E-028

深層学習による入力音響信号からの MIDI 音色パラメータ推定

MIDI tone parameter value estimation of the input instrument audio signal based on deep learning

瀧田 寿明† 櫛部 義之† 星野 准一† 矢澤 櫻子† 浜中 雅俊‡
Toshiaki Takita Yoshiyuki Kushibe Junichi Hoshino Sakurako Yazawa Masatoshi Hamanaka

1. はじめに

本研究では，深層学習の音楽情報処理研究への利用について検討する．深層学習のモデルの一つである Deep Belief Networks(DBN)に基づき，混合音中の対象楽器音を MIDI 音源のパラメータ調整により再現するシステムの構築を目指すことを試みる[1]． 本稿では，その第一段階として，単一楽器音の音響特徴を MIDI 音源で再現する事を試みる．

近年，深層学習(Deep Learning)が画像認識や音声認識などの分野で高い性能を示している．深層学習は深い層を持ったフィードフォワード型ニューラルネットワーク(NN)を実現する手法の総称で，多次元入力ベクトルから汎用性の高い高次特徴量を識別することが可能である．そこで我々は，入力音響信号に似通った音を出力するよう MIDI 音源のパラメータを深層学習 のモデルの一つである DBN を用いて学習することを試みる．このような学習が可能となれば，音楽音響信号を MIDI 音源などの特徴パラメータで表現することが可能となり，ある曲で用いた音に似通った音を MIDI 音源で再現して他の曲で再利用することが可能となる．

従来，音源分離の研究において，分離後の音を MIDI 音で再現することでノイズや歪みを含まない音を出力する事が試みられていたが，MIDI 音源のパラメータと音響信号との関係を重回帰分析により表現していたため，離散的なパラメータについては適切に学習することが困難であった[2].

本稿では，MIDI 音源のパラメータと，その MIDI パラメータの時に出力される音響信号との関係を学習する際，離散的な MIDI パラメータについても学習が可能となるよう DBN を用いて学習を行う．これにより，DBN が適切な学習を行うことができれば，オシレータの種類の切り替えなど離散的な未知パラメータがあった場合でも，入力音響信号に対して MIDI 音源で似通った音を再現可能となる．予備実験では適切な DBN の中間層構成と，教師なし学習回数を検討し，構築したモデルによって入力音響信号から MIDI 音源のパラメータ推定を行い，文献[2]と性能を比較した．

2. DBN を用いた解析モデル

解析モデルは，MIDI 音源のパラメータと音響特徴量との関係を DBN を用いて学習するモードと学習済みの DBM を用いて，音響信号から MIDI 音源のパラメータを推定するモードの 2 つからなる（図 1）．

† 筑波大学
‡ 京都大学

2.1 MIDI 音色パラメータ推定

入力音響信号から抽出した特徴量と，その時の MIDI 音源のパラメータの組を教師データとして DBN によって学習を行う(図 1b)．特徴量は，文献[2]を参考に MIDI で生成した音から窓長 0.025 秒，シフト長を 0.02 秒ごとに抽出した 12 種 32 次元のベクトルとする．パラメータの推定時には，入力音響信号から抽出した音響特徴量ベクトルを，学習済みの DBN に入力し，出力値を得る．DBN の出力値を MIDI 音源にパラメータとして設定すると，元の MIDI 音が出力される(図 1a).

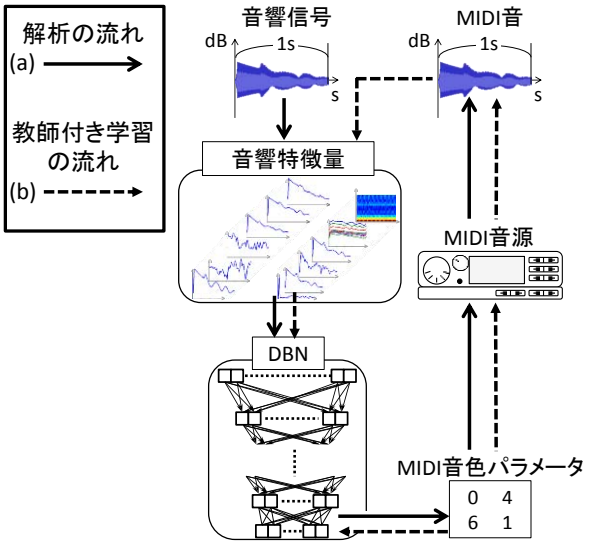

図 1. DBN を用いた解析モデル

2.2 Deep Belief Networks の構成

我々が構成した DBN の$i+1$層におけるデータL_{i+1}は，i層のデータL_iによって (1)式で表される．

$$L_{i+1} = sigmoid(W_i L_i + b_i) \quad (1)$$

$$sigmoid(x) = \frac{1}{1+e^{-x}} \quad (2)$$

W_iはi番目の層と$i+1$番目の層間の重み付け行列であり，b_iはバイアスである．フィードフォワード型 NN は階層的な構造になっていて，層数を増やすほど，複雑なモデルを識別出来るが，層数を増やすほど局所解に陥りやすいという問題がある．この問題の原因は NN の層数を増やすと，教師信号と DBN 出力値の差が入力層側に伝わらないため，誤差逆伝播法による学習が適切に行えない事である．従来から DBN では層数の多いフィードフォワード型 NN に，事前に Restricted Boltzman Machine (RBM) を用いて教師なし学習を行う事で，誤差逆伝播法による学習を実現しており，我々もその手法を用いる．

