4T-08

Cycle-Consistency に基づく音楽音響信号の自動採譜

柴田 健太郎 † 錦見 亮 † 中村 栄太 † 吉井 和佳 †
† 京都大学 大学院情報学研究科 知能情報学専攻

1. はじめに

音楽音響信号に対する自動採譜では，従来，時間周波数スペクトログラムの低ランク性やスパース性に着目し，非負値行列因子分解（NMF）[1] や確率的潜在要素解析（PLCA）[2] などの音源の加法性に基づくスペクトログラムの生成モデルを用いるアプローチが主流であった．最近では，ニューラルネットワークを用いて，スペクトログラムから楽譜（ピアノロール）への写像を直接学習するアプローチ [3, 4] が主流となりつつある．しかし，十分な量の音響信号と楽譜のペアデータをそろえるのは必ずしも容易ではない．楽譜から音響信号を生成する手法 [5] を用いれば学習データを増加させることができるが，それ自体の学習に十分な量のペアデータが必要であり，本質的な解決にならない．

本研究では，楽譜の認識器とスペクトログラムの生成器を同時に半教師あり学習する手法を提案する．具体的には，スペクトログラムを認識器に入力して楽譜を得て，さらに生成器に入力してスペクトログラムを再構成した際に，両者を近づける Cycle-Consistency 基準と，楽譜を生成器に入力してスペクトログラムを得て，さらに認識器に入力して楽譜を再構成した際に，両者を近づける Cycle-Consistency 基準をもとに学習を行う．これにより，音響信号と楽譜のペアデータに加えて，いずれかのみのデータを併用した半教師あり学習が可能になる．音声認識では，この種の Cycle-Consistency 学習が精度向上に寄与することが報告されている [6].

2. 提案法

提案法は，スペクトログラムをピアノロールに変換する認識ネットワーク（S2P-Net）と，ピアノロールをスペクトログラムに変換する生成ネットワーク（P2S-Net）から構成される．音響特徴量として，対数メルスペクトログラム $\mathbf{X} \in \mathbb{R}^{F \times T}$ を用いる．ここで，F は対数スケール周波数ビン数，T は時間フレーム数を表す．ピアノロールは，ピアノの音域に対応する MIDI ノート番号 21 から 108 の $\mathbf{Y} \in \{\mathbf{0}, \mathbf{1}\}^{\mathbf{88} \times \mathbf{T}}$ で表される．

2.1 認識ネットワーク・生成ネットワーク

認識ネットワーク S2P-Net を CNN で構成する．このネットワークは入力のスペクトログラム $\mathbf{X}$ を $(1 \times F \times T)$ の画像とみなしチャンネル毎に二次元のカーネルを畳み込むことで $(1 \times 88 \times T)$ のピアノロール $\mathbf{Y}^*$ を推定する．

$$\mathbf{Y}^* = \mathrm{S2P}(\mathbf{X}) \tag{1}$$

このネットワークは以下のバイナリクロスエントロピーを最小化させることで最適化される．

$$\mathcal{L}_{\mathrm{S2P}} = -\frac{1}{88 \times T} \sum_{m=21}^{108} \sum_{t=1}^{T} \hat{y}_{m,t} \log y^*_{m,t} \tag{2}$$

ここで $\hat{y}_{m,t}$ は時間フレーム t において正解ピアノロールの MIDI ノート番号 m の音が鳴っていれば 1，鳴っていなければ 0 をとり，$y^*_{m,t}$ は時間フレーム t における MIDI ノート番号 m の音高の出力確率を表す．

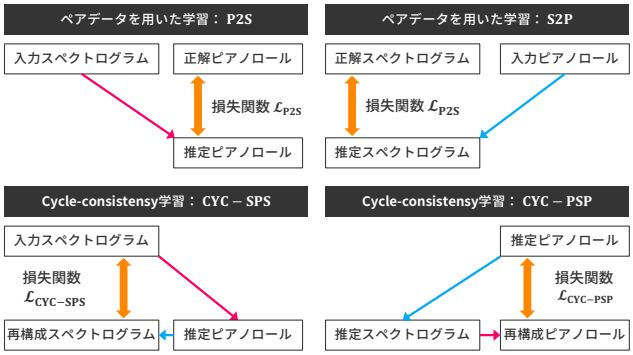

図 1: モデル全体像

P2S-Net は S2P-Net の逆の過程を表現するネットワークである．S2P-Net 同様に入力ピアノロール $\mathbf{Y}$ を $(1 \times 88 \times T)$ の画像とみなし，$(1 \times F \times T)$ のスペクトログラムを生成するネットワークを CNN で構成する．

$$\mathbf{X}^* = \mathrm{P2S}(\mathbf{Y}) \tag{3}$$

このネットワークは以下の平均二乗誤差を最小化することで最適化される．

$$\mathcal{L}_{\mathrm{P2S}} = \frac{1}{F \times T} \sum_{f=1}^{F} \sum_{t=1}^{T} (x^*_{f,t} - \hat{x}_{f,t})^2 \tag{4}$$

