長岡技術科学大学大学院
工学研究科修士論文

題　　目

自己注意機構を用いたマルチバンド ニューラルネットワークによる 楽器音分離

指導教員

杉田 泰則　　准教授

著　　者

電気電子情報工学専攻
19312289　　小熊 隼

令和 5 年 2 月 10 日

ABSTRACT

Music source separation using multi-band self-attention neural networks

Department of Electrical, Electronics
and Information Engineering
Author : 19312289 Shun OGUMA
Supervisor : Assoc. prof. Yasunori SUGITA

Music Source Separation(MSS) is the technology to separate the sounds of specific instruments from sounds which mix many musical instruments. As examples of usage, instruments' sound extraction/muting enables the creation of karaoke sounds, preprocessing for automated transcription, etc.

In recent years, Deep Neural Networks(DNNs) are considered the superior methods for MSS. They are particularly effective for the problem that the mixing path is unknown, and the number of observable sounds is less than the number of sound sources, in short, underdetermined blind source separation(UBSS). Existing methods, multi-scale and multi-band networks with Long-Short Term Memory(LSTM) achieve good separation. LSTMs model the temporal features of the input music, and multi-band structures learn the frequency features. Furthermore, a multi-scaled model using self-attention instead of LSTM shows a better score than it. Self-attention can model the longer context than LSTM.

This paper proposes a method using multi-scale and multi-band structured networks with self-attention to improve the separation quality. I expect that the networks model the temporal characteristics with self-attention mechanisms and learn the frequency characteristics with multi-band structure so that the method can produce better separation results.

I compare separating results with MMDenseLSTM as an existing method. Both existing and proposed methods use the MUSDB18 dataset to train and evaluate. As a result of experiments, I found that the proposed method is able to estimate better separation than existing for all instruments. In particular, tracks such that the length of phrases varies from instruments to other instruments achieved better results.

目次

第 1 章　はじめに 1
1.1　研究背景 . . . 1
1.2　研究目的 . . . 2
1.3　本論文の構成 . . . 2

第 2 章　関連研究 3
2.1　MMDenseNet . . . 3
2.1.1　Multi-scale 構造 . . . 3
2.1.2　Multi-band 構造 . . . 4
2.2　MMDenseLSTM . . . 4
2.3　Self-Attention Dense-UNet . . . 5
2.3.1　自己注意 . . . 6

第 3 章　提案手法 8
3.1　Self-Attention Dense-UNet の問題点 . . . 8
3.2　周波数帯による Self-Attention モデルの並列化 . . . 9
3.3　自己注意の挿入箇所の変更 . . . 11

第 4 章　実験 12
4.1　実験条件 . . . 12
4.1.1　モデル構造の詳細 . . . 12
4.1.2　学習条件 . . . 13
4.1.3　データセット . . . 13
4.1.4　前処理 . . . 13
4.1.5　後処理 . . . 13
4.1.6　評価指標 . . . 14
4.2　実験結果 . . . 14

第 5 章　おわりに 24
5.1　まとめ 24
5.2　今後の展望 24

謝辞 25

参考文献 26

第 1 章 はじめに

1.1 研究背景

近年，コンピュータ上でデスクトップミュージック (DTM) ソフトウェアやディジタルオーディオワークステーション (DAW) 等を用いて，個人で手軽に楽曲へアレンジを施すなど，音楽の楽しみ方に変化が生じている．こうした動向の中にあって，様々な楽器音が混合した音源からそれぞれの楽器の音を分離する技術である楽器音分離は，重要な技術である．楽器音分離によって音源から特定の楽器音のみを抽出あるいは削除できるため，ボーカルを取り除くことによるカラオケ音源の作成や，パート毎に分かれた譜面を自動で採譜するための前処理などに活用できる．

これまでに提案されてきた楽器音分離の手法としては非負値行列因子分解 [1] や多チャンネルウィーナーフィルタ [2] などがある．近年ではニューラルネットワーク (DNN) などの機械学習を用いた手法が盛んであり，特に混合経路が未知であるブラインド音源分離やマイク数が音源数より少ない劣決定音源分離においては，楽器や音声など大量のデータを用意しやすい分離対象の場合 DNN が有効な手法となる [3].

DNN を用いた例として Nugraha らや Urlich ら [4, 5] の研究があり，これらは標準的な全結合ネットワーク (FNN) を用いている．また Uhlich らは FNN と時系列データの学習が得意な長・短期記憶 (LSTM) との出力を結合させることで単一のネットワークを用いたモデルよりも性能を向上させており，異なる種類のネットワークの組み合わせが楽器音分離に効果的であることを示している [6]．また，MMDenseNet [7] は DenseNet [8] と呼ばれる畳み込みニューラルネットワーク (CNN) の一種を楽器音分離のために改良したモデルで，Multi-scale および Multi-band と呼ばれる構造によって Urlich らの手法 [6] よりも少ないパラメータ数でより高い性能を実現している．さらに，MMDenseLSTM [9] は MMDenseNet [7] と LSTM を組み合わせた手法であり，更なる分離性能を実現している．特に，ネットワークを並列化する Multi-band 構造はモデル化の対象となる音に特有の周波数特性を考慮した構造であり，分離性能の向上に大きく影響すると考えられる [7, 9].

さらに Liu らは，タスクをボーカルとそれ以外の楽器という 2 種類の分離に限定して，自己注意 (Self-Attention) と呼ばれるネットワークと Multi-scale な DenseNet とを組み合わせることで MMDenseLSTM [9] よりも高い精度での分離を行なってい

る [10]．自己注意機構は LSTM よりも広い範囲のコンテキストを学習することが可能であり，入力となる楽曲の時間的な変化を良くモデル化する．そのため，時系列データの学習の中でも，特に長期的な関係性を考慮する必要のある音楽のようなデータの学習においては，LSTM よりも自己注意機構を用いた方が性能の向上に寄与すると考えられる．

1.2　研究目的

本論文では，楽曲の時間的・周波数的特徴の双方について学習精度を向上させることを目的とし，自己注意機構と Multi-scale な DenseNet を組み合わせたモデルに Multi-band 構造を導入した手法を提案する．Multi-scale DenseNet は時間的な，Multi-band 構造は周波数的な特徴のモデル化に適していると考えられるため，これらの組み合わせによって分離性能が向上することを期待する．

1.3　本論文の構成

第 1 章では研究の背景および目的を述べた．第 2 章では提案手法に関連した楽器音分離の従来手法について述べる．また第 3 章では提案手法である Multi-band 化した自己注意を用いたネットワークの構造について詳細を示す．第 4 章では従来手法と提案手法の比較実験を行ない，提案手法の有用性を示す．そして第 5 章ではまとめと今後の課題を示し結びとする．

第 2 章 関連研究

2.1 MMDenseNet

MMDenseNet [7] は，画像認識の領域で提案された DenseNet [8] を発展させたモデルである．

DenseNet は畳み込み層 (CNN) を密に結合したネットワークを有し，各層の入力と出力とを結合して次の層へ入力する．

$$x^{(l)} = H([x^{(l-1)}, x^{(l-2)}, \ldots, x^{(0)}]) \tag{2.1}$$

ここで $x^{(i)}$ は i 層目の出力，$H(\cdot)$ は非線形関数，$[\ldots]$ は特徴量の連結 (concatenation) を表す．音源分離においてもこの構造が有用であり，最初の入力である混合音源や前層の出力から分離音源を効率よく推定できるとしている [7]．ただし，層数が増加するにつれて層の間の結合の数が指数関数的に増大し，多くのメモリを要求するという欠点がある．

MMDenseNet は，以下で説明する Multi-scale 構造によって必要なメモリ量を削減し，また Multi-band 構造によって分離性能の向上を実現している．

2.1.1 Multi-scale 構造

Multi-scale DenseNet(MDenseNet) は，図 2.1 に示す dense block とダウン/アップサンプリング層からなる．同図中の comp. layer(composition layer) は，1.) バッチ正則化層，2.) 活性化関数層，3.) 畳み込み層を合成したものを表している．ダウンサンプリング層によって特徴量のサイズを小さくすることでメモリの使用量を抑えつつネットワークを深くし，長時間・広い周波数帯域の特徴量の学習が可能になる [7]．小さいサイズにエンコードされた特徴量は，アップサンプリング層によって元のサイズまで復元される．このとき，ダウンサンプリング層へ入力する前の高解像度の特徴量を，同サイズのアップサンプリング層出力と連結することで，圧縮による情報の損失を抑えている．MDenseNet の全体は図 2.2 の通りである．

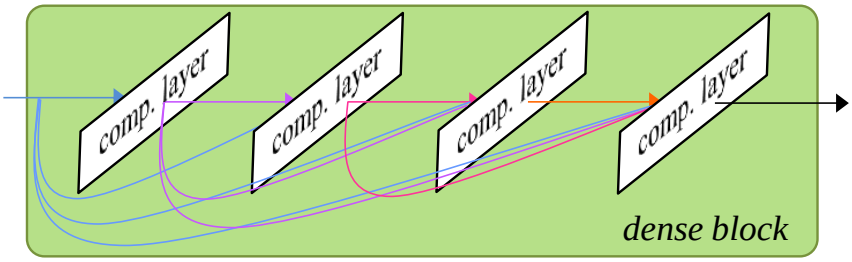


図 2.1: Dense block([7])

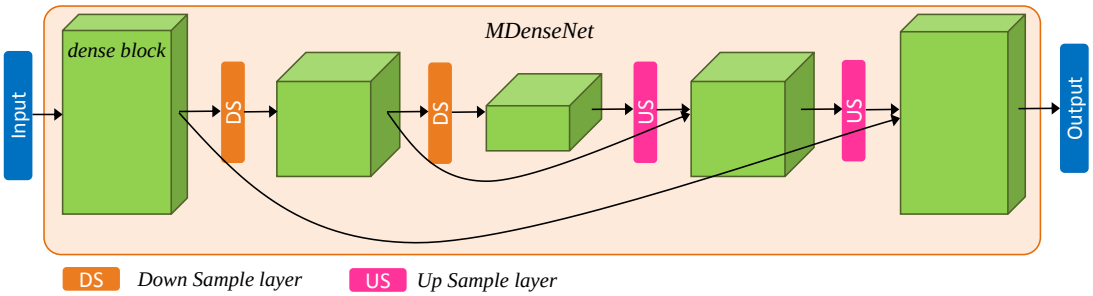


図 2.2: MDenseNet([7])

