E-016

スペクトル特徴に応じた時間・周波数分解能を学習可能な Multi Window Lengths U-Net による楽曲音源分離

Music source separation by Multi Window Lengths U-Net that can learn time and frequency resolution according to spectral features

坂藤 隆我 † Ryuga Sakato
荒井 秀一 † Shuichi Arai

1. はじめに

楽曲の音源分離は, 複数の楽器で構成される音源から 1 つ, または複数の楽器音源を分離する研究分野である. 音源分離は自動採譜 [1] や歌詞認識 [2] などの音楽情報処理の前処理として使える. そのため, 分離音は本来の音に近い品質が求められる.

2. 従来手法

楽曲の音源分離は, スペクトログラムを対象とした周波数領域での処理を用いた手法がある. この手法が扱うスペクトログラムは, 図 1 に示すように Mixture の時間域信号に対して, 短時間フーリエ変換 (Short-time Fourier transform, STFT) を用いて変換されたものである. 従来の周波数領域での処理を用いた手法は, 図 1 の音源分離器で不要な音源のスペクトルを低減し, 目標音源を抽出するためのマスクを生成する. そして, 生成したマスクを入力スペクトログラムに掛けて目標音源を分離する. このマスクは U-Net のような深層学習の手法で生成するのが主流である.

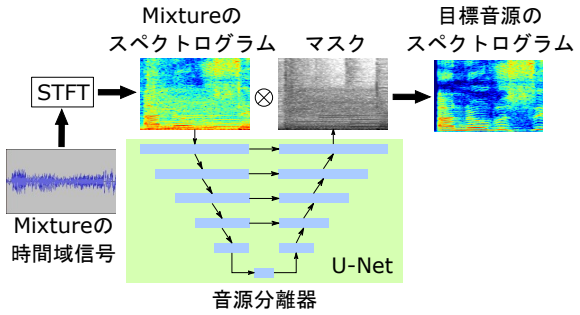

図 1: 周波数領域での処理を用いた音源分離

3. 提案手法

第 2 章で述べたように, 従来の周波数領域での処理を用いた手法は, STFT で変換されたスペクトログラムをネットワークアーキテクチャの入力とする. しかし, 従来手法が入力として扱うスペクトログラムは既に, 図 2 の白枠で囲われた部分のように立ち上がり, 立ち下がり部分や定常部で目標音源と不要な音源のスペクトルが混合してしまう. そして, 入力の時点で混合したスペクトルは従来手法が生成したマスクでは分離ができない. これは, STFT で変換されたスペクトログラムが分解能のトレードオフの関係を持っているにも関わらず, 従来手法が単一のフレーム長で周波数解析を行い, 単一の時間・周波数分解能を扱ってしまうからである. そこで, 異なる分解能を持つスペクトログラムからスペクトル特徴に応じて時間・周波数分解能を学習する Multi Window Length U-Net(MWL U-Net) を提案する.

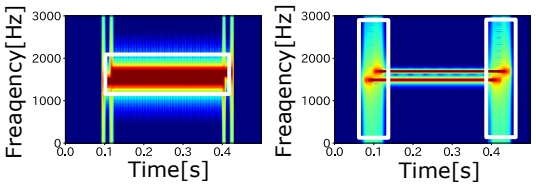

図 2: 1500[Hz] と 1700[Hz] の sin 波を 20[ms] ずらして鳴らした音源に対して, 異なるフレーム長で変換されたスペクトログラム. 縦軸は周波数, 横軸は時間を表す. 色は信号の振幅を表し, 赤くなるほど大きい.

MWL U-Net は, 図 3 に示すように入力時間域信号に対して異なるフレーム長で変換された複数のスペクトログラムから目標音源を分離する「分離部」と, 分離後の各スペクトログラムの長所を持つように 1 つに合成するための重みを生成する「重み生成部」から構成される. 分離部では, まず入力時間域信号をフレーム長 512, 1024, 2048 の STFT で変換する. (なお, 今後はフレーム長 512, 1024, 2048 で変換されたスペクトログラムをそれぞれ Short, Medium, Long とする.) その後, 各 U-Net にそれぞれのスペクトログラムを入力し分離をする.

分離後の各スペクトログラムは, Short が高い時間分解能を持つので立ち上がり, 立ち下り部分の分離性能が良く, Long は高い周波数分解能を持つので定常部の分離性能が良い. 重み生成部では, このような各スペクトログラムの局所的な分離性能の優れた部分を 1 つに合成するための重みを生成する. そのために, 重み生成部では 1 層目の 5×5 Convolution で分離後の各スペクトログラムから局所的な特徴 (立ち上がり, 定常部など) を抽出し, 2, 3 層目の 1×1 Convolution で各解像度における分離性能が良い特徴を選択する. MWL U-Net は分離部が分離したスペクトログラムと, 重み生成部が生成した重みを用いて最終的な出力を合成する.