ここで，$\hat{x}_{f,t}$ は正解スペクトログラムの対数スケールビン f，時間フレーム t の要素である．

2.2 Cycle-Consistency コスト

二つのネットワークを統合的に最適化するために Cycle-Consistency 学習を行う．Cycle-Consistency 学習では S2P-Net に P2S-Net を連絡し，スペクトログラム $\mathbf{X}\rightarrow$ 推定ピアノロール $\mathbf{Y}^*\rightarrow$ 再構成スペクトログラム $\mathbf{X}^{**}$ の循環型ネットワークを構成し，以下の Cycle-Consistency 損失関数 $\mathcal{L}_{\text{cyc-sps}}$ を最小化する．

$$\mathbf{X}^{**} = \mathrm{P2S}(\mathrm{S2P}(\mathbf{X})) \tag{5}$$

$$\mathcal{L}_{\text{cyc-sps}} = \frac{1}{F \times T} \sum_{f=1}^{F} \sum_{t=1}^{T} (x^{**}_{f,t} - x_{f,t})^2 \tag{6}$$

また，同様に P2S-Net に S2P-Net を連絡し，ピアノロール $\mathbf{Y}\rightarrow$ 生成スペクトログラム $\mathbf{X}^*\rightarrow$ 再構成ピアノロール $\mathbf{Y}^{**}$ の循環型ネットワークを構成し，以下の Cycle-Consistency 損失関数 $\mathcal{L}_{\text{cyc-psp}}$ を最小化する．

$$\mathbf{Y}^{**} = \mathrm{S2P}(\mathrm{P2S}(\mathbf{Y})) \tag{7}$$

$$\mathcal{L}_{\text{cyc-psp}} = -\frac{1}{88 \times T} \sum_{m=21}^{108} \sum_{t=1}^{T} y_{m,t} \log y^{**}_{m,t} \tag{8}$$

2.3 学習手順

具体的な学習手順を述べる．まずペアデータを用いて P2S-Net と S2P-Net を独立に学習する．

Automatic music transcription based on Cycle-Consistency: *Kentaro Shibata, Ryo Nishikimi, Eita Nakamura, and Kazuyoshi Yoshii (Kyoto Univ.)*

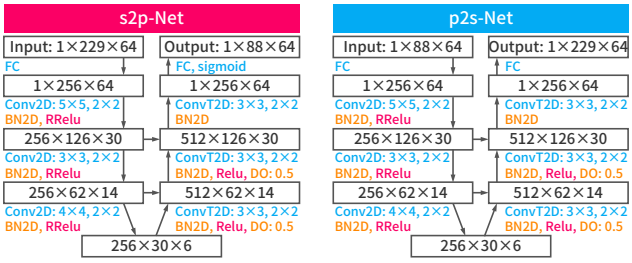

図 2: ネットワーク構成

1. S2P-Net の学習（式 (2) を最適化，ペアデータ）
2. P2S-Net の学習（式 (4) を最適化，ペアデータ）

次に，ペアデータに加えて，ペアを持たない音響・楽譜データを用いてネットワーク全体を最適化する．

3. S2P-Net→P2S-Net の Cycle-Consistency 学習（式 (6) を最適化，非ペアデータ）
4. P2S-Net→S2P-Net の Cycle-Consistency 学習（式 (8) を最適化，非ペアデータ）
5. S2P-Net の学習（式 (2) を最適化，ペアデータ）
6. P2S-Net の学習（式 (4) を最適化，ペアデータ）

3. 評価実験

予備的な実験を行った．MAPS データベース [7] の学習用データ（”ISOL”, ”RAND”, ”UCHO”のラベルがついたもの）を教師あり学習用と半教師あり学習用に二分割し，それぞれ 80% を学習，10% を検証，10% をテストに用いた．入力音響信号は対数 メルスペクトログラムとし，サンプリング周波数は 16 kHz，窓幅は 1048 点，シフト幅は 512 点，メルフィルタバンク数は 229 とした．本手法では CNN を用いたためスペクトログラムを 64 フレーム（約 2 秒）毎に区切り入力とした．ピアノロールは MIDI ノート番号 21-108 の 88 次元の 0，1 のベクトルをフレーム時間に対応させたものとした．