2.1.2　Multi-band 構造

また高橋ら [7] は，画像を対象とするタスクでは画像中の様々な箇所に同様のパターンが出現することが多いが，楽器音のスペクトログラムは一般的に周波数帯域ごとに大きく異なるパターンを持つことに着目して，入力をいくつかの周波数帯域で分割してそれぞれに独立した MDenseNet を学習させる手法を提案した．各畳み込み層のカーネルが学習する帯域を制限することでより効率的に特徴を学習できるとして，この構造を Multi-band MDenseNet(MMDenseNet) と呼んでいる．図 2.3 に示すように入力を N 個の帯域に分割してそれぞれの MDenseNet に入力し，それらの出力を再び連結して最後の dense block へと入力する．

2.2　MMDenseLSTM

MMDenseLSTM は，MMDenseNet [7] と双方向長・短期記憶 (BLSTM) とを組み合わせることで分離性能を向上させたモデルである [9]．Uhlich らの研究 [6] によれば異なる種類のネットワークの組み合わせが楽器音分離の性能向上に有効であり，LSTM が楽曲の長期的なコンテキストを学習することで高い性能を発揮するとしている．MMDenseLSTM の構造を図 2.4 に示す．

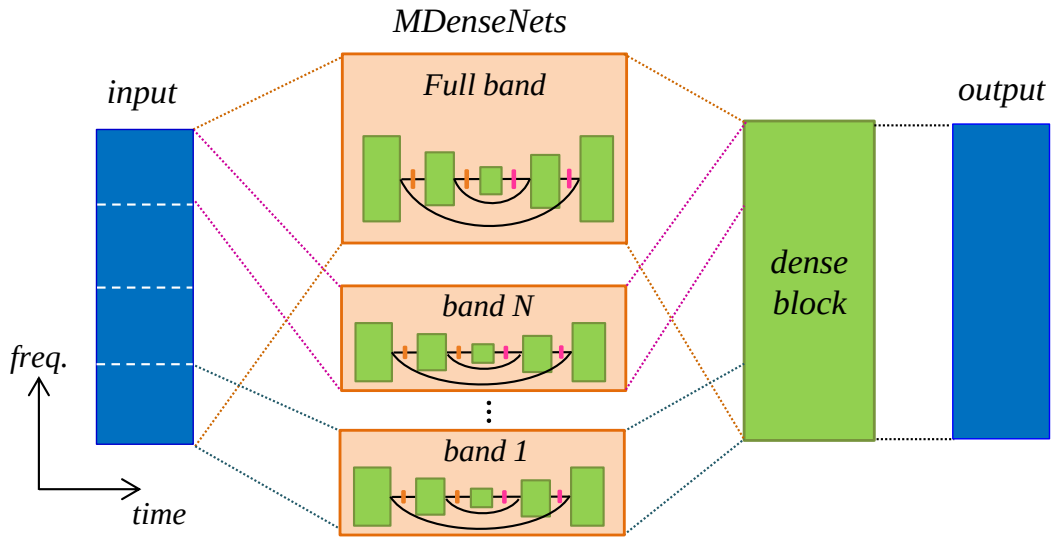


図 2.3: MMDenseNet([7])

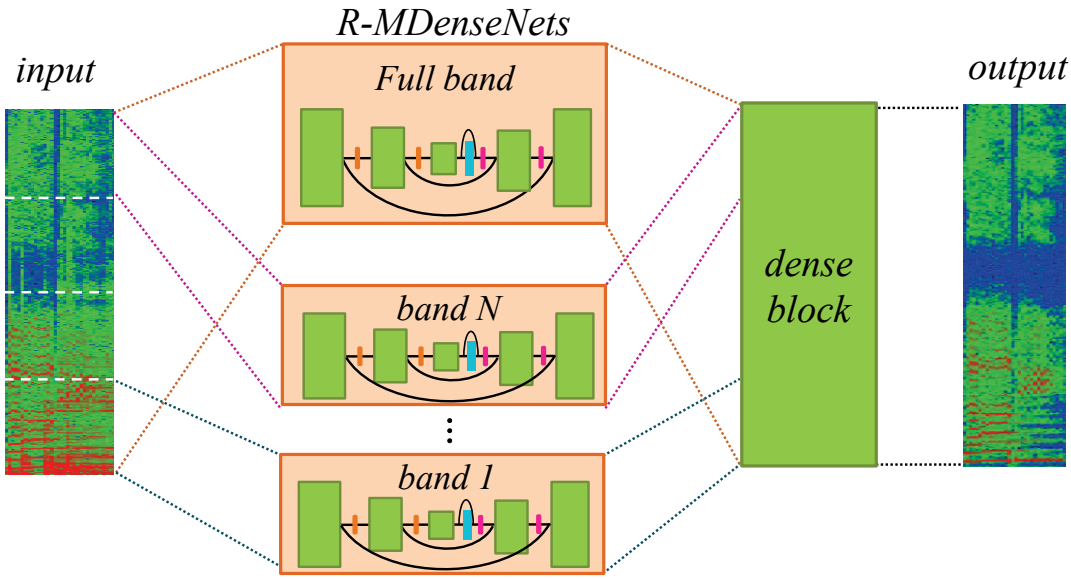


図 2.4: MMDenseLSTM([9])

2.3　Self-Attention Dense-UNet

Liu らは MMDenseNet および MMDenseLSTM [7,9] をさらに発展させ，LSTM の代わりに自己注意機構を用いた，Self-Attention Dense-UNet を提案した [10].

楽曲を構成する音のうち，ボーカル音声は絶えず変化する歌詞を歌い，その他の楽器は比較的短いパターンを繰り返す傾向にある．そのため，ある時間セグメントにおけるパターンを他の箇所に見つけられれば，互いを参照することで楽器音分離の精度

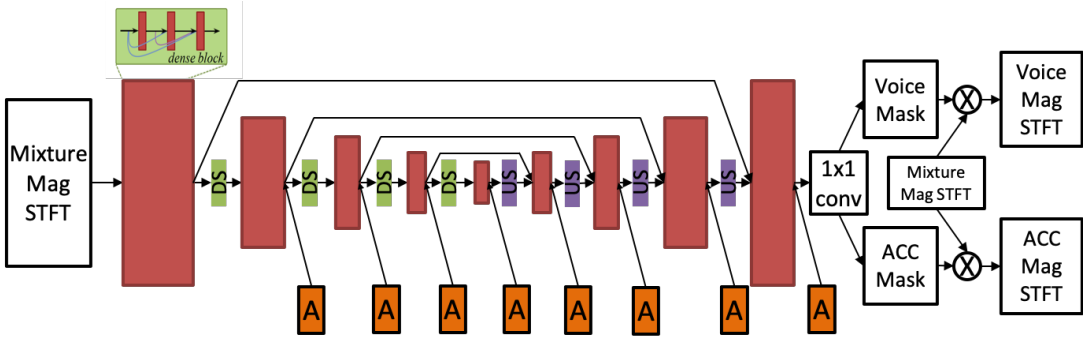


図 2.5: Self-Attention Dense-UNet([10])

が向上することが期待できる．しかし，このような楽器のリズムや繰り返しは比較的長い時間的コンテキストを考慮しなければならず，回帰型ニューラルネットワーク (RNN) や LSTM などでは大域的な関係性を学習しきれないという問題があった [10].

2.3.1　自己注意

自己注意 (Self-Attention) は自然言語処理の分野で提案され，既存の RNN や LSTM, CNN を置き換える形で非常に高い性能を発揮しているモデルである [11]．同モデルは RNN および LSTM よりも長期的な依存関係の学習が可能であるため，音源分離の精度向上に効果がある [10, 12].

注意機構 (Attention) は

- Query
- Key
- Value

の 3 つの行列を入力にとり，Query と Key の行列積に softmax 関数を適用して各行に関して重み付けを行ない，入力のどの部分に注意すべきかを決定して，最後に Value に掛けることで重み付き出力を得る (式（2.2）).

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.2}$$

ここで Q，K，V はそれぞれ Query，Key，Value，d_k はスケーリング項である．

また注意機構は入力の性質によって

- Source-Target Attention: Query と Key-Value に異なる値を入力
- Self-Attention: Query と Key-Value に同じ値を入力

の 2 種類に大別できる．自己注意はある系列自身の特徴を学習する機構だと言え

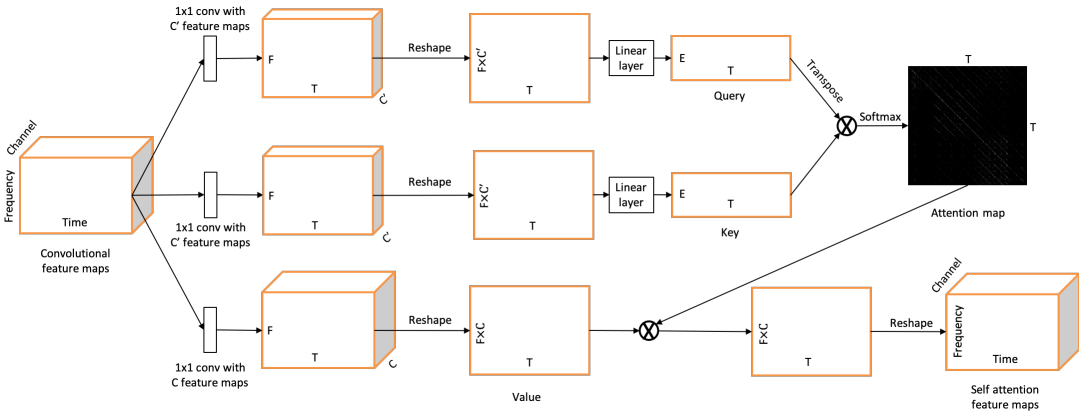


図 2.6: 自己注意機構 ([10])

る [11]．[10] で提案された，畳み込み層と全結合層による，時間方向への自己注意機構を図 2.6 に示す．

自己注意層では 1 つの入力を Query，Key，Value の 3 つに分岐して，それぞれを 1×1 畳み込み層で変形する．このとき，Query と Key のチャンネル数は C から C' へと削減される．さらに各特徴量を変形して，$T \times C'F$ または $T \times CF$ のサイズにする．T は時間，F は周波数方向のサイズを指す．全結合層に Query と Key を通してチャンネルと周波数の積を固定の埋め込み次元 E まで削減して，それぞれ $T \times E$ の行列を得たのち，それらの行列積を計算することで自己注意マップを計算する．自己注意マップは $T \times T$ であり，どの時間フレーム間に関係があるかをマッピングしていると考えられる．