4. 実験及び実験結果

本実験では, MUSDB18[4] を使用した. MUSDB18 には, 楽器ごと (Vocals, Bass, Drums, Other) の音源とそれらの Mixture が学習用に 100 曲, 評価用に 50 曲

†東京都市大学 総合理工学研究科

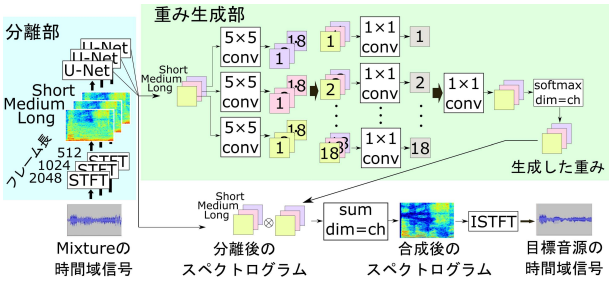

図 3: Multi Window Lengths U-Net

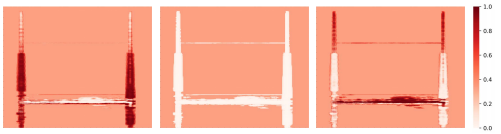

図 4: Multi Window Lengths U-Net・重み生成部が生成する重みの例. 色は重みの大きさを表し, 赤くなるほど値が大きい.

含まれている. 本実験では Vocals を分離対象とした.

4.1. 分離音源の評価

提案手法の分離音源を評価するために, 本実験ではテストデータの Signal to Distortion Ratio(SDR)[5] の平均値である Mean SDR を用いる. SDR は式 (1) で定義され, s_{target} は抽出した音源の正解データ, e_{interf} は分離結果において抽出した音源以外の音源, 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を表す.

$$SDR = 10log_{10}\frac{\| s_{target} \|^2}{\| e_{interf} + e_{noise} + e_{artif} \|^2} \quad (1)$$

SDR は分母がノイズとして計算されるため, 正解データと分離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. 一方, ノイズが大きい場合は分母が大きくなるため値が小さくなる, または負の値をとる. そのため, SDR は値が大きいほど良い分離結果であるといえる. 本実験では性能のベースラインとして U-Net を使用した. U-Net と Multi Window Lengths U-Net の Mean SDR を表 1 に示す.

表 1 を見ると, U-Net と MWL U-Net の Mean SDR 比較において, MWL U-Net は U-Net より 0.687pt 向上した.

表 1: Vocals に対する U-Net と Multi Window Lengths U-Net の Mean SDR

	Mean SDR
U-Net	3.940
MWL U-Net	4.627

4.2. 局所的な分離音源の評価

第 3 章で述べたように, MWL U-Net は分離後の各スペクトログラムの局所的な分離性能の優れた部分を 1 つに合成するので, 立ち上がり, 立ち下り部分や定常部において U-Net よりも分離性能に優位性があると考えられる. そこで, 局所的な分離音源に対して SDR を算出する Framwise SDR を用いる. Framwise SDR とは, 予め決めたフレーム長で分離音源を区切り, それぞれの分離音源ごとに SDR を算出する評価法である. 本実験では 1 秒ごとに分離音源を区切り, U-Net と MWL U-Net の分離音源の内, 立ち上がり, 立ち下り部分と定常部の Framwise SDR を比較する. 立ち上がり, 立ち下り部分の Framwise SDR の差分と定常部の Framsise SDR の差分を図 5 に示す.

図 5 より, MWL U-Net は立ち上がり・立ち下り部分や, 定常部において, U-Net よりも Framewise SDR が高いことが確認できる. この結果から, MWL U-Net は局所的な部分において U-Net よりも分離性能に優位性があることがわかる. このことより, 異なる分解能を持つスペクトログラムからスペクトル特徴に応じた時間・周波数分解能を学習する MWL U-Net が有用であることがわかった.

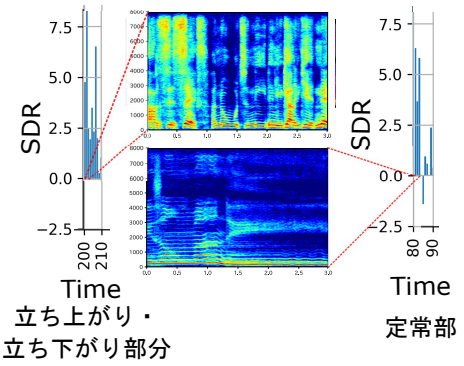

図 5: MWL U-Net と U-Net の Framewise SDR の差分. 左右のグラフは縦軸は SDR, 横軸は時間を, 中央のスペクトログラムは正解 Vocals を表す.

5. おわりに

異なる分解能を持つスペクトログラムからスペクトル特徴に応じた時間・周波数分解能を学習可能な MWL U-Net による音源分離を行った結果, 従来手法の U-Net に比べ SDR と Framwise SDR での有効性が確認できた.

参考文献

[1] Annamaria Mesaros et al, EURASIP Journal on Audio, Speech, and Music Processing, 2010.

[2] Mark D Plumbley et al, Automatic music transcription and audio source separation, Cybernetics Systems 2002.

[3] Andreas Jansson et al, SINGING VOICE SEPARATION WITH DEEP U-NET CONVOLUTIONAL NETWORKS, ISMIR 2017.

[4] Zafar Rafii et al, MUSDB18-a corpus for music separation, 2017.

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