5Q-04

Deep Neural Network を用いた楽器演奏区間検出

金田　響†　　岩野　公司†
東京都市大学†

１．はじめに

近年，楽曲検索のニーズが高まり，歌詞やメロディの一部をクエリとした検索サービスも登場している．しかし，「演奏楽器」の情報を積極的に利用した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，楽器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．更に詳細な楽器の時間構造情報（楽器の登場順や演奏のタイミングなど）を導入できれば，ユーザは「冒頭はドラムがメインで，その後ギターが入る曲」といった，より「こだわった」クエリによる検索が可能となる．このような検索の実現のためには対象の楽曲から，各楽器の演奏区間の時間情報を自動的かつ高精度に検出する必要がある．

そこで本研究では，高精度な楽器演奏区間検出手法の実現を目指し，多階層のニューラルネットワークである Deep Neural Network（DNN）を利用した手法を提案する．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]を利用し，高精度な検出を試みる．

２．DNN-HMM 法による楽器演奏区間検出

図 1 に，提案する楽器演奏区間検出の流れを示す．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前後複数フレームの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．この 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，それぞれの状態の出力確率となる．このモデル（DNN-HMM）と検出対象楽器を登録した辞書，演奏中の楽器出現のルール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決定する．その結果中の，検出

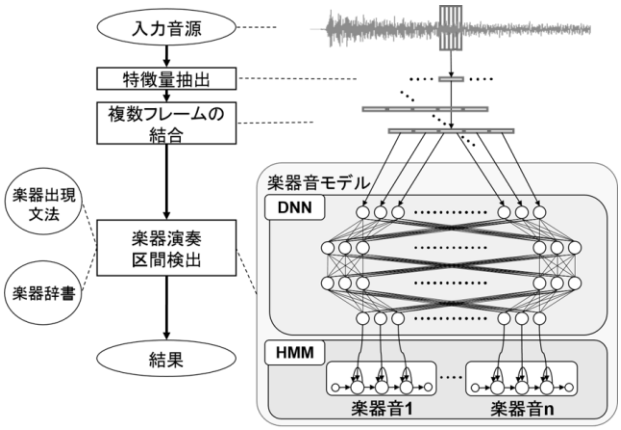

図 1　楽器演奏区間検出の流れ

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られることになる．

なお，DNN の学習のためには学習用音源に対する楽器演奏区間の時間ラベルが必要となるが，それは混合正規分布（GMM）に基づく HMM（GMM-HMM）で構築した標準的な楽器音モデルを用いた強制切り出しによって作成する．

３．楽器演奏区間検出の性能評価

3.1　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，ストリングス，ボーカルの 5 種とした．

学習には「RWC 研究用音楽データベース：楽器音[3]」から選択した音源（約 12 時間分）を利用した．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は存在しない．評価用には同データベースから選んだ単一楽器の演奏楽曲と，音楽 CD から選曲した一般楽曲の 2 種を使用する．前者は 15 音源（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，学習には使用されていない．

3.2　実験内容と条件

従来まで音声認識では，音響モデルに GMM-HMM を用いて（GMM-HMM 法），音響特徴量としてメル周波数ケプストラム係数（MFCC）を

Musical instrument activity detection using deep neural network
Hibiki Kaneda†, Koji Iwano†, †Tokyo City University

利用することが標準であった．本実験では，この手法をベースラインとする．まず，MFCC を用いた DNN-HMM 法の性能をベースラインと比較し，DNN の識別性能を検証する．MFCC は，その前段階で抽出される高次元の「フィルタバンク対数パワー（FBANK）」に対し，各種の正規化処理を行って次元を削減した，音声認識用の特徴量である．そこで次に，DNN が持つ特徴抽出の効果を検証するため，MFCC と FBANK を用いた DNN-HMM 法の検出性能を比較する．