3. 実験結果

本節では，まず MIDI 音源のパラメータと，その時に MIDI 音源から出力される音響信号の特徴量との関係を学習するのに適した DBN の設計について検討し，次に，構築したモデルに MIDI 音色パラメータの推定性能の評価を行う．

3.1 実験条件

実験には，MIDI 音色パラメータのうち我々が指定した 4 つのパラメータをランダムに変更したときの MIDI 音源の出力音を用いた．文献[2]と結果と比較を可能とするために以下の 5 つの実験条件を設定する．

① パラメータ 4 種をランダムに変更し，MIDI 音源で 100 の学習データ，10 のテストデータを作成する．
② ファミシンセⅡ[3]という MIDI 音源を用いる．
③ 解析対象の MIDI 音は，中心周波数 440Hz，音長は 1 秒のものに限定する．
④ パラメータごとに正規化する．
⑤ DBN の出力値は離散的でないので閾値を設定し，最も近いパラメータ値を推定値とする．

今回は比較的低い表現力でも表現可能だと予想される音長に関するパラメータの DecayTime を対象に評価した．評価方法として文献[2]を参考に誤差 e を(3)式のように定義した．

$$e = \frac{\sum_i | p_{est,i} - p_{ref,i} |}{P} \quad (3)$$

Pは推定するパラメータ数，$p_{est,i}$は推定されたパラメータ，$p_{ref,i}$は正解のパラメータである．

3.2 中間層の構成と学習回数の検討

DBN の中間層の構成と教師なし学習回数を検討するために 2 つの予備実験を行った．

まず，中間層構成の検討を行った．ニューロン数がそれぞれ 700，300，100 の中間層を持つ DBN と，ニューロン数がそれぞれ 300，100，60 の中間層を持つ DBN を比較した．これら 2 つの DBN に教師なし学習を 50,000 回ずつ行い，教師付き学習を 20,000 回行ったところ表 1 の結果が得られた．ニューロン数がそれぞれ 700，300，100 の中間層構成の方がテストデータに対する実験結果の誤差 e が小さいので，より性能が高いと判断した．

次に，教師なし学習回数の検討を行った．ニューロン数がそれぞれ 700，300，100 の中間層を持つ DBN に，0，1,000，5,000，50,000 回ずつ教師なし学習を行い，教師付き学習を 20,000 回行ったところ表 2 の結果を得た．教師なし学習 5,000 回の DBN がテストデータに対する実験結果の誤差 e が最も小さいので，より性能が高いと判断した．

以上から，3.3 節の実験ではニューロン数がそれぞれ 700，300，100 の中間層を持つ DBN に教師なし学習を 5,000 回行う．

表 1．DBN 中間層構成の比較評価

DBN 中間層構成	300，60，20	700，300，100
テストデータに対する実験結果の誤差 e	0.588	0.175

表 2．教師なし学習回数の比較評価

教師なし学習回数[回]	0 回	1,000 回	5,000 回	50,000 回
テストデータに対する実験結果の誤差 e	0.188	0.588	0.038	0.175

3.3 推定性能の評価

DBN に対して 366,000 回教師付き学習を行ったところ学習データに対するパラメータ正解率 98%に達したので，教師付き学習が十分に行われたとみなし，テストデータへの性能を評価した．テストデータに対するパラメータ正解率は，学習回数 172,000 回の時に 80%となり最大であった．学習回数 172,000 回以降もテストデータに対するパラメータ正解率は増減したが，最大値を更新せず，過学習であった．

テストデータに対するパラメータ正解率が最大値をとった時と教師付き学習終了時の，テストデータに対する実験結果の誤差 e を評価した(表 3)．テストデータの誤差 e は最小で 0.025 となった．文献[2]の実験では，学習データ 900 で誤差 e は 0.037 であったので，文献[2]よりも未知のパラメータが多く，少ない学習データで，同等以上の性能を発揮するモデルを実現した．

表 3．DBN の過学習

教師付き学習回数[回]	172,000 回	366,000 回
テストデータに対するパラメータ正解率[%]	80%	50%
テストデータに対する実験結果の誤差 e	0.025	0.075

4. まとめ

本研究では深層学習モデルの一つの DBN によって MIDI パラメータを MIDI 音源から出力した音から逆推定し，誤差 0.025 の推定結果を得た．これは従来の重回帰分析を用いた手法[2]の性能を上回るものであった．我々は，この事は音楽情報処理研究に対して DBN が有効である可能性を示す一例であると考えている．

今後，混合音響信号中の対象楽器音を MIDI 音で再現するシステム の実現や過学習を防止する手法について研究を行っていく．

参考文献

[1]Y. Bengio, P.Lamblin, D. Popovici and H. Larochlle, “Greedy Layer-Wise Training of Deep Networks.” Neural Information Processing Sysytems Foundation, 2006

[2]糸山克寿，奥野博，”楽器音に対する仮想音源のパラメータ推定”，情報処理学会研究報告，Vol.2013-MUS-100, No.5

[3]ファミシンセ II (FAMISYNTH-II)@mu-station VST LABO http://www.geocities.jp/mu_station/vstlabo/famisynth.html