論文 / 著書情報
Article / Book Information

論題	全層ゲート付き2次元畳み込みネットワークによる多重音信号の音高認識
著者	生田目 敬弘, 亀岡 弘和, 篠田 浩一
出典	研究報告音声言語情報処理（SLP）, vol. 120, no. 12, pp. 1-7
発行日	2018, 2
権利情報	本著作物の著作権は情報処理学会に帰属します。本著作物は著作権者である情報処理学会の許可のもとに掲載するものです。ご利用に当たっては「著作権法」ならびに「情報処理学会倫理綱領」に従うことをお願いいたします。

全層ゲート付き2次元畳み込みネットワークによる多重音信号の音高認識

生田目 敬弘[1,a)]　亀岡 弘和[2,b)]　篠田 浩一[1,c)]

概要：音楽は音高方向（和音構成，調波構造）と時間方向（旋律，リズム）の 2 次元構造を有する．我々は，音楽音響信号の音高認識の問題を音響スペクトログラムに対する音高ラベルの 2 次元的な配置問題と捉え，多重音信号の対数周波数スペクトログラムから直接音高認識を行う全層ゲート付き 2 次元畳み込みネットワークを提案する．全層がゲート付き 2 次元畳み込みネットワークで構成され，楽音の音響スペクトログラムと音楽の 2 次元構造を各層で表現する．従来の確率的潜在成分分析手法と比較し，Bach10 データセットにおいて従来手法の音符単位 F1 スコア 65.0％を 8.3％ポイント上回る 73.3％の性能を得た．また，室内楽データセットを新たに構築し，モデルの学習に用いた．

1. はじめに

近年，音楽情報処理は急速な発展を遂げており，中でも複数の楽音から楽譜へ変換する自動採譜システムは，音楽検索システムや著作権管理において重要な役割を果たすことが期待されている．自動採譜の問題は音高推定，楽器種推定，リズム・拍推定，テンポ推定，調推定などの要素問題からなるが，本稿ではその中でも複数楽器による多重音の音高推定問題に焦点を当てる．

楽音は音高に対応する基本周波数の成分以外に倍音と呼ばれる周波数成分を含み，その成分比は楽器種により異なる．楽音スペクトルは基本周波数にピークを持つとは限らず，従って単音を対象とした音高推定であっても楽音のスペクトル全体を手がかりにする必要がある．多重音信号を対象とする場合は，複数の楽音が混在したスペクトルの各成分がどの楽音に由来するかに関する情報が欠落するため，より困難な問題となる．

多重音を対象とした音高推定のアプローチとして，これまで非負値行列因子分解（Non-negative Matrix Factorization; NMF） [14] に基づく手法が提案されている．このアプローチには教師あり NMF[7], [17], [22] や，NMF と同形の手法である確率的潜在成分分析（Probabilistic Latent Component Analysis; PLCA）[20] に基づく手法 [1] が含まれる．Benetos らは PLCA の生成モデルに隠れマルコフモデルを導入することで楽音スペクトルの時間依存性を捉えることを可能にしている [1]．この手法は，Bach10 データセット [5] と TRIOS データセット [6] に対する音高推定において最も高い性能を示している．ただし，NMF アプローチは各音高に対応する楽音スペクトルのテンプレートを用いてスペクトログラムを各楽音スペクトログラムに分解する手法であるため，楽音テンプレートが実際の楽音スペクトルと合致する場合には高い性能を示す一方で，合致しない場合に脆弱であるという難点がある．

これに対し，ニューラルネットワーク（Neural Network; NN）がもつ識別能力を活かし，スペクトルから直接的に音高を推定するプロセスを NN によりモデル化した深層学習アプローチが近年検討されている．多種多様な楽音を用いて学習を行うことで個体差や奏法の違いに対し頑健な音高推定器を得られる可能性がある．これまで NN を用いたアプローチとしては，再帰型ニューラルネットワーク（Recurrent NN; RNN）を用いたピアノ音を対象とした音高推定手法 [3] や，楽器種と音高の同時推定手法 [24] が提案されている．

RNN は出現音高の時間依存性を捉えることを可能にするが，実際の音楽は，旋律やリズムのような時間方向のパターンだけでなく，和音構成や調波構造などの音高・周波数方向のパターンが存在し，2 次元の構造を持つ．従って，時間方向の依存性だけでなく和音構成の規則や周波数方向のスペクトル構造も同時に捉えられる 2 次元畳み込みネットワーク（Convolutional NN; CNN）が音高推定タスクに