最後に Value と自己注意マップの行列積を取り特定のフレームを強めたり弱めたりしたうえで，入力のサイズへと復元して出力する．[10] では，この出力は入力と結合して次の層へと伝播するよう提案されている．

第 3 章　提案手法

Self-Attention Dense-UNet は，multi-scale な DenseNet に図 2.6 で示した自己注意を組み合わせることで，入力の時間的なコンテキストを LSTM よりも広範囲にわたって学習するモデルであった．また，MMDenseLSTM は Multi-band 構造によって楽曲の周波数的な特性を効率的に学習する．

本論文で提案するモデルは，時間的な特徴の学習に長けた自己注意を用いたモデルである Self-Attention Dense-UNet を，MMDenseLSTM の Multi-band 構造に倣って周波数帯域ごとに並列化する．提案するモデルの各帯域におけるネットワーク構造および並列化したモデルを，図 3.1，図 3.2 に示す．

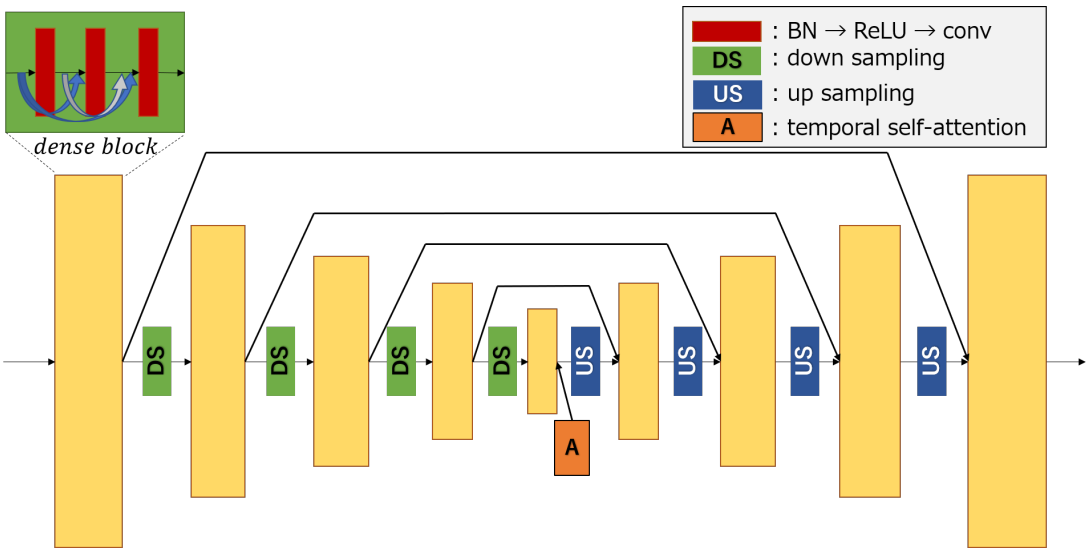


図 3.1: 各帯域のネットワーク構造

3.1　Self-Attention Dense-UNet の問題点

自己注意の基幹となる処理は，系列内における注目度の導出である．これは 2.3.1 で示したように，Query と Key との行列積と softmax 関数による自己注意マップを導出したのち Value との行列積を計算することで得られる [10]．このとき自己注意マップは時間 × 時間のような 1 つの次元にのみ着目した関係性のマッピングを行ない，強調あるいは減衰させる．すなわち，例えば図 3.3 のようにベースギターが鳴っている

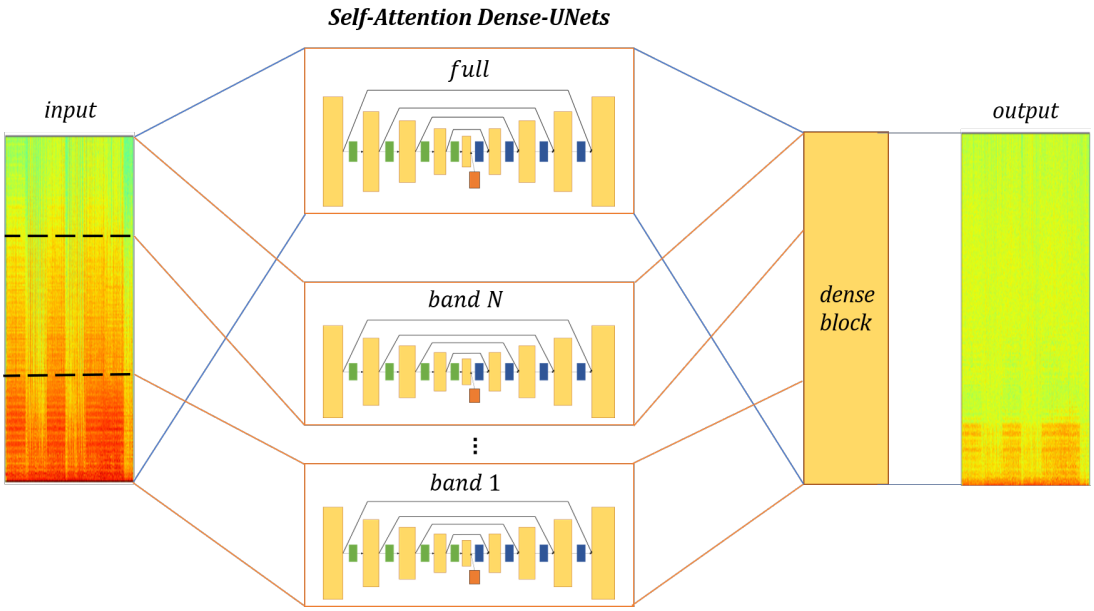


図 3.2: 提案するモデル全体の構造

時間に注目したとき，そのフレームを全周波数にわたって強調するため，本来強調したい低周波帯域だけでなくベースギターとは関係ない高周波帯域まで強調してしまい，結果として同時に鳴っている他の楽器音が混ざってしまい分離性能が低下するということが考えられる．

高橋ら [7,9] による Multi-band 構造は，図 3.4 で示すように分離対象となる音源の周波数特性に基づいて各ネットワークが学習する帯域を制限するため，余計な帯域まで強調してしまうことを防げると考えられる．また 2.1.2 で述べたような MMDenseNet および MMDenseLSTM [7, 9] で提案された Multi-band 構造の本来的な目的である，音の周波数帯域に特有なパターンに対する畳み込み層の学習も帯域毎に独立して行えるため，Self-Attention Dense-UNet 全体の学習の効率化が期待できる．

3.2　周波数帯による Self-Attention モデルの並列化

ネットワークへの入力は，各楽曲の時間波形を短時間フーリエ変換したスペクトログラムであり，チャンネル × 周波数 × 時間の次元を持っている．この入力特徴量を周波数によって分割して，帯域ごとに独立したネットワークに学習させる．また，入力を分割せずそのまま学習する全帯域用のネットワークも用意する．つまり，入力を N 個の帯域に分割するとすれば，計 $N+1$ 個のサブネットワークを同時に学習することになる．各帯域ごとのネットワーク出力 N 個を周波数方向に結合して入力と同じだけの周波数を含む特徴量へと復元して，さらに全帯域用ネットワークの出力とチャンネ

