Vol.2016-MUS-111 No.33
2016/5/21

演奏楽器の時間構造情報に基づく楽曲検索と DNN を用いた楽器演奏区間検出

金田　響[†1*]　岩野公司[†1]

概要：近年，楽曲検索システムにおいて多様な検索クエリが利用されているが，演奏楽器の登場順や演奏割合といった時間構造情報をクエリとして利用することで，より利便性の高い楽曲検索が可能になると考える．本研究では，まず，このようなクエリを利用した楽曲検索のシミュレーションを 30 名の被験者で主観評価することによって，検索の有用性を示す．このような検索の実現のためには，登場楽器の演奏区間情報を高精度に検出する必要がある．そこで次に，DNN（Deep Neural Network）を用いた楽器演奏区間検出手法の提案と性能評価を行った．提案手法には，音声認識において高い性能が確認されている DNN-HMM 法を用いる．この手法では，検出対象となる楽器音を隠れマルコフモデル（HMM）でモデル化した上で，各状態の出力確率が DNN によって推定される．楽器演奏区間を同定は，入力楽曲に対して最尤となるモデル系列を探索することで行われる．約 12 時間の楽器音データを学習に，約 10 分の単一楽器による演奏楽曲を評価に利用した実験を行ったところ, 提案手法により約 9 割の検出性能が達成された．

キーワード：楽曲検索，演奏楽器情報，Deep Neural Network，楽器演奏区間検出

Music retrieval based on time structure information of musical instruments and musical instrument activity detection using Deep Neural Network

HIBIKI KANEDA[†1, †2]　KOJI IWANO[†2]

Abstract: Recently, various search queries are utilized in music retrieval systems. We propose a useful music retrieval using queries based on the time structure information of musical instruments, i.e., appearance order/ratio of the musical instruments. In this research, we first show the usefulness of the proposed music retrieval by subjective evaluations (with thirty subjects) of a simulation system. In order to realize the retrieval system, a method for high precision extraction of musical instrument information, namely “musical instrument activity detection”, is needed. Therefore, we next propose a DNN(Deep Neural Network)-based musical instrument activity detection method and evaluate the performance. The proposed method is based on the DNN-HMM framework, which achieves high performance in speech recognition tasks. In this framework, acoustic models of musical instruments are represented by HMMs (Hidden Markov Models) and the output probabilities of the states are calculated by the DNN. The detection is conducted by searching for the best acoustic model sequence yielding the maximum likelihood for input music data. Experiments were conducted using twelve hours of training data and ten minutes of evaluation data. The results show that the proposed method achieves approximately 90% detection performance.

Keywords: Music Retrieval, Musical Instrument Information, Deep Neural Network, Musical Instrument Activity Detection

1. はじめに

近年，楽曲検索システムの実用化が急速に進んでおり，歌詞やメロディの一部をクエリとした楽曲検索サービスや，音楽をスマートフォンに直接聞かせることで楽曲検索を行うアプリケーションなどが登場している．

一方で,「演奏楽器」の情報を積極的に利用する楽曲検索システムの研究・開発例はそれほど多くない．文献[1]では演奏に使われている楽器の種類（構成）の推定とそれに基づく楽曲検索について論じており，その有効性を示している．楽器の種類といった情報だけでなく，楽器の登場順や演奏の割合といった，さらに詳細な「楽器の時間構造情報」をクエリとして利用することができれば,「冒頭はドラムで，その後ギターが入る曲」「ピアノソロがメインで，ときどきドラムが入る曲」などといった,「こだわった」クエリによる検索が可能となる．このような楽器情報をクエリとした検索は，場の雰囲気にあった楽器構成の BGM を見つけようとするときや，楽器の演奏者が演奏楽曲を選択するといった場面で，有効に活用できるものと考える．しかし，このような検索のニーズをどれだけのユーザが持っているかについては不明確である．そこで本研究では，まず，提案する楽曲検索の有用性を被験者実験によって検証する．

また，このような楽曲検索を実現するためには，検索対象となる楽曲に対し，楽器演奏区間の時間情報を高精度に自動検出する必要がある．そこで本研究では，多階層のニューラルネットワークである Deep Neural Network（DNN）を利用した楽器演奏区間検出手法を提案し，その検出性能の評価を行う．具体的には，近年音声認識分野で高い性能が報告されている DNN-HMM 法[2]を利用することで，高精度な楽器演奏区間検出の実現を目指す．