[1] 東京工業大学
Tokyo Institute of Technology, Japan
[2] 日本電信電話株式会社 NTT コミュニケーション基礎研究所
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
[a)] namatame@ks.cs.titech.ac.jp
[b)] kameoka.hirokazu@lab.ntt.co.jp
[c)] shinoda@c.titech.ac.jp

おいて有効になりうる．実際，CNN は既に単一楽器多重音信号の音高推定 [8], [12], [19] や単一楽音信号の楽器種推定 [15] など，音楽情報処理の様々なタスクに適用されている．また，CNN を多楽器多重音信号の音高推定 [2] に適用した方法も提案されている．ただし，通常 CNN で広範囲の構造を捉えた識別を行うためには多層化が必須であるが，単純な CNN では多層化に伴って勾配消失が生じやすくなることが知られている．

そこで我々は，通常の CNN に比して勾配消失を生じにくい特長を持つゲート付き CNN （Gated CNN; GCNN） [4] を導入した音高推定法を提案する．GCNN は Dauphin らにより提案され，入力文章における後続単語を予測する言語モデルとしての能力が長・短期記憶（Long Term-Short Memory; LSTM）を凌駕することが報告されている．GCNN は，LSTM ネットワークと同様に線形出力を変調させる GLU（Gated Linear Unit）と呼ぶゲート構造を畳み込み層の活性化関数に導入することにより各層で通過させたい情報の制御を可能にしつつ勾配消失を防ぐことができる特長がある．従来の GCNN は時系列データの時間方向のモデリングに用いられていたが，我々は音楽の 2 次元構造を捉えることを可能にするため GCNN を 2 次元に拡張する．

深層学習手法は一般に，モデルを学習するための多量のデータを要する．一方，音楽のデータは著作権の問題やラベル生成コストが高いことから研究者間で共有される利用可能なデータが少ない．特に多楽器による多重音データは極めて少量で，多楽器の多重音音高推定に深層学習手法を適用した例は，我々の知る限り，Bittner ら [2] の 1 例に限られる．

そこで我々は，新たに 3 楽器による室内楽曲のデータセットを構築した．データセットは総曲数 54 曲，総曲長 108 分の三重奏室内楽曲で構成され，楽器ごとの音響信号と人手で付与された楽器別の音高ラベルが含まれ，音高推定システムの評価実験に用いることができる．

2. 全層ゲート付き 2 次元畳み込みネットワークを用いた音高認識

音楽音響信号のスペクトログラムを $\mathbf{X} \in \mathbb{R}^{F \times T}$（$F$ を周波数ビン数，T をフレーム数）とする．本稿では，$\mathbf{X}$ を入力とし，各時刻における各音高の生起確率を表した値を要素にもつ $\mathbf{Y} \in \mathbb{R}^{88 \times T}$ （88 はピアノの鍵盤数に相当）を出力する NN として，第 $l+1$ 層の出力 $\mathbf{H}_{l+1}$ が

$$\mathbf{H}_{l+1} = (\mathbf{W}_l * \mathbf{H}_l + \mathbf{b}_l) \odot \sigma(\mathbf{V}_l * \mathbf{H}_l + \mathbf{c}_l) \qquad (1)$$

で与えられるゲート付き畳み込み層を全層にもつ NN を考える．ただし，$\mathbf{H}_l \in \mathbb{R}^{D_l \times F_l \times T_l}$ は第 l 層の出力を表す．また，$\odot$ は要素ごとの積，$\sigma(\cdot)$ は要素ごとの標準シグモイド関数を表し，D_l および (F_l, T_l) は $\mathbf{H}_l$ のチャネル数とサイズ，$(\tilde{F}_l, \tilde{T}_l)$ は第 l 層のカーネル（あるいはフィルタ）のサイズを表す．ここで，すべての層の $\mathbf{W}_l \in \mathbb{R}^{D_{l+1} \times D_l \times \tilde{F}_l \times \tilde{T}_l}$, $\mathbf{b}_l \in \mathbb{R}^{D_{l+1}}$, $\mathbf{V}_l \in \mathbb{R}^{D_{l+1} \times D_l \times \tilde{F}_l \times \tilde{T}_l}$, $\mathbf{c}_l \in \mathbb{R}^{D_{l+1}}$ が学習すべきパラメータである．式 (1) を要素ごとに表記すると

$$h_{l+1,d,f,t} = \left(\sum_{d'=0}^{D_l-1} \sum_{f'=0}^{\tilde{F}_l-1} \sum_{t'=0}^{\tilde{T}_l-1} w_{l,d,d',f',t'} h_{l,d',f-f',t-t'} + b_{l,d} \right) \cdot \sigma \left(\sum_{d'=0}^{D_l-1} \sum_{f'=0}^{\tilde{F}_l-1} \sum_{t'=0}^{\tilde{T}_l-1} v_{l,d,d',f',t'} h_{l,d',f-f',t-t'} + c_{l,d} \right) \qquad (2)$$