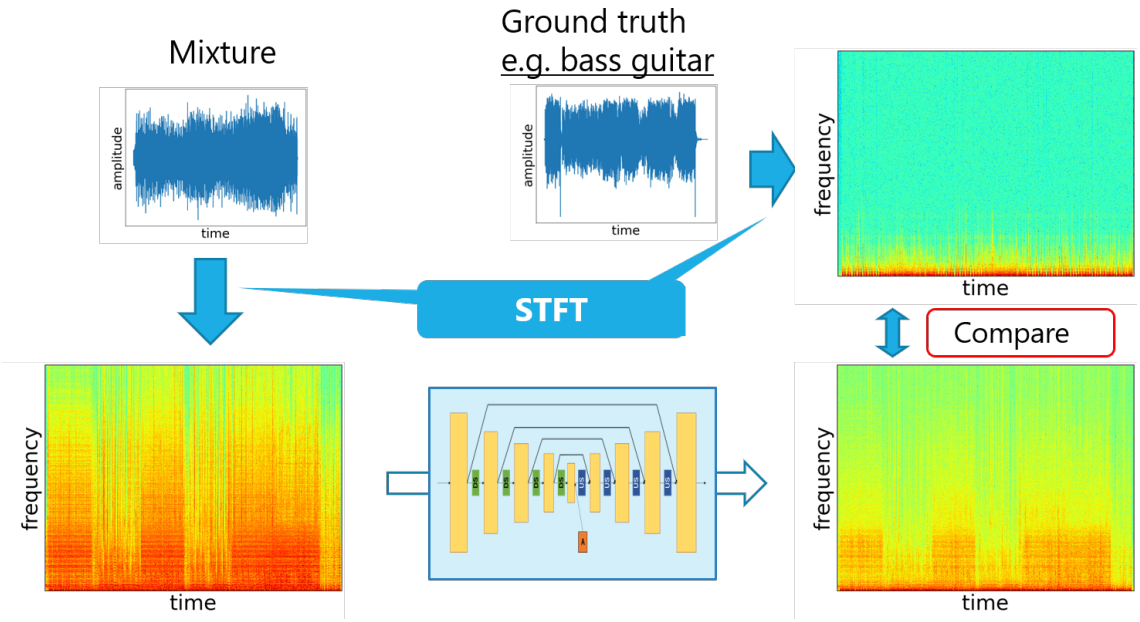


図 3.3: 分割を行なわない場合

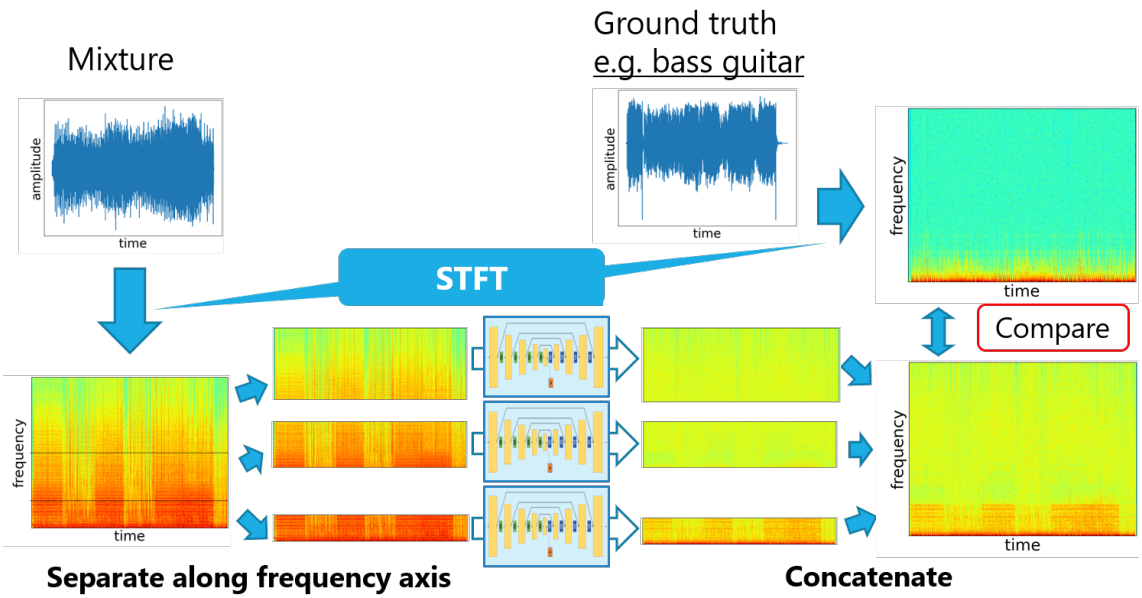


図 3.4: 周波数帯域による分割

ル方向に結合する．最後に，すべてのサブネットワークの出力が結合した特徴量を畳み込み層に通すことでチャンネル数を調整して，入力と同じ次元にして出力する．このような Multi-scale Multi-band 構造を持った Self-Attention Dense-UNet を，分離する楽器それぞれについて学習する．

3.3　自己注意の挿入箇所の変更

ただし，自己注意は入力の行列積を計算するにあたって，要求するメモリ量が入力特徴量の大きさに強く依存する．LSTM も同様に系列長の増加による計算量の増大が問題点として挙げられるが，高橋らは [9] において LSTM の挿入箇所について検証を行なっており，ダウンサンプリング層によって入力特徴量が小さくなった箇所にのみ LSTM を挿入することでメモリ量や計算量を抑えることができ，かつ分離精度もあまり低下しないという結果を示している．Liu ら [10] は図 2.5 で示したように最初の 1 層目を除くすべてのスケールにおいて自己注意を挿入していたが，本論文で提案するモデルでは高橋らと同様に，スケールが十分小さくなった箇所にのみ自己注意を挿入することで，要求するメモリ量を抑える．

第 4 章 実験

4.1 実験条件

4.1.1 モデル構造の詳細

モデルの各パラメータの詳細を，表 4.1 に示す．

表 4.1: モデルの各パラメータ

band	k	scale	d1	d2	d3	d4	d5	u4	u3	u2	u1
low	14	l	5	5	5	5	-	-	5	5	5
		e	-	-	-	10	-	-	-	-	-
mid	4	l	4	4	4	4	-	-	4	4	4
		e	-	-	-	10	-	-	-	-	-
high	2	l	1	1	-	-	-	-	-	1	1
		e	-	-	10	-	-	-	-	-	-
full	7	l	3	3	4	5	5	5	4	3	3
		e	-	-	-	-	20	-	-	-	-

ここで，k は成長率と呼ばれる dense block 内の畳み込み層の出力チャンネル数，l は dense block 内の畳み込み層の数，e は自己注意層の埋め込み次元数を表している．つまり，成長率 k，層数 l の dense block を通過すると，特徴量のチャンネル数は $k \times l$ だけ増加する．また，各自己注意層の中間チャンネル数 C' は 5 とした．

low， mid， high の各帯域はそれぞれ $0\tilde{4}.1$，$4.1\tilde{1}1$，$11\tilde{2}2.5$[kHz] である．また，スケールに’d’ とつく層は，dense block および自己注意層の後にダウンサンプリング層を通して特徴量の時間/周波数サイズを縮小し，’u’ とつく層ではアップサンプリング層を通してサイズを復元する．

4.1.2　学習条件

最適化アルゴリズムには RAdam [13] を用い，学習係数は 1.0×10^{-3} とした．また損失関数には平均二乗誤差を用いた．バッチサイズは 1 で，学習エポック数は 50 とした．

4.1.3　データセット

本論文では，モデルの学習および評価のために MUSDB18 [14] を用いた．同データセットは 150 曲を含み，各楽曲は vocal，bass，drums，other の 4 つの楽器音のトラックからなる STEM ファイル形式で提供されている．またすべての楽曲は 2 チャンネルのステレオ音源で，44.1[kHz] でサンプリングされている．本論文では 86 曲を学習に，14 曲を学習中の確認 (バリデーション) に，50 曲を評価に使用した．さらに学習とバリデーションのデータには 2 種類のデータ拡張を施した．1 つはランダムに 0.25-1.25 倍の範囲で振幅を増減するもので，もう 1 つはランダムにチャンネル間を交換するものである [6].

4.1.4　前処理

時間波形を短時間フーリエ変換 (STFT) し，絶対値を取って振幅スペクトログラムへ変換して入力特徴量とした．窓長は 4096 サンプル，シフト長は 25% の 1024 サンプルで，窓関数は hann 窓とした．これによって周波数ビン数は 2049 となる．また一度にネットワークへ入力する特徴量の時間フレーム長は，リズムや楽器の繰り返しを学習できるよう 10 秒程度を含むように，512 とした．

4.1.5　後処理

モデルから出力した各楽器音の振幅スペクトログラム推定値を用いて多チャンネルウィーナーフィルタ (MWF) [2, 6] を作成する．混合音源の振幅スペクトログラムを作成した MWF でフィルタリングすることで出力となる振幅スペクトログラムを計算し，混合音源の位相スペクトログラムとあわせて逆短時間フーリエ変換 (ISTFT) することで時間波形へと戻す．

4.1.6　評価指標

評価には，客観的指標として SDR(Signal-to-Distortion Ratio) を用いた．SDR は推定された信号 $\hat{S}$ と教師信号 S について

$$SDR(\hat{s}, s) := 10 \log \frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \tag{4.1}$$

と定義され，s_{target} は $\hat{s}$ のうちの信号成分，e_{interf}，e_{noise}，e_{artif} はそれぞれ干渉，雑音，アーティファクトによる歪み成分の大きさを表す．SDR が大きいほど教師信号との類似度が高く，よりよい分離結果が得られていることを示す．

また SDR は，データセット内の各楽曲に対して，かつそれぞれの曲を特定の時間で区切った範囲ごとに計算するが，1 曲ごとの中央値を計算し，さらにそれらの中央値を計算してデータセット全体に対する評価指標とした．

4.2　実験結果

提案手法と，既存手法である MMDenseLSTM における実験結果の比較を表 4.2 に示す．Ideal Binary Mask は理想 2 値マスクと呼ばれる既知の分離音源に基づくフィルタを用いた結果であり，mixture は混合音源を全く処理せず各楽器音と比較した結果である．

表 4.2: 実験結果

Methods	#params[$\times 10^6$]	SDR in dB			
		vocals	bass	drums	other
UB(Ideal Binary Mask)	-	11.1	7.84	8.30	8.90
MMDenseLSTM([9])	1.1	5.686	4.485	5.555	3.568
proposed	0.28	6.182	4.966	5.575	3.769
LB(mixture)	-	-0.28	-0.19	0.005	0.49

表 4.2 より，すべての楽器において SDR の向上を確認した．drums の向上は 0.02dB 程度であるのでほぼ同等であると言えるが，other は 0.2dB，vocals や bass については 0.5dB 程度の向上が確認でき，人間の耳で判別できるレベルの差であると言える．またネットワークのパラメータ数は MMDenseLSTM [9] と比較して 1/3 倍程度まで少なくなり，学習および推論時に要求するメモリ量も削減できた．ただし DenseNet 構造

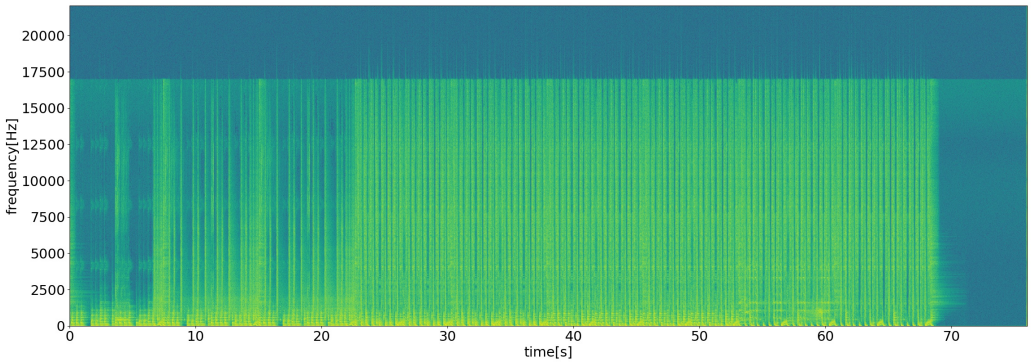
は単一の畳み込み層の重みや出力を複数回にわたって使用するために勾配情報などもその都度増加するので，学習時に必要なメモリ量はいずれもシーケンシャルなネットワークに対しては比較的大きい．

また図 4.1 から図 4.8 にかけて，各楽器に関して最も SDR に改善がみられた楽曲と，逆に最も SDR が低下した楽曲のスペクトログラムを示す．

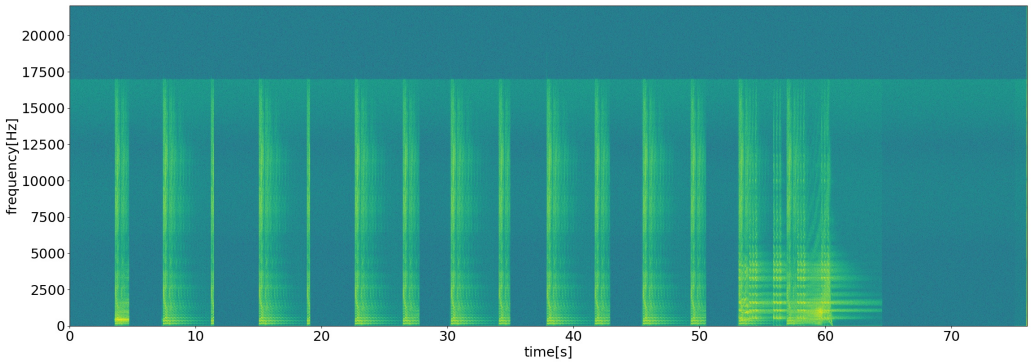
track30 は全体的に顕著な SDR の向上がみられた．この楽曲は比較的単純なフレーズを繰り返すものであったため，時系列の特徴を読み取るのが得意な自己注意がうまく作用してリズムパターンを学習出来たと考えられる．