†1 東京都市大学
Tokyo City University
* 現在，日本電気株式会社

ⓒ2016 Information Processing Society of Japan

2. 楽器情報をクエリとする検索の有用性検証

提案する「楽器の登場順や演奏割合といった時間構造情報をクエリとした楽曲検索」のニーズを検討するため，この検索の有用性を被験者実験によって評価する．そこで，楽曲検索のシミュレーションを行う Web システムを試作し，複数の被験者に使用してもらうことで主観評価実験を実施した．これらのクエリは，特に楽器演奏者にとってメリットがあることが想定されるため，評価結果が「被験者の楽器演奏経験の有無」の影響を受けるかどうかについても考察を行う．

2.1 シミュレーションシステムの構築

以下の 9 種類のクエリ（楽器の登場順に関するもの 6 種類，演奏割合に関するもの 3 種類）を利用することができる楽曲検索のシミュレーションシステムを Web システムとして構築した．

【楽器の登場順に関するクエリ】

- ドラムから演奏が始まり，次にピアノが登場する曲
- ギターから演奏が始まり，次にボーカルが登場する曲
- ピアノから演奏が始まり，次にドラムが登場する曲
- ピアノから演奏が始まり，次にストリングスが登場する曲
- ストリングスから演奏が始まり，次にピアノが登場する曲
- ボーカルから演奏が始まり，次にギターが登場する曲

【楽器の演奏割合に関するクエリ】

- ギターが他の楽器よりも長く演奏される曲
- ピアノがメインで，ドラムも登場する曲
- ストリングスがメインで，他の楽器も登場する曲

検索対象の楽曲は人手で収集しており，システムには J-POP を中心に合計で 24 曲の情報が登録されている．各クエリを選択すると，該当する複数の楽曲の情報がタイトルと共に表示され，結果の楽曲を試聴することができる．

2.2 評価方法

本学学生 30 名を対象に被験者実験を行った．そのうち，楽器演奏経験者は 20 名，未経験者は 10 名であった．被験者には，システムによる検索を体験してもらった上で，次の 2 項目について 5 段階（1: 全くそう思わない～5:とてもそう思う）で評価してもらった．

- 「楽器の登場順に関するクエリ」による楽曲検索は有用であると思うか
- 「楽器の演奏割合に関するクエリ」による楽曲検索は有用であると思うか

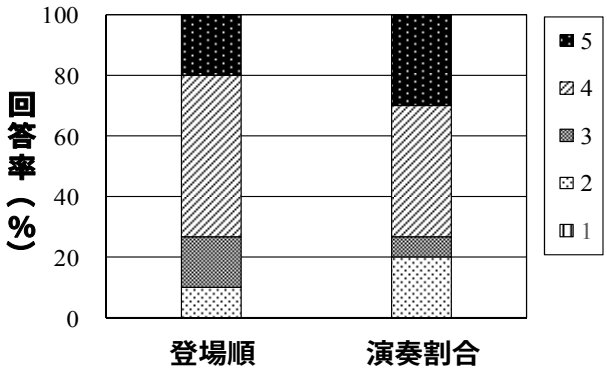

図 1　楽器の時間構造情報をクエリに用いた楽曲検索の有用性の 5 段階評価結果

図 1 に，「登場順」「演奏割合」それぞれのクエリの検索に対する各評価値の回答率を示す．双方のクエリともに，4 点以上の評価をつけた被験者の割合が 7 割を超え，平均評価値は 3.8 という高い評価となったことから，これらのクエリによる検索の有用性が示唆された．なお，クエリ間での評価の傾向の違いは特に見られなかった．

次に，楽器の演奏経験者と未経験者の評価の違いについて考察する．両者の評価結果の比較を，楽器の登場順，楽器の演奏割合のクエリの評価ごとに，それぞれ図 2, 3 に示す．有意水準 5%で平均評価値の差を検定したところ，どちらのクエリに対しても，演奏経験者と未経験者の間で有意差は確認されなかった．したがって，これらのクエリによる検索が経験者にとって特に有用という傾向は確認されず，幅広いユーザに有益である可能性が示された．