S2P-Net と P2S-Net にはそれぞれスキップコネクションを持つ U-Net を用いた．提案手法で用いたネットワークの構造を図 2 に示す．ここで，入出力及び中間層はチャンネル数×ビン数×フレーム数で表記し，全結合層，畳み込み層，逆畳み込み層，バッチ正則化層，ドロップアウト層を FC，Conv2D，ConvT2D，BN2D，DO で表記した．パラメータの更新はミニバッチサイズ 64 として，学習率 0.006 の Adam [8] によって行った．

事前学習のみを行ったネットワークによる推定結果と，事前学習に加え Cycle-Consistency 学習を行ったネットワークによる推定結果の例を図 3 に示す．図 3 の例ではピアノロールの推定誤りが減っていることを確認できる．ペアデータのみを用いた学習では再構成スペクトログラムがぼやけているが，Cycle-Consistency 学習を行ったネットワークではより入力に近いスペクトログラムを再構成できていることが分かる．また，どちらのネットワークでも推定ピアノロールのフレーム番号 10 前後で極端に短い挿入誤りをしている．このように音楽的に不自然な誤りを減らすためには，Generative Adversarial Network（GAN）を用いてピアノロールの尤もらしさを測る損失関数を加える拡張が考えられる．

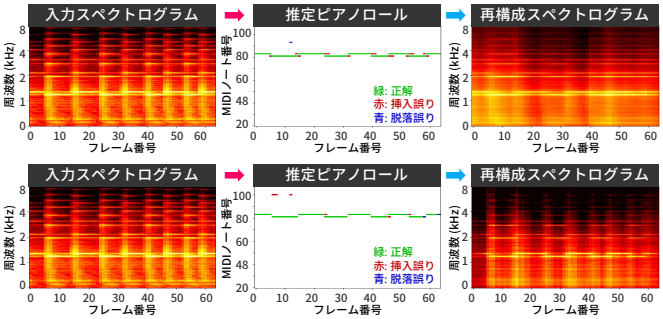

図 3: 推定・再構成結果の例．上段がペアデータのみで学習を行った結果で，下段がペアデータでの学習に加え Cycle-Consistency 学習を行った結果．

4. おわりに

本稿では音響信号からピアノロールを推定するネットワークを Cycle-Consistency 学習を用いて最適化する枠組みを提案した．ペアデータとペアを持たないデータを同時に用いて半教師あり学習をすることで採譜精度が改善することを確認した．これは，音楽の採譜のようにペアデータの量が豊富でない状況でモデルを学習する上で有効であると考えられる．今回は単純なデータのみを用いて実験を行ったが，実際の曲を用いてより実践的なデータに対する有効性を確認する必要がある．また，提案するネットワークに GAN の枠組みを統合し [9]，Cycle-Consistency 学習において，スペクトログラムおよびピアノロールそれぞれの最もらしさを評価する損失関数を加えることでさらなる採譜精度の向上を目指す．

謝辞 本研究の一部は、JSPS 科研費 16J05486，16H01744 および JST ACCEL No. JPMJAC1602 の支援を受けた．

参考文献

[1] P. Smaragdis *et al.* Non-negative matrix factorization for polyphonic music transcription. In *WASPAA*, 177–180, 2003.

[2] E. Benetos *et al.* An efficient temporally-constrained probabilistic model for multiple-instrument music transcription. In *ISMIR*, 701–707, 2015.

[3] R. M. Bittner *et al.* Deep salience representations for F0 estimation in polyphonic music. In *ISMIR*, 63–70, 2017.

[4] C. Hawthorne *et al.* Onsets and Frames: Dual-objective piano transcription. In *ISMIR*, 50–57, 2018.

[5] J. Engel *et al.* Neural audio synthesis of musical notes with wavenet autoencoders. *ICML*, 1068–1077, 2017.

[6] T. Hori *et al.* Cycle-consistency training for end-to-end speech recognition. *arXiv :1811.01690*, 2018.

[7] V. Emiya *et al.* Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE TASLP*, 18(6):1643–1654, 2010.

[8] D. P. Kingma *et al.* Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[9] J. Zhu *et al.* Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.