特徴量抽出のフレーム間隔は 10ms，フレーム幅は 25ms とした．実際の音響特徴量は，GMM-HMM 法で MFCC を用いる場合には，「12MFCC＋12ΔMFCC＋Δ 対数パワー」の 25 次元ベクトル，DNN-HMM で MFCC を用いる場合には，それに「対数パワー」を加えた 26 次元ベクトルとした．FBANK を利用する場合には「24FBANK＋24ΔFBANK＋対数パワー＋Δ 対数パワー」の 50 次元ベクトルとした．DNN への入力は，前後 2 フレームの特徴量を結合した高次元ベクトル（MFCC: 130 次元，FBANK: 250 次元）とした．

なお，HMM は 3 状態であり，GMM-HMM の正規分布の混合数は予備実験で最適化を行って 32 とした．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．

3.3 DNN の層数・ノード数の最適化

提案手法については，中間層の層数を 1 から 6，ノード数を 64，128，256，512 と変化させて性能評価を行い，単一楽器楽曲，一般楽曲の双方で性能が良好となる値を最終評価に利用する．図 2 に，FBANK を用いた DNN-HMM 法における，層数とノード数を変化させたときの検出性能を示す（単一楽器楽曲）．この手法については，一般楽曲の結果も考慮して，4 層，512 ノードを最適値として採用した．

3.4 実験結果

図 3 に，単一楽器楽曲と一般楽曲に対する演奏区間検出結果を示す．特徴量に MFCC を利用した GMM-HMM 法と DNN-HMM 法の性能を比較すると，DNN を利用することで両楽曲ともに約 10%の性能改善がみられる．また，FBANK を用いることでさらに検出性能が向上することから，DNN による特徴抽出効果も確認することができる．今回，最高検出性能は単一楽器楽曲で 90.3%となったが，一般楽曲では性能は 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わかった．

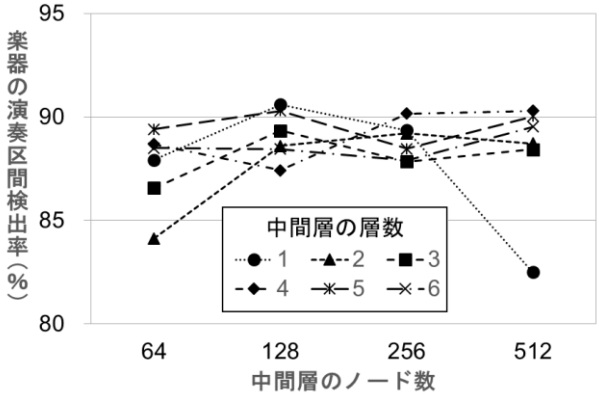

図 2　単一楽器楽曲の演奏区間検出結果

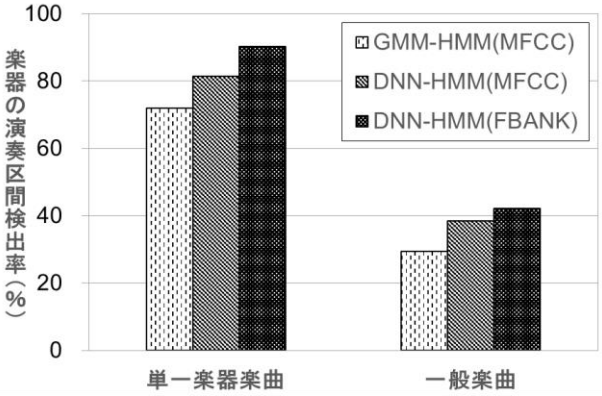

図 3　楽器演奏区間検出の性能比較

4．まとめ

本研究では， 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，単一楽器楽曲では約 9 割の検出性能を得たが，一般の楽曲では性能は約 4 割にとどまった．今後は，演奏重複区間への対処による性能改善を図るとともに，提案手法を実装した楽曲検索システムの構築を目指す．

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] 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.