3. DNN-HMM 法による楽器演奏区間検出

2 章において，楽器の時間構造情報をクエリとした検索の有用性が示されたことから，本章では，その検索の実現に必要となる,楽器演奏区間の自動検出手法の提案を行う．高精度な区間検出を目指し，近年音声認識で高い性能が報告されている DNN-HMM 法[2]の枠組みを利用する．

図 4 に，提案手法の流れを示す．まず，入力音源から音響特徴量をフレーム単位で抽出する．次に，時間方向に連続した複数フレームの音響特徴量を結合して高次元ベクトルを作成し, それを事前に学習した DNN に入力する. DNN の出力ノードは，各楽器と無音区間をモデル化した隠れマルコフモデル（HMM）の各状態に対応づけられており，DNN の出力値は HMM の各状態における出力確率となる．この楽器音モデル（DNN-HMM）と検出対象楽器を登録した辞書，登録楽器が等確率で出現することを記述した文法を利用し, 音声認識で一般に利用される Viterbi 探索の枠組みで，入力音源に対する最尤楽器音系列を推定する．その結果中から検出対象楽器に割り当てられた区間を，その楽器の演奏区間として検出する．したがって，複数楽器の演奏重複区間に対しても，最尤となる単一の楽器が割り当て

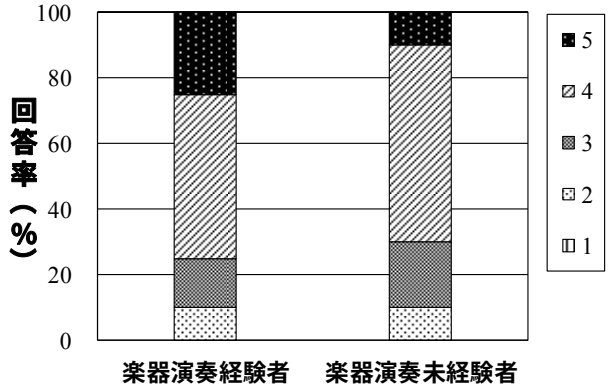

図 2　登場順のクエリに対する演奏経験者と未経験者の有用性の評価結果の比較

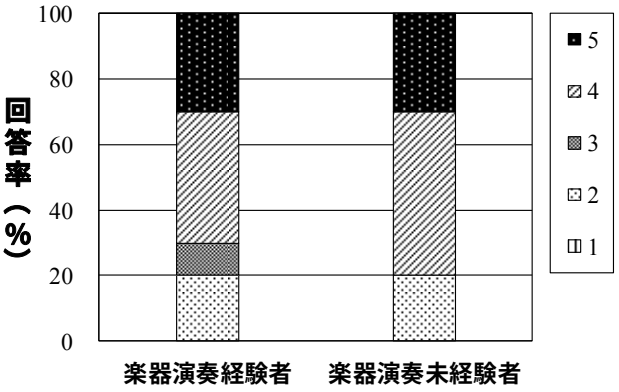

図 3　演奏割合のクエリに対する演奏経験者と未経験者の有用性の評価結果の比較

られることになる．

DNN の学習は，教師なしの事前学習（pre-training）と教師ありの再学習（fine-tuning）の 2 段階で実施する．まず，事前学習において，制約つきボルツマンマシン（RBM: Restricted Boltzmann Machine）を一層ずつ学習する．出来上がった RBM を積み重ね，そこに乱数で初期化したソフトマックス層を追加することで, DNN の初期ネットワークを構成する[3]．得られた初期ネットワークに対し，確率的勾配降下法（STD: Stochastic Gradient Descent）による学習を行うことで最終的な DNN を得る．再学習のためには，学習用の音源データに対する，フレームごとの正解ラベルが必要となる．大量の学習用データに対して人手でラベルを付与することは現実的ではないことから，本研究では，事前に学習データを用いて各楽器音のモデルを混合正規分布（GMM）に基づく HMM（GMM-HMM）で構築し，そのモデルを用いた強制切り出し（Forced Alignment）を学習データに対して実行することで，正解ラベルを作成する．