となる．ここで，$\mathbf{H}_0$ が入力データに対応し，$\mathbf{H}_0 = \mathbf{X}$ である．畳み込み演算として Strided 畳み込みと Dilated 畳み込みのいずれかまたは両方を用いることができる．Strided 畳み込みはフィルタの畳み込みの適用間隔（ストライド幅と呼ぶ）を 1 以外にすることを許容した畳み込みで，ストライド幅が S のとき畳み込みの出力のサイズは入力サイズの $\frac{1}{S}$ 倍になる．よって S が 2 以上のときはダウンサンプリングの役割も担った畳み込みとなる．Dilated 畳み込みはパラメータを増やさずに受容野の範囲を大きくするよう適当なフィルタの係数を 0 に固定した畳み込みである．なお，各層の出力サイズは，入力 $\mathbf{H}_l$ に対して適当なゼロ埋めを行うことで $D_l \times F_l \times T_l$ となるよう調整可能である．ネットワークの出力 $\mathbf{Y}$ は

$$\mathbf{Y} = \sigma(\mathbf{H}_L) \qquad (3)$$

のようにシグモイド関数を適用し，各要素を区間 $[0, 1]$ に収まるようにすることで，本章冒頭で述べたように $\mathbf{Y}$ を各時刻における各音高の生起確率を表した値を要素にもつ行列と見なすことができる．

式 (1), (2) は 2 次元のゲート付き畳み込み層を記述したものであるが，1 次元版も含む表現となっている．1 次元 GCNN は，入力スペクトログラム $\mathbf{X}$ をチャンネル数が $D_0 = F$ でサイズが $1 \times T$ の画像，出力 $\mathbf{Y}$ をチャンネル数が $D_{L+1} = 88$ でサイズが $1 \times T$ の画像と見なす $(F_0 = \tilde{F}_l = 1)$ 場合に相当し，2 次元 GCNN は，入力 $\mathbf{X}$ をチャネル数が 1 でサイズが $F \times T$ の画像，出力 $\mathbf{Y}$ をチャネル数が $D_{L+1} = 1$ でサイズが $88 \times T$ の画像と見なす場合にそれぞれ相当する．

所与のスペクトログラムと音高ラベル行列（各時刻において各音高が存在するか否かを表したバイナリ行列）のペア $\{\mathbf{X}_j, \hat{\mathbf{Y}}_j\}_j$ を教師データとすることで以上の NN のパラメータ $\boldsymbol{\theta}$ を学習することができる．本稿では学習規準として交差エントロピー

$$\mathcal{J}(\boldsymbol{\theta}) = \sum_{f,t} \left\{ \hat{y}_{f,t} \log y_{f,t} + (1 - \hat{y}_{f,t}) \log(1 - y_{f,t}) \right\} \qquad (4)$$