逆に SDR が悪化した楽曲について， track28 は vocals が最も悪化しており，これは図 4.2b のように楽曲がほぼインストゥルメンタルに近く，サンプリングされた音声が特定のパターンで楽器のように繰り返されているため，自己注意による学習結果がうまく作用しなかったと考えられる．また track13 の bass は他の楽曲と比較して多彩なメロディに変化しており，時間的な変化が一定ではない．そのため自己注意が学習した bass のパターンから外れて，SDR が低下したと予想する．drums が最も悪化した track9 はロック調の楽曲だが，ドラムパターンが比較的短いスパンで変化したり，特徴的なフィルインが用いられたりしており，これも学習したパターンから外れていると考えられる．そして other については track34 が最も SDR が低下したが，楽曲がラップ調かつビートにサンプリングした悲鳴のような音声が使われており，時間的特徴が vocals と区別しづらかったことが精度低下の原因だと考えられる．

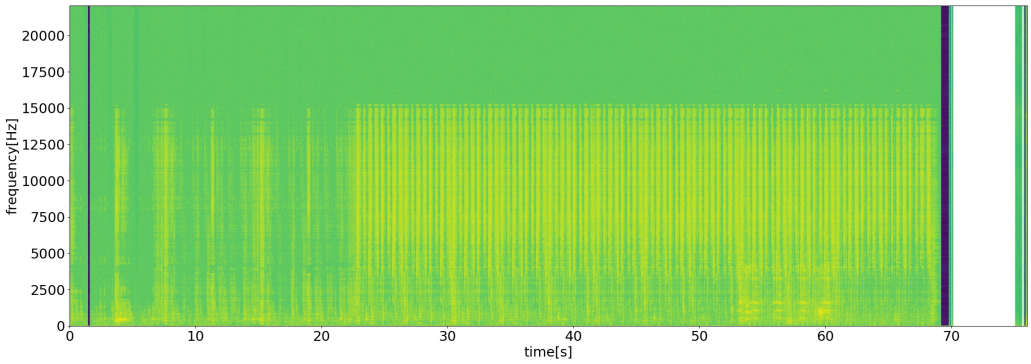
これらの結果から，自己注意を用いるねらいであったボーカルと楽器のリズムやフレーズの違いという前提に沿った，絶えず変化するボーカルと比較的短いフレーズの繰り返しである楽器の音という組み合わせ [10] の楽曲については LSTM を用いたモデルよりもうまく分離ができ，前提から大きく逸脱した楽曲については分離がしづらくなると言える．