4. 楽器演奏区間検出の性能評価

4.1 使用データ

本研究では，検出対象楽器をドラム，ギター，ピアノ，

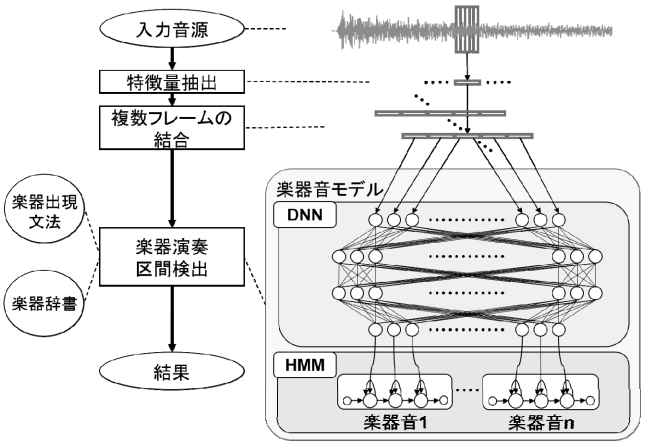

図 4　提案手法に基づく楽器演奏区間検出の流れ

ストリングス，ボーカルの 5 種類とした．

学習データには「RWC 研究用音楽データベース：楽器音[4]」から選択した音源（約 12 時間分）を利用する．これらは単一楽器による演奏で，複数楽器による演奏重複区間は存在しない．なお学習時には，これらの音源を全て単音に分割したデータを利用する．評価データには，同データベースから選曲した単一楽器の演奏楽曲と，J-POP や Rock に分類される一般の音楽 CD から選曲した楽曲の 2 種類を使用する．前者は約 10 分（1 楽器あたり約 2 分），後者は 5 楽曲の冒頭 15 秒（計 75 秒）であり，いずれも学習には使用されていない．

4.2 実験の内容と条件

従来までの音声認識の枠組みでは，音響特徴量にメル周波数ケプストラム係数（MFCC）を，音響モデルに GMM-HMM を利用することが標準的であった．そこで，この枠組みを演奏区間検出に適用したものを本研究ではベースライン（以下, MFCC による GMM-HMM 法）とする．

その上で，以下に示す 2 種類の提案手法の性能を評価する実験を行う．

1. 音響特徴量にベースラインと同じ MFCC を用い，音響モデルに DNN-HMM を用いる手法（以下，MFCC による DNN-HMM 法）．ベースラインとの性能比較により，音響モデルへの DNN の導入効果（識別的学習の効果）を検証することができる．
2. 音響特徴量として高次元のフィルタバンク対数パワー（FBANK）を用い，音響モデルに DNN-HMM を用いる手法（以下，FBANK による DNN-HMM 法）．MFCC は，その前段階で抽出される高次元の FBANK に対して各種正規化処理を行って次元を削減した，音声認識に特化した特徴量といえる．したがって，2. の性能を 1. の性能と比較することにより，DNN が持つ特徴量抽出器としての性能を評価することができ，DNN によって楽器音の認識により適した特徴量の抽出が実現できているかを検証することができる．