を用いた．学習した NN パラメータを用いてテスト信号の

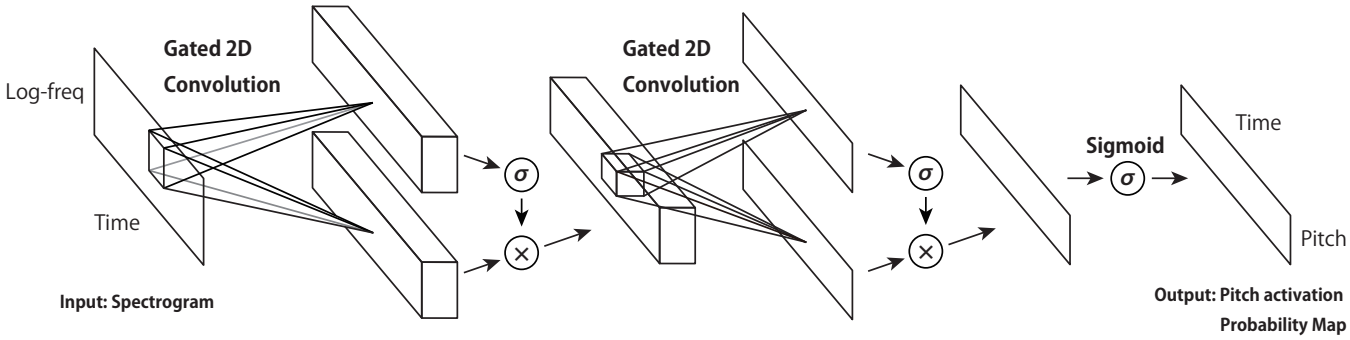

図 1 提案する全層ゲート付き 2 次元畳み込みネットワーク

音高推定を行う際は，$\mathbf{Y}$ の各要素をしきい値 τ により 2 値化したものを音高推定結果とする．

後処理として，従来手法 [1] と同様，時間方向に連続して検出された一連の音をまとめて 1 つの音符として扱い，長さが 80 ms に満たない音符を検出結果から取り除くこととした．これにより，音高推定結果に短い音符が含まれることを防げるが，一方で高速なフレーズを含む楽曲には適用できない問題が生じる．

3. 関連研究

音声音響特徴量の時系列のモデル化を目的としてゲート付き畳み込みネットワークが用いられている [11]．また，周波数方向の規則を捉える用途ではないが，音響イベント検知システムにおける音響特徴量の時系列のモデル化を目的としてゲート付き 2 次元畳み込み層が用いられている [23]．

4. 新規室内楽データセットの構築

提案モデルの学習に用いるため，室内楽曲を収録した代々木室内楽データセットを新たに構築した．代々木室内楽データセットは 5 曲，総曲長 108 分の三重奏楽曲で構成され，各楽曲はヴァイオリン，フルート，クラリネット，ファゴットのうち 3 楽器によって演奏された．演奏誤りの発生を抑制するため，演奏が止む部分で曲を分割し，各セグメントごとに収録を行った．また，データが似た音響信号になる曲の繰り返し部分について収録を行わなかった．各楽器は異なる部屋で同時に演奏され，個別のマイクで収録された信号を単純に加えることで 3 重奏楽曲を生成した．3 部屋を隔てる壁の一部がガラス窓になっており，演奏者らは互いの姿を視認できるほか，ヘッドホンを通して他の楽器の演奏を聞きながら演奏できるため，各演奏の同期性が確保されている．各楽曲には楽器別の音高ラベルが人手で付与されており，音高推定システムの評価実験に用いることができる．収録された楽曲の詳細は表 1 に示す通りである．各楽器の信号が独立な形で収録されているため，音高推定問題だけでなく，音源分離問題等にも用いることができる．このデータセットは音楽情報処理の研究者が利用できるよう公開を予定している．

5. 評価実験

5.1 使用データ

評価用に Bach10 データセット [5] ならびに TRIOS データセット [6] を用いた．Bach10 データセットはヴァイオリン，クラリネット，ファゴット，サックスによって演奏されたバッハ作曲の 4 声コラール 10 曲計 5 分で構成される．また，TRIOS データセットは三重奏の室内楽曲 5 曲計 3 分で構成される．楽曲はピアノ，ヴァイオリン，ヴィオラ，チェロ，クラリネット，ファゴット，トランペット，ホルン，サックスと幅広い楽器により演奏されており，これらの大半が訓練データに含まれていないことと，ピアノが 1 楽器で多重音を演奏するため，音高推定が他のデータセットより難しいことが特徴である．ドラムを含む楽曲 1 曲についてはドラム音を除いた上で実験に利用した．

提案モデルの学習に代々木室内楽データセットの楽曲 ‘mozart’ 及び ‘haydn’ を，ハイパーパラメータ探索時の検証に ‘huguenin’ を，評価に ‘vanhal’ を用いた．‘london’ は楽器構成が異なるため本実験では除外した．

5.2 実験条件

ネットワークへの入力特徴量として，1 オクターブあたりの周波数ビン数を 48（すなわち半音ごとのビン数は 4），最低周波数を 27.5 Hz，特徴次元数は 424 次元とし，フレーム幅は 10 ms として得られた CQT スペクトログラムを用いた．CQT スペクトログラムの生成には librosa ライブラリ [16] を用い，全層ゲート付き 2 次元畳み込みネットワークの実装には Chainer [21] を用いた．パラメータの学習手法として，Adam [13] を用いた．