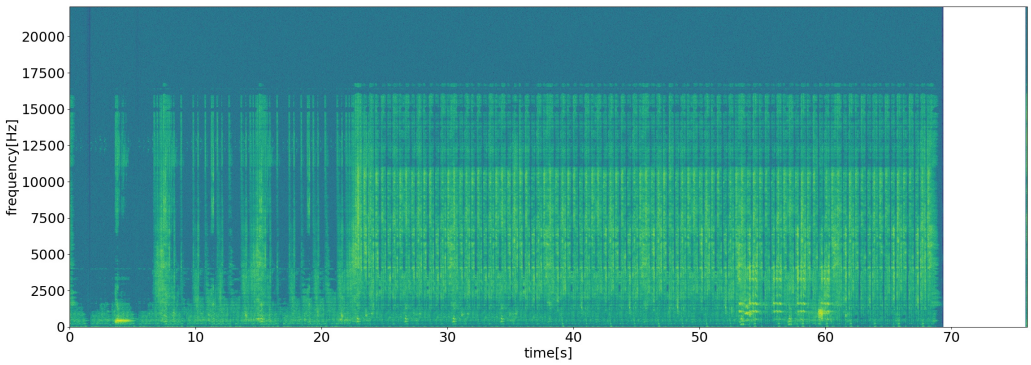
(a) mixture



(b) source

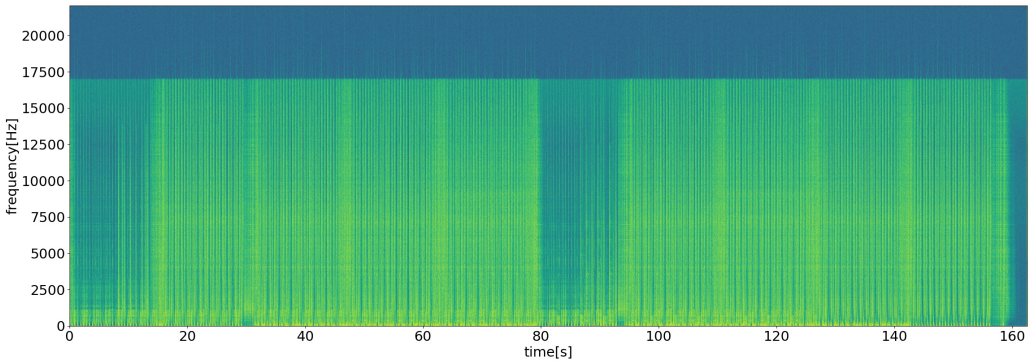


(c) baseline

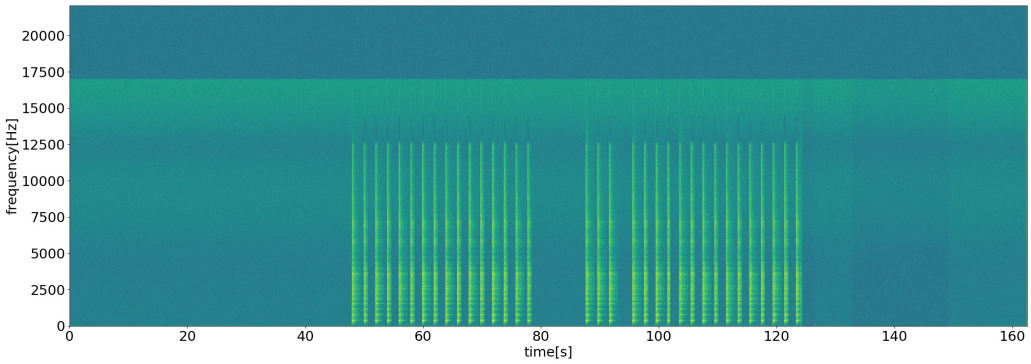


(d) proposed

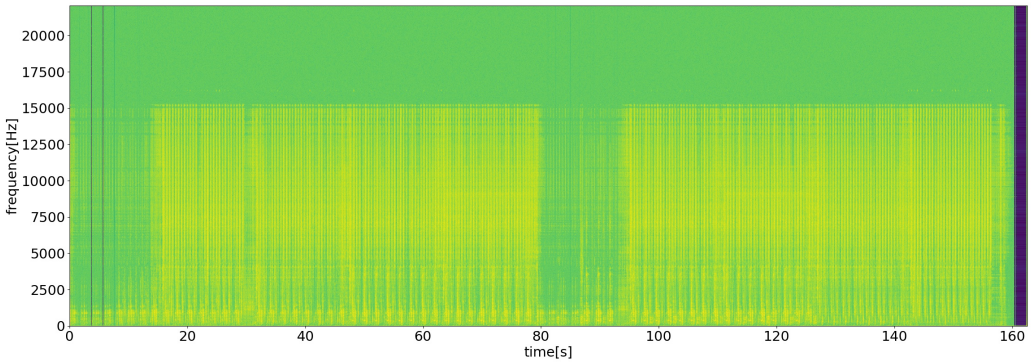
図 4.1: track29-vocals



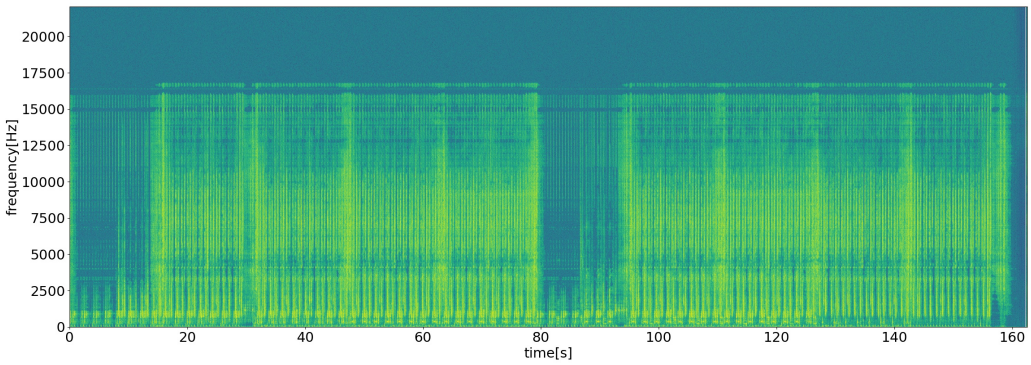
(a) mixture



(b) source

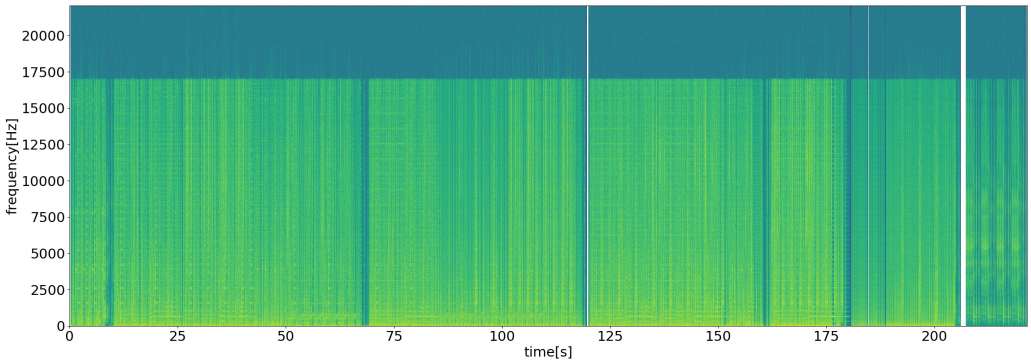


(c) baseline

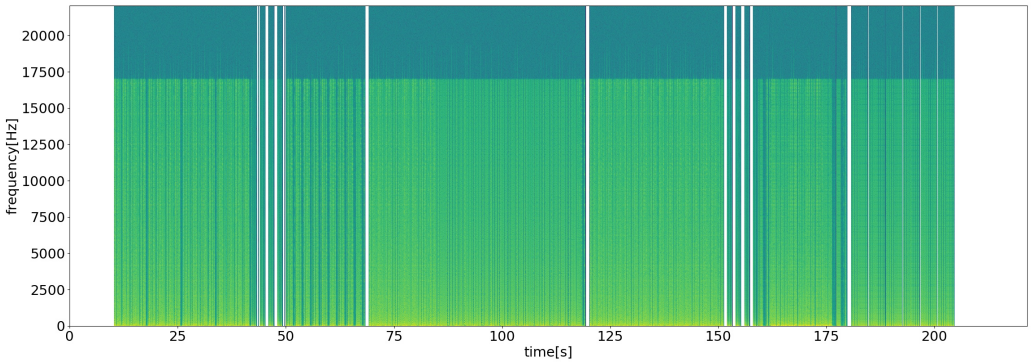


(d) proposed

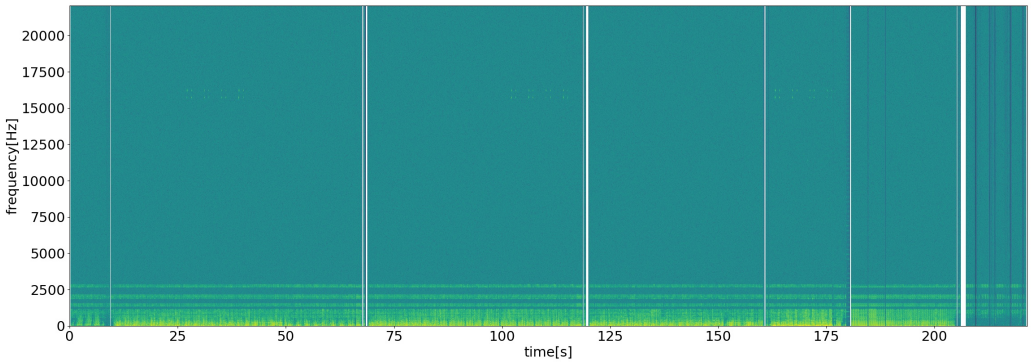
図 4.2: track28-vocals



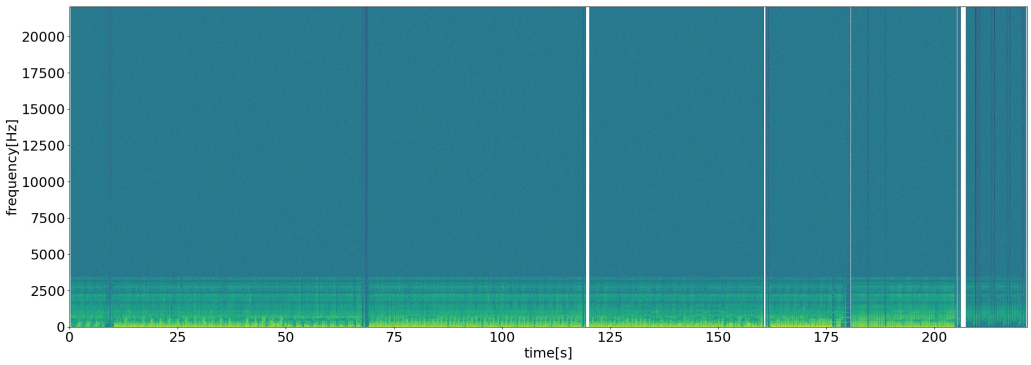
(a) mixture



(b) source

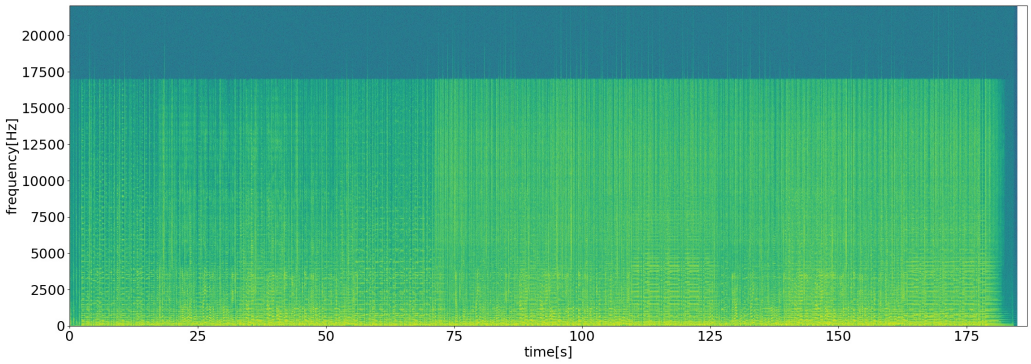


(c) baseline

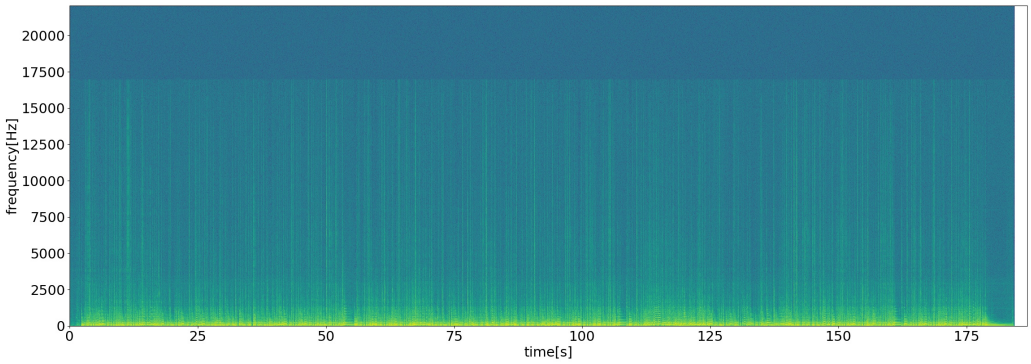


(d) proposed

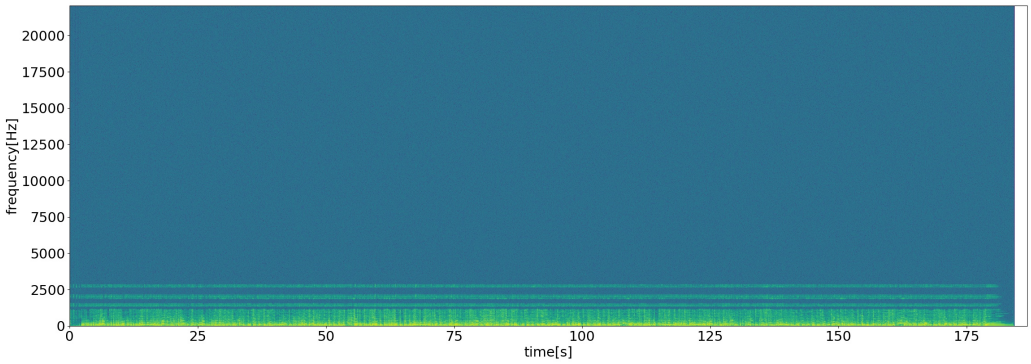
図 4.3: track30-bass



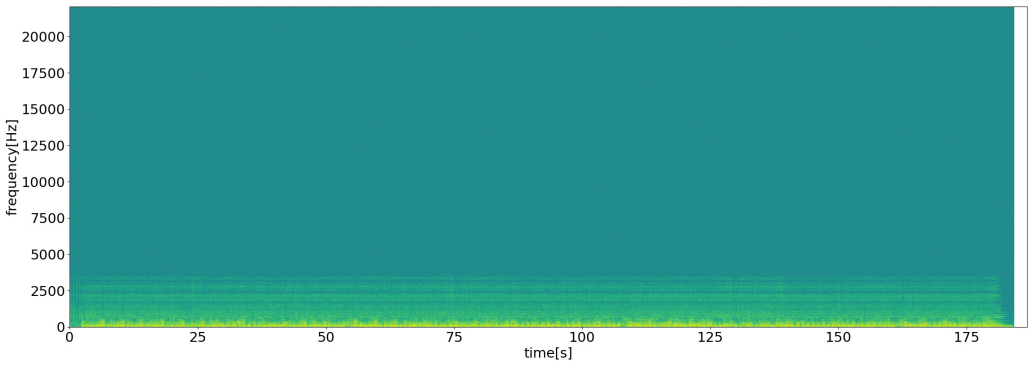
(a) mixture



(b) source

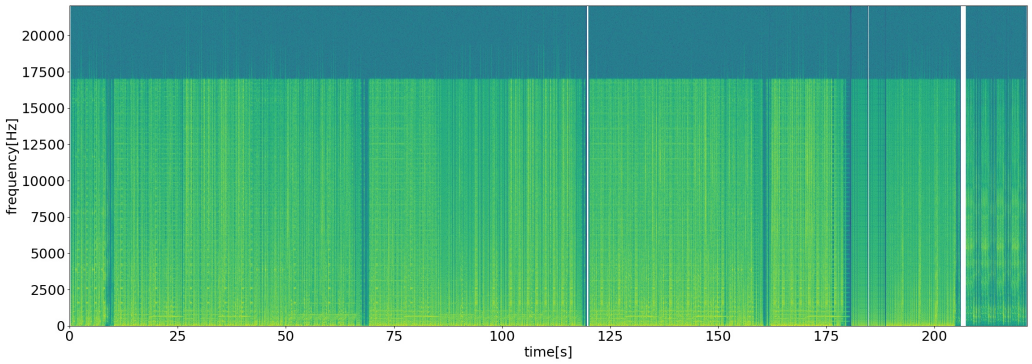


(c) baseline

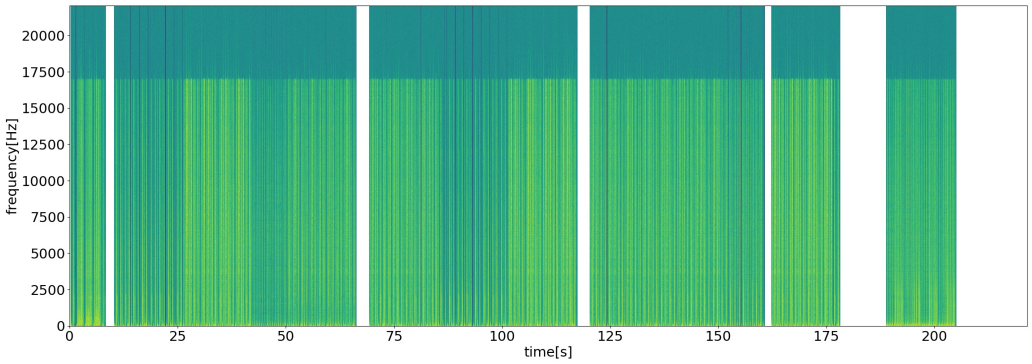


(d) proposed

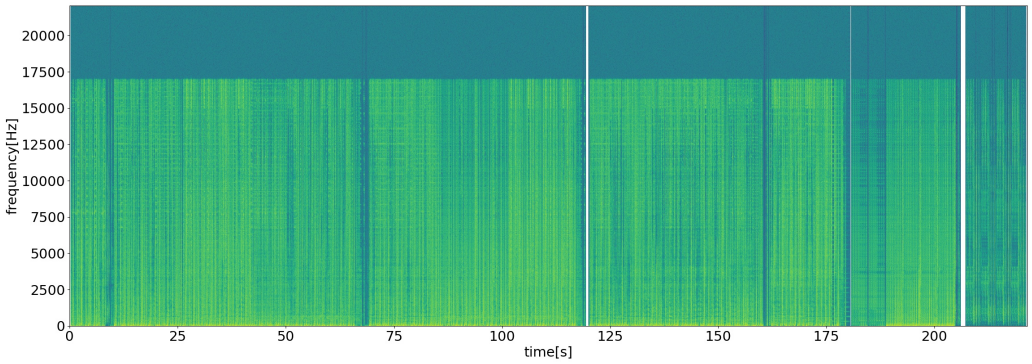
図 4.4: track13-bass



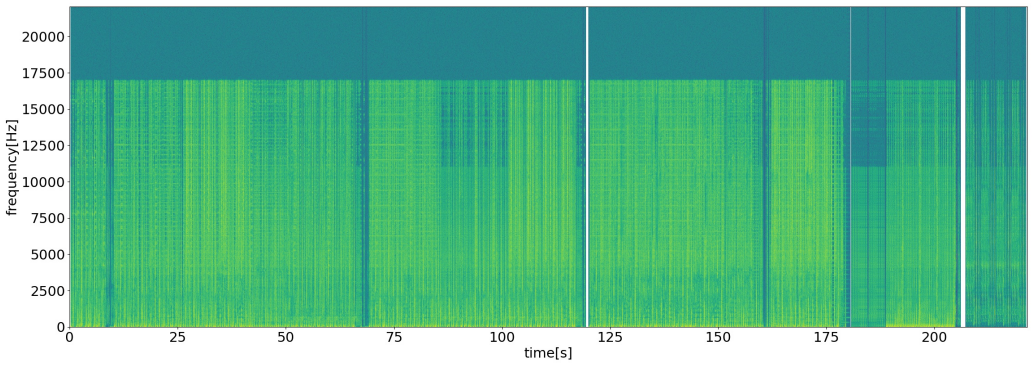