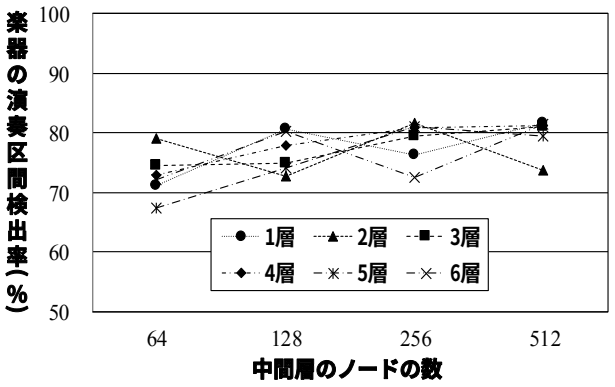

図 5　提案手法による単一楽器楽曲の演奏区間検出性能
（MFCC による DNN-HMM 法）

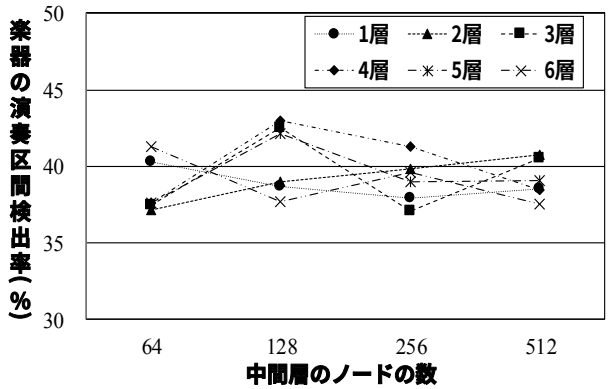

図 6　提案手法による一般楽曲の演奏区間検出性能
（MFCC による DNN-HMM 法）

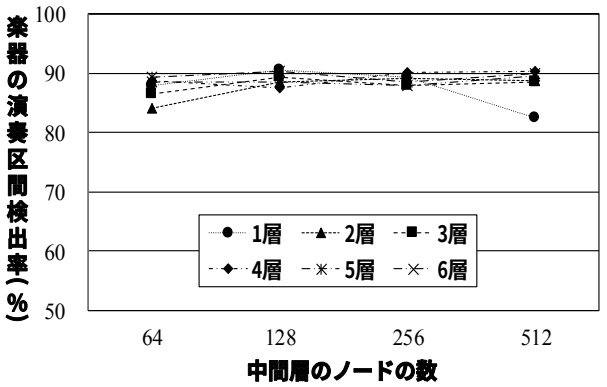

図 7　提案手法による単一楽器楽曲の演奏区間検出性能
（FBANK による DNN-HMM 法）

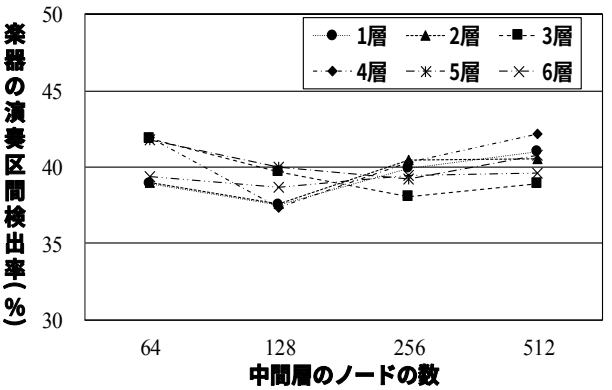

図 8　提案手法による一般楽曲の演奏区間検出性能
（FBANK による DNN-HMM 法）

MFCC, FBANK 共に，特徴量抽出のフレーム間隔は 10ms，フレーム幅は 25ms とした．実際の音響特徴量としては，ベースラインの MFCC による GMM-HMM 法では，「12MFCC＋Δ12MFCC＋Δ対数パワー」の 25 次元ベクトル，MFCC による DNN-HMM 法では，それに「対数パワー」を加えた 26 次元ベクトルとした．また，FBANK による DNN-HMM 法では「24 次元 FBANK＋Δ24 次元 FBANK＋対数パワー＋Δ対数パワー」の 50 次元ベクトルを利用した．DNN には前後 2 フレーム（合計 5 フレーム）の特徴ベクトルを結合した高次元ベクトルを入力する．したがって，MFCC による DNN-HMM 法では 130 次元，FBANK による DNN-HMM 法では 250 次元の特徴ベクトルが DNN に入力される．

なお，HMM は 3 状態とし，GMM-HMM の正規分布の混合数は予備実験で評価データに対する性能が最高となった 32 に設定した．また，性能評価は楽器演奏区間のフレーム単位の検出率（F 値）で行う．

4.3 中間層の層数・ノード数の変化に対する性能調査

性能比較に先立ち，提案手法（DNN-HMM 法）において，DNN の中間層の層数とノード数を変化させたときの性能評価を行い，評価データ（単一楽器楽曲と一般楽曲）に対する楽器演奏区間検出性能が良好となる条件を調査しておく．本実験では，中間層の層数は 1～6 層，ノード数は 64, 128, 256, 512 と変化させて性能評価を行った．