国際的コンペティションである MIREX [9] に従い，フレーム単位と音符単位による 2 種類の評価を行った．フレーム単位の評価では，システムは各フレームでの音高を推定し，システムの推定音高と正解音高ラベルが一致した数を N_{TP}，総正解音高ラベル数を N_{ref}，システムが推定した音高の総数を N_{sys} として以下のように適合率（$\mathcal{P}$），再現率（$\mathcal{R}$）を定め，これらの調和平均である F1 スコア（$\mathcal{F}$）

表 1 代々木室内楽データセット収録楽曲一覧．表中の楽器構成において Vn. はヴァイオリン，Bn. はファゴット（バスーン），Cl. はクラリネット，Fl. はフルートを指す．

識別子	楽曲名	作曲者	楽器構成	データ数 [曲]	総曲長 [秒]
mozart	5 Divertimentos, K.Anh.229/439b	W. A. Mozart	Vn., Bn., Cl.	37	3718
huguenin	Trio for Oboe, Clarinet and Bassoon No.1, Op.30	C. Huguenin	Vn., Bn., Cl.	5	445
haydn	Keyboard Sonata in G major, Hob.XVI:40	F. J. Haydn	Vn., Bn., Cl.	7	842
vanhal	6 Trios, Op.10	J. B. Vanhal	Vn., Bn., Cl.	7	690
london	London Trios	F. J. Haydn	Vn., Bn., Fl.	10	763

表 2 実験に用いた全層ゲート付き 2 次元畳み込みネットワーク構造．上が入力側，下が出力側を表す．GC+BN は畳み込み演算後にバッチ正規化を行うゲート付き畳み込み層を表し，GC はバッチ正規化を行わないゲート付き畳み込み層を表す．次の項目は順にチャンネル数，重みパラメータテンソルのサイズを表す．S はストライド値を表す．

Gated 1D CNN	Gated 2D CNN
1D GC+BN: 256 × (424, 21)	2D GC+BN: 15 × (1, 200, 21), $S = 2$
1D GC+BN: 192 × (256, 21)	2D GC: 1 × (15, 100, 21), $S = 2$
1D GC: 88 × (192, 21)	Element-wise sigmoid
Element-wise sigmoid	

による評価を行った．

$$\mathcal{P} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{sys}}}, \mathcal{R} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{ref}}}, \mathcal{F} = \frac{2\mathcal{P}\mathcal{R}}{\mathcal{P}+\mathcal{R}} \qquad (5)$$

音符単位の評価では，推定した音符の音高と正解の音符の音高が同じでかつ立ち上がり時刻の差が 50 ms 以下である総音符数を N_{TP} とする．総正解音符数を N_{ref}，システムが推定した総音符数を N_{sys} と定め，フレーム単位の評価と同様に F1 スコアを計算する．評価に際して mir_eval ライブラリ [18] を用いた．

学習を停止するエポック数や後処理で用いるしきい値 τ 等のハイパーパラメータやネットワーク構造は代々木室内楽データセットの検証セット上で F1 スコアが最も高くなる組を選択した．選択したハイパーパラメータを表 2 に示す．また，検証の結果，以下の 2 点の工夫を行った．1 点目に，過学習を防ぐため最終層を除いた各層の畳み込み演算後にバッチ正規化 [10] を適用した．バッチ正規化を適用するタイミングは畳み込み演算後，シグモイド関数後，ゲート関数後の 3 通りが考えられ，実験から畳み込み演算後を選択した．2 点目に，入力行列へのゼロ埋めの方法を工夫した．時間方向には前後均等にゼロ埋めを行うが，周波数方向には，1 層目では調波構造を捉えるため高周波数方向にのみ，2 層目では和音構造を捉えるため高低均等にゼロ埋めを行うこととした．

5.3 実験結果

提案する全層ゲート付き 2 次元畳み込みネットワークの有効性を確認するため，1 次元版のゲート付き畳み込みネットワーク（以後 Gated 1D CNN），提案する全層ゲート付き 2 次元畳み込みネットワーク（以後 Gated 2D CNN），従来手法 [1] の音高推定性能を比較した．

訓練データと異なる環境で収録された Bach10 データセットでの音高推定実験を行い，その音符単位評価を表 3 にまとめた．この実験において，提案手法の音符単位 F1 スコアは従来手法の音符単位 F1 スコア 65.0%を 8.3%ポイント上回る 73.3%となり，これまで提案されてきた音高推定手法の中で最も高い音高推定性能を示した．