(a) mixture



(b) source

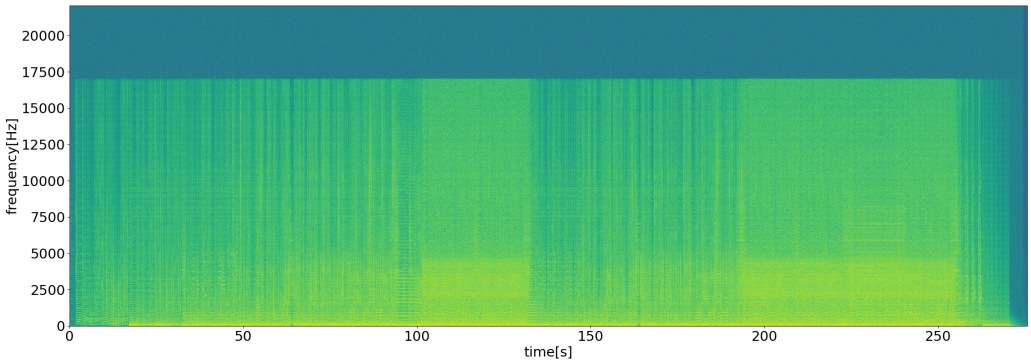


(c) baseline

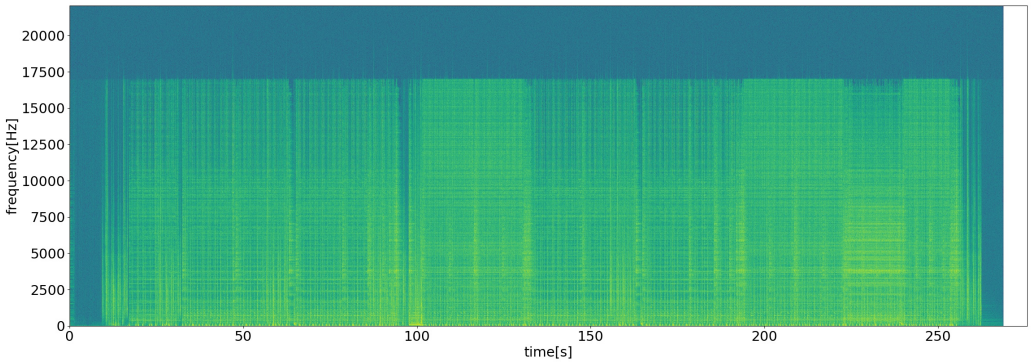


(d) proposed

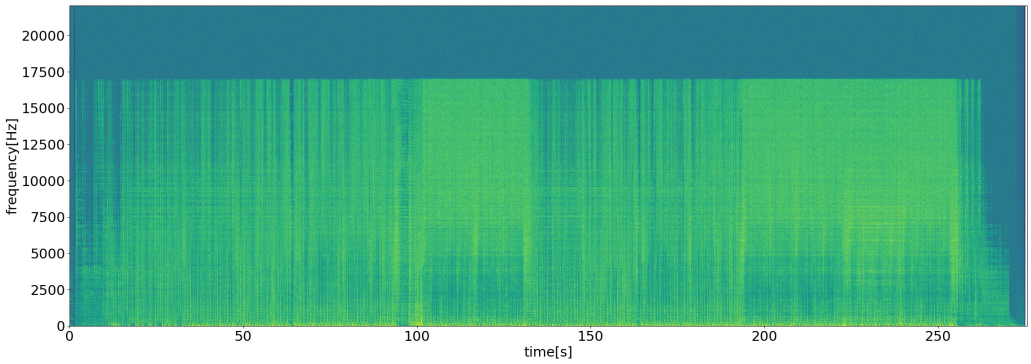
図 4.5: track30-drums



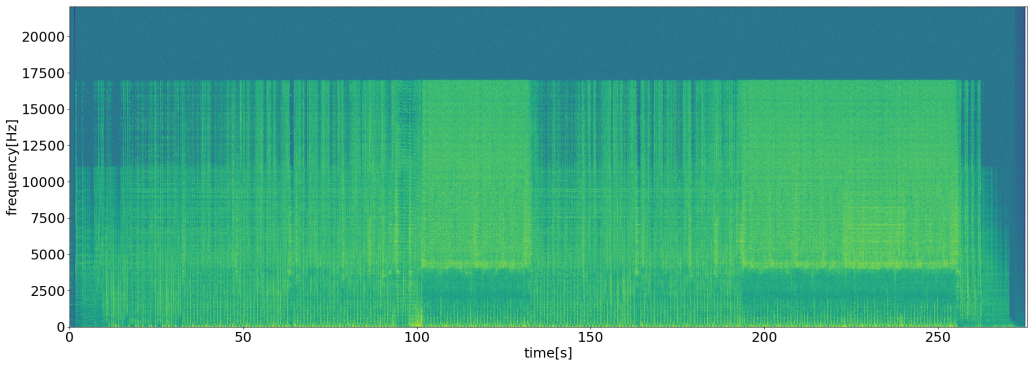
(a) mixture



(b) source

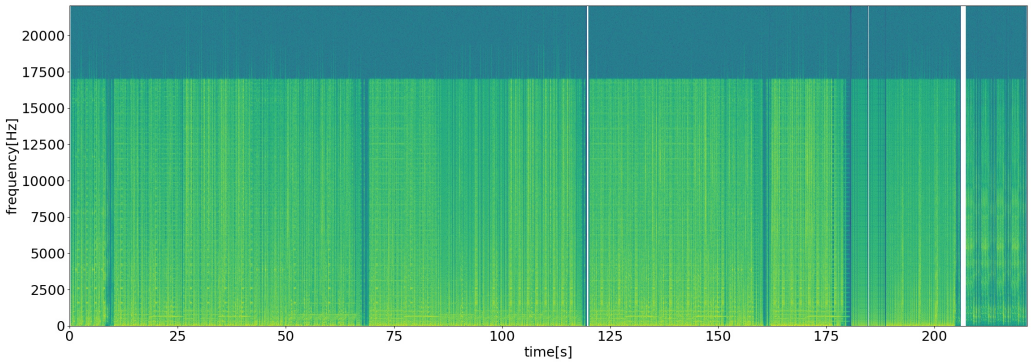


(c) baseline

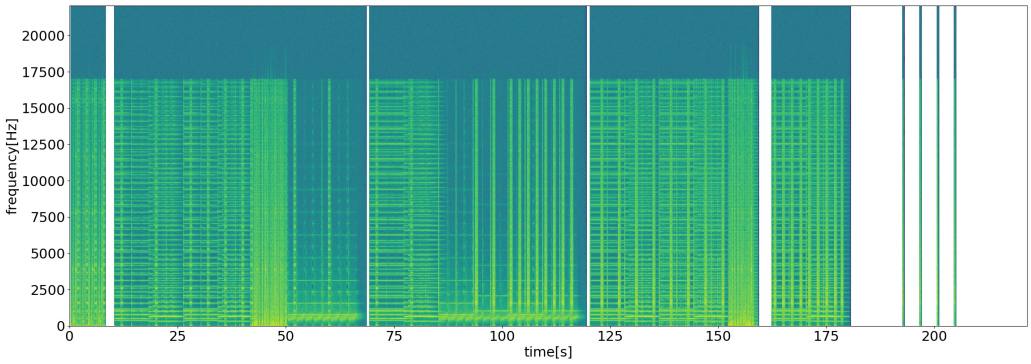


(d) proposed

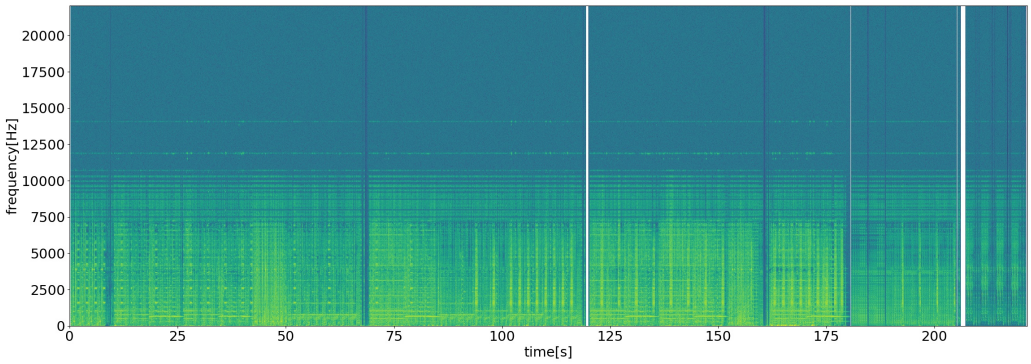
図 4.6: track9-drums



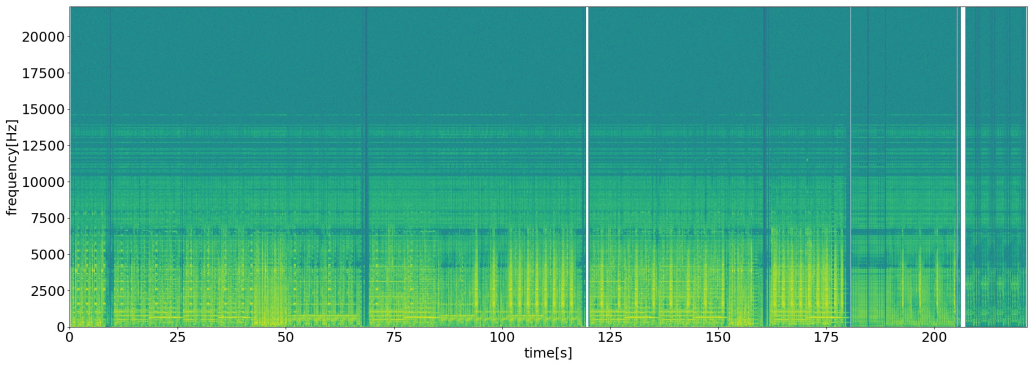
(a) mixture



(b) source

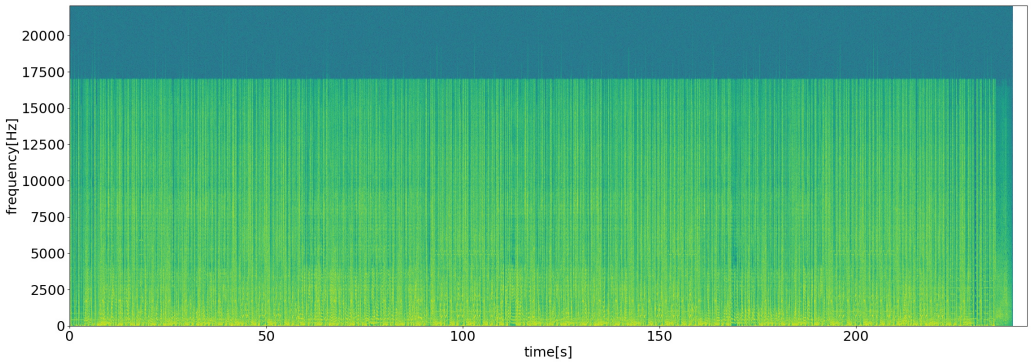


(c) baseline

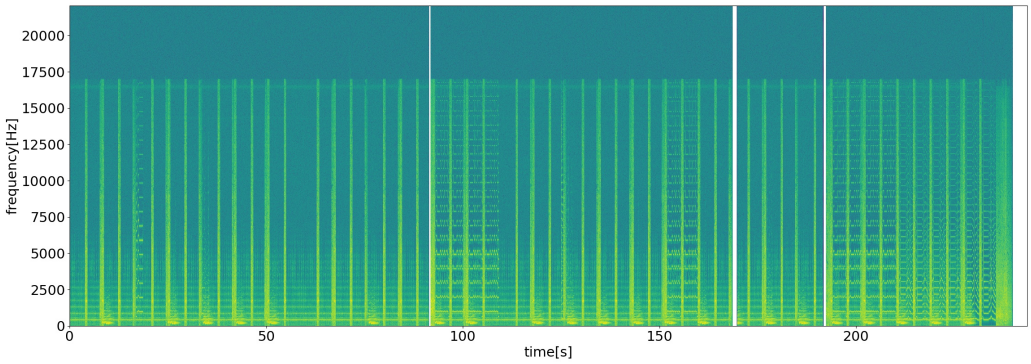


(d) proposed

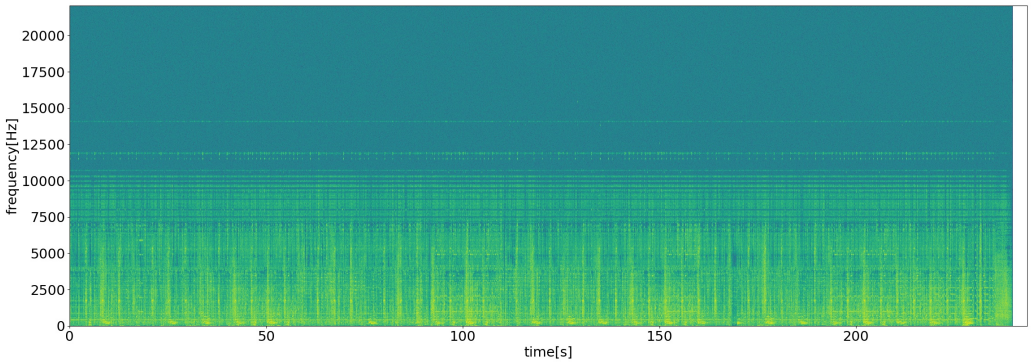
図 4.7: track30-other



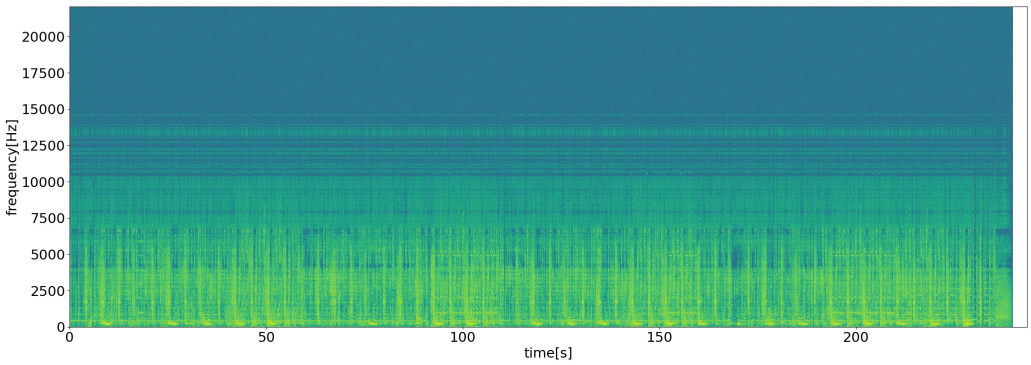
(a) mixture



(b) source



(c) baseline



(d) proposed

図 4.8: track35-other

第 5 章 おわりに

5.1 まとめ

本論文では，楽曲の時間的・周波数的特徴の双方について学習精度を向上させることを目的とし，自己注意機構と Multi-scale な DenseNet を組み合わせたモデルに Multi-band 構造を導入した手法を提案した．第 1 章では近年の楽器音分離の動向および既存手法の性質について言及し，本論文の目的および提案手法を示した．