MFCC による DNN-HMM 法の単一楽器楽曲と一般楽曲に対する検出性能の変化の様子をそれぞれ図 5, 6 に示す．横軸は中間層のノード数，縦軸は楽器の演奏区間検出率（F 値）であり，各線が層数の違いを表している．図 5 の単一楽器楽曲の結果を見ると，中間層の層数を 3～5 とし，ノード数を 256, 512 といった大きい値に設定すると安定して良好な性能が得られていることがわかる．一方，図 6 の一般楽曲の結果を見ると，層数，ノード数の変化に対して性能がばらついており，ノード数 128 で局所的に最高性能が得られているが，多くの条件において 40%を下回る性能となっていることがわかる．

次に，FBANK による DNN-HMM 法の単一楽器楽曲と一般楽曲に対する検出性能の変化の様子をそれぞれ図 7, 8 に示す．図 7 の単一楽器楽曲の結果を見ると，層数，ノード数の変化に対する性能変化は全体的に小さく，層数が十分大きければ，ノード数 128 以上で安定して良好な性能が得られていることがわかる．図 8 の一般楽曲の結果を見ると，ノード数 512 のときに比較的性能が良好になり，ほとんどの層数の条件に対して 40%を超える検出性能となった．

ⓒ2016 Information Processing Society of Japan

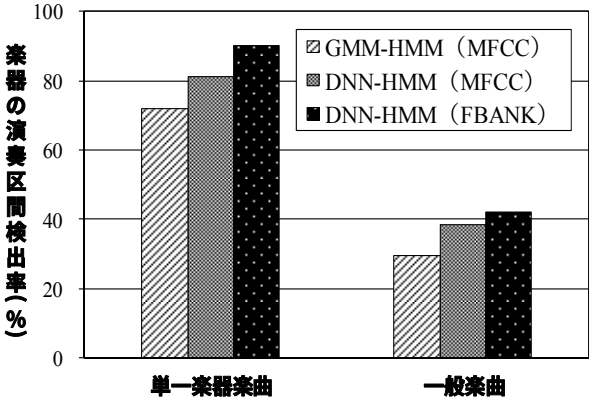

図 9　楽器演奏区間検出性能の比較

4.4 検出性能の比較

ベースラインである MFCC による GMM-HMM 法と，提案手法である MFCC, FBANK による DNN-HMM 法の３つの楽器演奏区間検出手法の性能比較を行う．なお，DNN-HMM 法における中間層の層数とノード数は，4.3 節の実験結果から，単一楽器楽曲と一般楽曲の双方に対して比較的良好な性能を示す条件を選ぶこととし,結果として，MFCC, FBANK の両方ともに「4 層，512 ノード」の結果を採用することとした．

図 9 に，単一楽器楽曲と一般楽曲に対する楽器演奏区間検出性能の比較結果を示す．MFCC による GMM-HMM 法（ベースライン）と MFCC による DNN-HMM 法の性能を比較すると，識別的学習で構築された DNN を利用することで，両楽曲ともに約 10%の性能改善が得られることがわかる．また，MFCC による DNN-HMM 法と FBANK による DNN-HMM 法を比較した場合，後者の方が高い検出性能となった．これは音声認識で標準的な特徴量である MFCC を利用するよりも，それよりも前段階で抽出される高次元特徴量 FBANK から DNN によって特徴抽出を行う方が楽器音の認識に有効であることを示唆している．結果として，提案手法の検出性能は単一楽器楽曲で 90.3%に到達した．しかし，一般楽曲における最高検出性能は 42.2%にとどまり，複数楽器による演奏重複への対応が必要であることが明らかになった．

4.5 楽器ごとの検出性能の検証

提案手法 DNN-HMM 法における楽器ごとの演奏区間検出性能を検証する．図 10 に単一楽器楽曲，図 11 に一般楽曲に対する検出結果を，使用した音響特徴量ごとに示す．

図 10 の単一楽器楽曲に対する実験結果を見ると，ボーカル（音声）については MFCC も FBANK と同等の検出性能を有しているのに対し，他の楽器音では FBANK を用いることで検出性能が有意に改善していることがわかる．このことから，音声区間の検出については MFCC で十分対応できたが，楽器音の検出に対しては MFCC の性能は不十分であり，DNN の導入により楽器音向けの特徴抽出が実現され，検出性能が改善したものと考えられる．