表 4，表 5 は，代々木室内楽データセットを用いたフレーム単位，音符単位の評価実験結果をそれぞれ表している．提案手法の音高推定性能は従来手法に比べ，フレーム単位 F1 スコアにおいて 13.0%ポイント向上し，音符単位 F1 スコアは従来手法に比べて 23.5%ポイント向上した．提案手法がこれほど高い音高推定性能を示した大きな要因として，訓練データと同じ楽器構成，同じ環境で演奏された楽曲で評価実験を行ったことが考えられる．

最後に，TRIOS データセットでの音高推定実験の結果を表 6 にまとめた．従来手法の音符単位 F1 スコア 59.4%に対し，提案手法の音符単位 F1 スコアは 2.6%下回った．この大きな要因として，TRIOS データセットの楽曲に含まれる楽音の大半が，訓練データにない楽器によって演奏されていることが考えられる．従来手法は TRIOS データセット中に含まれる全ての楽器について予め楽音テンプレートを学習している一方で，提案手法では未知の楽器であるため，TRIOS データベースの楽器構成が提案手法に不利に働いたと考えられる．この仮説を裏付けるように，適合率 $\mathcal{P}$ が従来手法より高く，再現率 $\mathcal{R}$ が従来手法よりも低いことがわかる．もしこの要因が大きいならば，今後利用可能なデータが増加していくにつれ，訓練データに多種多様な楽器，奏法，旋律，楽曲の構成などが含まれるようになるため，TRIOS データセットにおいても提案手法の音高推定性能が向上していくことが期待できる．

各実験における Gated 1D CNN と Gated 2D CNN の結

表 3　Bach10 データセットにおける音符単位 F1 スコア（%）

	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
PLCA [1]	65.0	57.4	**75.1**
Gated 1D CNN	47.9	50.5	45.6
Gated 2D CNN	**73.3**	**73.6**	73.0

表 4　代々木室内楽データセットにおける
フレーム単位 F1 スコア（%）

	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
PLCA [1] (Reimpl.)	76.2	75.0	77.3
Gated 1D CNN	80.3	84.1	76.9
Gated 2D CNN	**89.2**	**91.2**	**87.3**

表 5　代々木室内楽データセットにおける
音符単位 F1 スコア（%）

	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
PLCA [1] (Reimpl.)	60.6	52.1	72.5
Gated 1D CNN	70.7	74.8	67.0
Gated 2D CNN	**84.1**	**88.8**	**79.9**

表 6　TRIOS データセットにおける音符単位 F1 スコア（%）

	$\mathcal{F}$	$\mathcal{P}$	$\mathcal{R}$
PLCA [1]	**59.4**	60.2	**59.5**
Gated 1D CNN	30.6	47.7	22.5
Gated 2D CNN	56.8	**63.4**	51.4

果を比較すると，提案する Gated 2D CNN の音高推定結果が Gated 1D CNN の音高推定結果を全てのデータセットにおいて大きく上回っていることがわかる．この結果から，多重音の音高推定においては音楽が持つ時間方向の構造を捉えるだけでなく，音高・周波数方向の構造の活用も重要であることを示している．

従来手法と提案手法が推定した音高系列の例として，Bach10 データセット中の楽曲 '01-AchGottundHerr' の推定結果の最初の 10 秒分を図 2 に示した．3 秒付近や 7 秒付近の推定結果を比較すると，従来手法では音の検出個数は正しくも音高が誤っていたところが，提案手法では正しく音の高さを推定できており，多重音信号の音高推定で発生しがちな倍音誤りに強い手法になっていると考えられる．また，提案手法が誤った例として，音符を丸々検出に失敗している例が確認できる．これは，訓練データが三重奏楽曲のみで構成されるため入力が 3 重音以下であると捉えやすくなり，4 つ目の音を見逃しやすくなったと考えられる．

6. 結論と今後の課題

本稿では，全層ゲート付き 2 次元畳み込みネットワークによる多重音信号の音高認識手法を提案した．従来の 1 次元系列データのモデル化に用いられるゲート付き畳み込みネットワークを 2 次元へ拡張することで，音楽の周波数（音高）方向と時間方向の 2 次元構造を良く捉えることができる．

音高推定性能の評価実験を行った結果，Bach10 データセットにおいて，従来手法の音符単位 F1 スコア 65.0 %に対して提案手法の音符単位 F1 スコアは 8.3 %ポイント向上した 73.3 %を示し，より高精度な音高推定を行えることを確認した．

今後，利用可能な訓練データが増加し，多種多様な楽器，旋律，楽曲構成を網羅できるようになれば，提案手法の音高推定性能は向上すると期待できる．また，今後の課題として，生成モデルに基づく手法に関連して研究されてきた調波構造に関する制約を明示的に深層学習手法に取り入れる工夫や，大規模な楽譜データを活用する言語モデルの導入 [19] が挙げられる．また，楽器種の推定と音高の推定は密接に関係していることから，楽器種と音高を同時に推定する深層学習システムの考案を検討している．

参考文献

[1] Benetos, E. and Weyde, T.: An efficient temporally-constrained probabilistic model for multiple-instrument music transcription, *In Proc. of ISMIR* (2015).

