. 8th International Conference on Music Information Retrieval* (*ISMIR 2007*), pp.173–174 (2007).
[36] Hamasaki, M. and Goto, M.: Songrium: A music browsing assistance service based on visualization of massive open collaboration within music content creation community, *Proc. 9th International Symposium on Open Collaboration* (*ACM WikiSym + OpenSym 2013*), pp.1–10 (2013).
[37] Yoshii, K. and Goto, M.: Music Thumbnailer: Visualizing musical pieces in thumbnail images based on acoustic features, *Proc. 9th International Conference on Music Information Retrieval* (*ISMIR 2008*), pp.211–216 (2008).
[38] Goto, M. and Goto, T.: Musicream: Integrated music-listening interface for active, flexible, and unexpected encounters with musical pieces, *IPSJ Journal*, Vol.50, No.12, pp.2923–2936 (2009).
[39] Foote, J.: Visualizing music and audio using self-similarity, *Proc. 7th ACM International Conference on Multimedia* (*ACM Multimedia 1999*), pp.77–80 (1999).
[40] Cooper, M. and Foote, J.: Automatic music summarization via similarity analysis. *Proc. 3rd International Conference on Music Information Retrieval* (*ISMIR 2002*) (2002).
[41] Goto, M.: A chorus section detection method for musical audio signals and its application to a music listening station, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.14, No.5, pp.1783–1794 (2006).
[42] Müller, M. and Jiang, N.: A scape plot representation for visualizing repetitive structures of music recordings, *Proc. 13th International Society for Music Information Retrieval Conference* (*ISMIR 2012*), pp.97–102 (2012).
[43] Goto, M., Yoshii, K., Fujihara, H., Mauch, M. and Nakano, T.: Songle: A web service for active music listening improved by user contributions, *Proc. 12th International Society for Music Information Retrieval Conference* (*ISMIR 2011*), pp.311–316 (2011).
[44] Jeong, D. and Nam, J.: Visualizing music in its entirety using acoustic features: Music flowgram, *Proc. International Conference on Technologies for Music Notation and Representation*, pp.25–32 (2016).
[45] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167 (2015).
[46] He, K., Zhang, X., Ren, S. and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, arXiv preprint arXiv:1502.01852 (2015).
[47] Sechidis, K., Tsoumakas, G. and Vlahava, I.: On the stratification of multi-label data, *Machine Learning and Knowledge Discovery in Databases*, pp.145–158 (2011).
[48] Peeters, G.: A large set of audio features for sound description (similarity and classification) in the CUIDADO project, Technical Report, IRCAM (2004).

高橋 卓見

2017 年京都工芸繊維大学工芸科学部設計工学域情報工学課程卒業．2019 年筑波大学大学院システム情報工学研究科修士課程修了．現在，産業技術総合研究所契約職員．

深山 覚 （正会員）

2013 年東京大学大学院情報理工学系研究科システム情報学専攻博士課程修了．博士（情報理工学）．日本学術振興会特別研究員（DC2），産業技術総合研究所研究員を経て，2017 年より同研究所主任研究員．専門は音楽情報科学．2009 年度情報処理学会山下記念研究賞受賞．情報処理学会音楽情報科学研究会計算論的生成音楽学（GMI）ワーキンググループ 2 代目代表．

後藤 真孝 （正会員）

1998 年早稲田大学大学院理工学研究科博士後期課程修了．博士（工学）．現在，産業技術総合研究所首席研究員兼メディアコンテンツ生態系プロジェクトユニット代表．2016 年から JST ACT-I「情報と未来」研究総括，2017 年から日本学術会議連携会員等を兼任．JST ACCEL 研究代表者．日本学士院学術奨励賞，日本学術振興会賞，ドコモ・モバイル・サイエンス賞基礎科学部門優秀賞，FIT 船井業績賞，科学技術分野の文部科学大臣表彰若手科学者賞，星雲賞等，51 件受賞．