E-018

各楽器音の時間-周波数特徴の変化に追従可能な Deform-Conv Dense U-Net による音源分離法

A Source Separation Method Using Deform-Conv Dense U-Net Capable of Tracking Fluctuations in Time-Frequency Features of Each Instrument Sound

竹田 舜 †
Shun Takeda

荒井 秀一 †
Shuichi Arai

1. はじめに

楽曲音源分離とは複数の楽器で構成される音源から目標音源を分離する研究分野である．楽曲の音源分離では畳み込み層で構成された Convolutional Neural Network(CNN) を用いた手法が主流であり，U-Net[1] などの深層学習法が成果を挙げている．しかし，それらの従来法は急激に時間-周波数特性が変化する箇所で分離性能が低下する問題を抱えている．そして，その問題は楽器音のスペクトログラムの違いが関係している．そこで本稿ではそれを改善するために楽器音と既存の CNN の関係に着目した．前述のように各楽器音によってスペクトログラムに差異があるが，従来法はその差異を無視して均一な処理をする．そのため，楽器音に適した畳み込みが施されていないと考えた．そこで楽曲構成音の時間-周波数空間での局所的な特徴変化に追従して注視領域の形状を変形することで，分離性能を改善できるのではないかと考えた．具体的には，各楽器音の局所的な特徴に追従するために Deformable Conolution(Deform-comv)[2] を用いた方法を提案する．

2. 楽器音の時間-周波数特性

楽器音の時間-周波数特性には，演奏音が最大音量に達するまでの Attack, 音が一定値まで減衰するまでの Decay, 音が持続する Sustain, 演奏をやめたときの残響である Release がある．瞬発音とは Sustain がなく，瞬間的に鳴るもので，Drums などの打楽器が分類される．継続音は Sustain があり，徐々に減衰する音で，Vocals, Bass などが分類される．このような性質から，音高がなく周波数帯に広く分布する Drums では縦に広く周波数スペクトログラムが広がり，Vocals のような音高の連続的変化がある継続音では周波数方向，時間軸方向ともに周波数スペクトログラムが広がるべきである．しかし，U-Net などの従来手法では，予め決められたサイズで畳み込みが施される．

本稿では，楽曲の時間-周波数特性の変化に応じて畳み込みの受容野を変形すべきだと考えた．

3.Dense U-Net

図 1 に示す Dense U-Net は，音源分離タスクで高い性能を達成している定評ある手法で，本稿ではこれをベースアーキテクチャとして用いる．Dense U-Net は振幅スペクトログラムを入力とし，分離に用いるマスクを出力する．U-Net はボトルネック構造を有し，音源を抽出するための時間-周波数マスクを学習する．U-Net の各部を構成する DenseBlock[3] は図 1 にあるように同ブロック内の畳み込み層をすべて結合する．これにより，勾配消失の軽減と特徴量伝達を強化を行う．

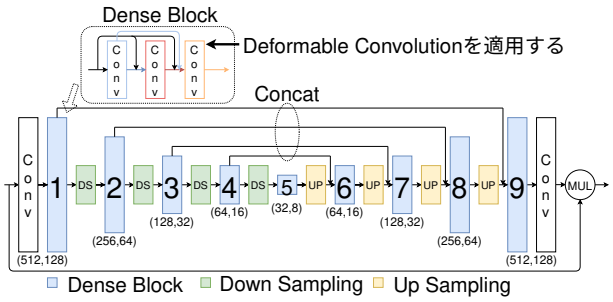

図 1: DenseU-Net．Block 8 に deform-conv を適用

4. 提案手法

2 章で述べたように各楽器音は時間-周波数ごとに特性が変化するので，それに追従して畳み込み受容野を変形できるよう Deform-Conv を導入したアーキテクチャを提案する．Deform-Conv は不定形な領域の特徴抽出を目的とした提案された畳み込みである．Deform-Conv の処理を図 2 に示す．Deform-Conv は従来の畳み込みでは固定であったサンプリング位置の offsets ベクトルを用いた移動で，入力の特徴に応じてサンプリング位置を変える可変受容野を実現している．Deform-Conv を式 1 に示す．

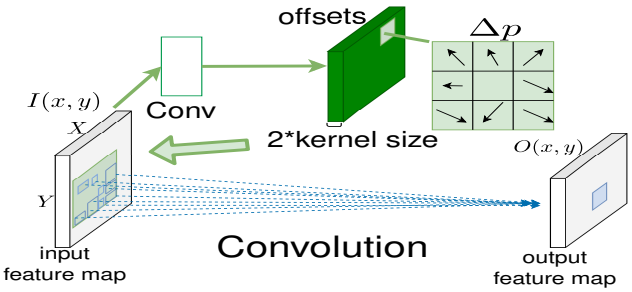

図 2: Deformable Convolution の概要

$$O(x,y) = \sum_{m=-M}^{M} \sum_{n=-N}^{N} K(m,n) * I(x+m+\Delta p_x, y+n+\Delta p_y) \quad (1)$$

出力を $O(x,y)$，入力マップを $I(x,y)$，畳み込み受容野 (サンプリング範囲) を $K(m,n)$ で表す．入力マップのサイズは (X,Y) で表され，受容野のサイズは (M,N) で表される．式 1 中の x, y, m, n はそれぞれ $(m = -M,\ldots,M)$ $(n = -N,\ldots,N)$ $(x =$