[2] Bittner, R. M., Mcfee, B., Salamon, J., Li, P. and Bello, J.: Deep salience representations for F0 estimation in polyphonic music, *In Proc. of ISMIR* (2017).

[3] Böck, S. and Schedl, M.: Polyphonic piano note transcription with recurrent neural networks, *In Proc. of ISMIR*, pp. 612–618 (2016).

[4] Dauphin, Y. N., Fan, A., Auli, M. and Grangier, D.: Language modeling with gated convolutional networks, *In Proc. of ICML* (2016).

[5] Duan, Z., Pardo, B. and Chang, C.: Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions, *IEEE Transactions on Audio, Speech, and Language Processing* (2010).

[6] Fritsch, J.: High quality musical audio source separation (2012).

[7] Gao, L., Su, L., Yang, Y. H. and Lee, T.: Polyphonic piano note transcription with non-negative matrix factorization of differential spectrogram, *In Proc. of ICASSP*, IEEE, pp. 3112–3116 (2014).

[8] Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S. and Eck, D.: Onsets and frames: Dual-objective piano transcription, *arXiv preprint arXiv:1710.11153* (2017).

[9] IMIRSEL: Music Information Retrieval Evaluation eX-change (MIREX), `http://music-ir.org/mirex/` (2017).

[10] Ioffe, S. and Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, *In Proc. of ICML* (2015).

[11] Kaneko, T., Kameoka, H., Hiramatsu, K. and Kashino, K.: Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks, *Proceedings of the 18th Annual Conference of the International Speech Communication Association* (2017).

[12] Kelz, R., Dorfer, M., Korzeniowski, F., Böck, S., Arzt, A. and Widmer, G.: On the potential of simple framewise approaches to piano transcription, *In Proc. of ISMIR* (2016).