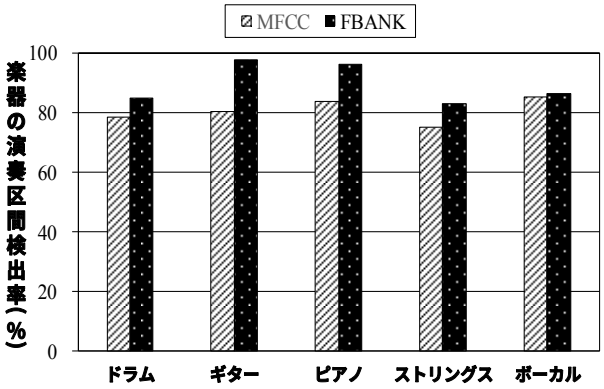

図 10　提案手法による検出対象楽器別の検出性能
（単一楽器楽曲）

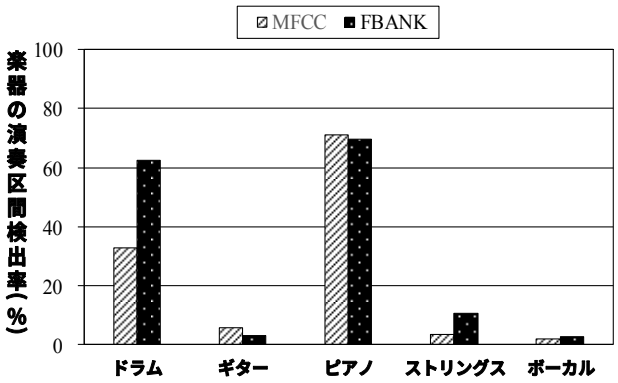

図 11　提案手法による検出対象楽器別の検出性能
（一般楽曲）

一方，図 11 の一般楽曲に対する実験結果を見ると，楽器ごとの検出性能に大きなばらつきが見られ，FBANK を利用した場合でも，ギター，ストリングス，ボーカルなどでは検出性能が非常に低い値に留まっていることがわかる．これらの楽器は比較的演奏重複の影響を受けやすいと考えられることから，検出性能の改善のためには，演奏重複区間に対する直接的な対処手法の導入が望まれる．

5. おわりに

本研究では，まず，楽曲検索に対して楽器の登場順や演奏割合といった「時間構造情報」をクエリとして利用することを提案し，その検索の有用性を，シミュレーションシステムを用いた被験者実験によって検証した．その結果，登場順，演奏割合に関するクエリともに高い有用性があることが確認された．次に，このような検索の実現に向け，DNN-HMM 法を利用した楽器演奏区間検出手法の提案を行い，その性能評価を行った．その結果，単一楽器による演奏楽曲では約 9 割という高い検出性能を示したが，複数楽器の演奏重複が存在する一般的な楽曲に対しては約 4 割という性能にとどまった．

ⓒ2016 Information Processing Society of Japan

今後は, 演奏重複区間への対処による性能改善に向けて, 演奏重複区間が存在する学習データの利用や，複数楽器の同時検出手法の検討, DNN のトポロジーの再考などによる区間検出の性能改善などに取り組む必要がある．そのうえで，検出性能が改善された提案手法を実装した，最終的な楽曲検索システムを構築することが強く望まれる．

謝辞　本研究の一部は JSPS 科研費 基盤研究（B）25280058 の助成を受けたものです．

参考文献

[1] 戸谷他，“楽器構成に着目した楽曲サムネイルとプレイリスト生成機能つき音楽プレイヤー，” 情報処理学会シンポジウム論文集，vol. 2008, no.4, pp.173-174, 2008.
[2] G.E. Dahl, et al., “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” IEEE Trans. Audio, Speech, & Language Proc., vol. 20, no. 1, pp. 30-42, 2012.
[3] 三村他，“CSJ を用いた日本語講演音声認識への DNN-HMM の適用と話者適応の検討，” 情報処理学会研究報告，2013-SLP-97, vol.2013, no.9, pp.1-6, 2013.
[4] 後藤他，“RWC 研究用音楽データベース：音楽ジャンルデータベースと楽器音データベース，” 情報処理学会研究報告，2002-MUS-45-4, vol.2002, no.103, pp.19-26, 2002.

ⓒ2016 Information Processing Society of Japan