†東京都市大学 総合理工学研究科

$1,\ldots,X)$ $(y=1,\ldots,Y)$ の範囲で動く．Deformable Convolution は従来の Convolution に offsets($\Delta p_x, \Delta p_y$ を加えて，受容野を可変にしている．Δp_x は x 軸移動量, Δp_y は y 軸移動量を表す．offsets の変化量は畳み込み受容野の重みとして学習・最適化されていく．

この重みをもとにスペクトログラム上での音の形状に合わせて受容野を変形する．また，Dense U-Net における Deform-Conv の適用箇所は Vocals 音源を用いた予備実験によって決めた．その結果を表 1 に示す．

表 1: 適用箇所を決める予備実験の結果

	Median SDR
Dense U-Net	5.22
DConv DU-Net(No.6)	5.07
DConv DU-Net(No.7)	5.61
DConv DU-Net(No.8)	**5.62**

表 1 を見ると図 1 の No. 8 Block に適用した場合にもっとも効果があった．そのため，以降の実験では全て No.8 Block に Deform-Conv を適用した．

5. 実験及び結果

まず，各アーキテクチャの定量評価を SDR で行う．次に offsets について考察するために offsets の値を集計した．実験では MUSDB18[4] をデータセットに使用した．MUSDB18 には楽器ごと (Bass,Drums,Vocals,Other) の音源が学習用 100 曲，評価用 50 曲含まれている．データセットは予め短時間フーリエ変換し，振幅スペクトログラムにした．テストセットを分離し，分離音源と正解音源を比較して評価する．

5.1. 分離音源の評価

評価基準として Source-to-Distortion Ratio(SDR)[5] を用いる．SDR は下記の式 2 で定義される．

$$SDR = 10\log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}+e_{noise}+e_{artif}\|^2} \quad (2)$$

比較のため，Median Source-to-Distortion Ratio(Mdn SDR) を算出した．Mdn SDR は SDR(式 2) の中央値である．S_{target} は抽出したい音源の正解データ, e_{interf} は推定結果において抽出したい音源以外の音源, 例えば, Vocals を評価したい場合は Bass, Drums, Other を指す．e_{noise} はセンサーノイズ，e_{artif} は丸め誤差などの表現エラーである．本実験ではこれらの値のうち，e_{noise}，e_{artif} は非常に小さく，0 として扱える．SDR は分母がノイズとして計算されため，教師データと推定結果が似ている場合，すなわちノイズが小さい場合は分母が 0 に近づくため評価値は大きくなる．そのため，SDR が大きいほどいい分離結果であるといえる．性能のベースラインとして Dense U-Net を使用した．表 2 に実験結果を示す．

表 2 を見るとベースラインと提案手法の SDR 比較において，Drums が 0.25pt, Vocals が 0.40pt, Other が 0.1pt 向上した．ベースモデルと提案手法を比較すると， 3 種類の楽器において SDR の向上が確認できた．この結果から，受容野の可変化は楽器の音源分離において効果があるとわかった．

表 2: 従来手法 (上段) と提案手法 (下段) の Mdn SDR 比較

	Bass	Drums	Vocals	Other
Dense U-Net	3.14	4.74	5.22	1.92
DConv DU-Net	2.87	5.00	5.62	2.02

5.2.Offsets ベクトルの評価

test set における offsets の x 軸方向と y 軸方向の平均と標準偏差を表 3, 表 4 に示す．

表 3: offset の x 軸方向における平均と標準偏差

	x 軸平均 (周波数)	x 軸標準偏差 (周波数)
Drums	61.64	94.22
Bass	0.66	1.20
Vocals	9.97	29.84
Other	0.39	1.14

表 4: offset の y 軸方向における平均と標準偏差

	y 軸平均 (時間)	y 軸標準偏差 (時間)
Drums	0.47	0.94
Bass	0.28	0.58
Vocals	4.80	12.48
Other	6.12	12.87

表 3，表 4 より Drums の offsets の y 軸方向へ広がりを確認できる．これは音高がなく，瞬発音である Drums の特徴と一致している．Vocals の値より周波数を捉えるため広がりつつも時間方向にも広がっているため，継続音の特徴を捉えている．以上の 2 つから Deform-Conv の楽器音に追従した変形が確認できた．一方で Bass の場合は, SDR が低下する結果となった．

これは Bass に分類される音と Other に分類されている音の類似が原因だと考えられる．

6. おわりに

CNN の一部の畳み込み受容野を変形させて楽器音に追従する手法の有効性が確認できた．今後の課題は offsets の明示的な制御, 変形した受容野の可視化である．

参考文献

[1] Jansson et al. Singing voice separation with deep u-net convolutional networks. *ISMIR*, 2017.

[2] Dai et al. Deformable convolutional networks. In *ICCV*, 2017.

[3] Takahashi et al. Multi-scale multi-band densenets for audio source separation. In *2017 WASPAA*. IEEE, 2017.

[4] Rafii et al. Musdb18-a corpus for music separation. 2017.

[5] Vincent et al. Performance measurement in blind audio source separation. *IEEE/ACM*, 2006.