[13] Kingma, D. P. and Ba, J. L.: Adam: A method for

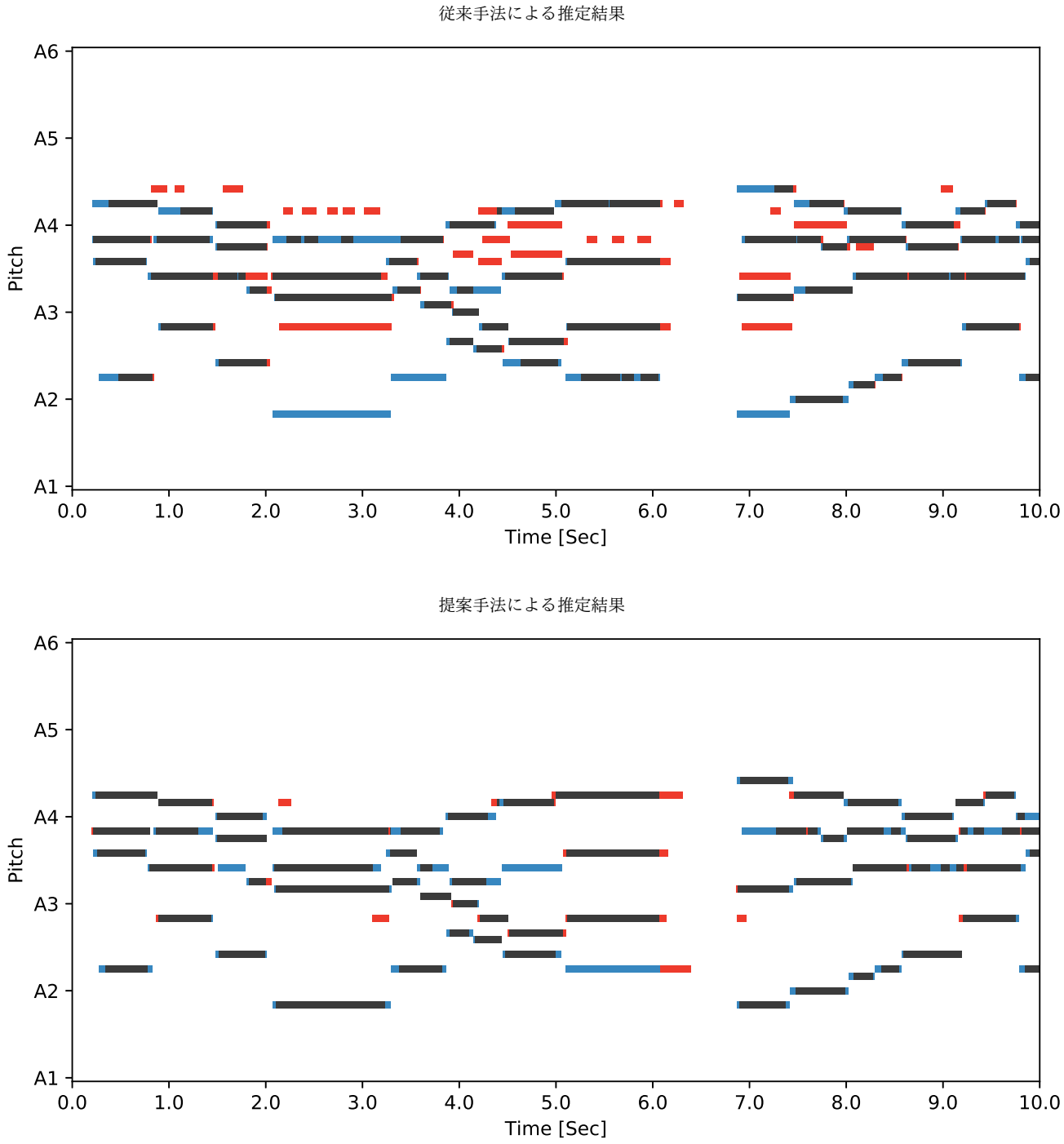

図 2 Bach10 データセット中の楽曲 '01-AchGottundHerr' より，最初の 10 秒間の各手法による音高推定結果ならびに対応する正解音高．音高 'A4' は 440 Hz の「ラ」に対応している．図中の黒い領域は音高推定システムが正解した音高，青い領域は検出しなかった音高，赤い領域は誤検出した音高を表している．

stochastic optimization, *In Proc. of ICLR* (2014).

[14] Lee, D. D. and Seung, H. S.: Learning the parts of objects by non-negative matrix factorization, *Nature* (199).

[15] Lostanlen, V. and Cella, C.-E.: Deep convolutional networks on the pitch spiral for music instrument recognition, *In Proc. of ISMIR*, pp. 612–618 (2016).

[16] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E. and Nieto, O.: librosa: Audio and music signal analysis in python, *In Proceedings of the 14th Python in Science Conference* (2015).

[17] O'Hanlon, K. and Plumbley, M. D.: Polyphonic piano transcription using non-negative matrix factorisation with group sparsity, *In Proc. of ICASSP*, IEEE, pp. 3112–3116 (2014).

[18] Raffel, C., McFee, B., Humphrey, E. J., Salamon, J., Nieto, O., Liang, D., Ellis, D. P. and Raffel, C. C.: mir_eval: A transparent implementation of common MIR metrics, *In Proc. of ISMIR*, Citeseer (2014).

[19] Sigtia, S., Benetos, E. and Dixon, S.: An end-to-end neural network for polyphonic music transcription, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 5 (2015).

[20] Smaragdis, P., Raj, B. and Shashanka, M.: A probabilistic latent variable model for acoustic modeling, *In Proc. of NIPS* (2006).

[21] Tokui, S., Oono, K., Hido, S. and Clayton, J.: Chainer: a next-Generation open source framework for deep learning, *In Proc. of Workshop on Machine Learning Systems in The Twenty-ninth Annual Conference on NIPS* (2015).

[22] Vincent, E., Bertin, N. and Badeau, R.: Adaptive harmonic spectral decomposition for multiple pitch estimation, *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 18, No. 3, pp. 528–537 (2010).

[23] Xu, Y., Kong, Q., Wang, W. and Plumbley, M. D.: Large-scale weakly supervised audio classification using gated convolutional neural network, *arXiv preprint arXiv:1710.00343* (2017).

[24] 生田目敬弘，亀岡弘和，篠田浩一：楽器と音高の同時認識のための RNN 音響モデル，第 111 回音楽情報科学研究会音学シンポジウム，Vol. 111, No. 46 (2016).