第 2 章では提案手法のベースとなる MMDenseNet，MMDenseLSTM および Self-Attention Dense-UNet について述べた． 第 3 章では提案手法である DenseNet と自己注意機構を用いたモデルの並列化について述べ，構造を示した．第 4 章では MUSDB18 データセットを用いた既存手法と提案手法との比較実験を行なった．

結果としてすべての楽器において精度の向上が確認できた．特に楽器ごとにフレーズの長さにばらつきがある楽曲については顕著な精度の向上がみられ，自己注意機構が得意とする時間的特徴の学習が楽器音分離において有効であることを確認した．

5.2 今後の展望

Self-Attention Dense-UNet が時間方向にのみ自己注意を計算していることから，本論文では性能を向上すべくネットワークの並列化を提案したが，他にも周波数方向にも自己注意を計算するなどの手法でも性能の向上が見込める．また，田野崎ら [15] は楽曲のメタ的な特徴量であるジャンルを挿入することで，分離の精度が向上することを示しており，同様の手法を自己注意モデルに適用することでも結果の改善が期待できる．

謝辞

本論文の執筆を行なうにあたって，多くのご指導を頂いた杉田泰則准教授に厚くお礼申し上げます．また，本論文の審査にあたり適切なご指摘，ご指示を下さった岩橋政宏教授，圓道知博教授ならびに原川良介助教に，心より感謝を申し上げます．

最後に，本論文にかかる研究について，ゼミ発表や研究室での会話の中で多くのアドバイスを頂いたり，研究が行き詰った際に相談に乗ったりして下さった信号処理応用研究室の皆様へ深く感謝の意を示して，謝辞とさせていただきます．

令和 5 年 2 月

参考文献

[1] J. Le Roux, J.R. Hershey, and F. Weninger, “Deep nmf for speech separation,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.66–70, 2015.

[2] N.Q.K. Duong, E. Vincent, and R. Gribonval, “Under-determined reverberant audio source separation using a full-rank spatial covariance model,” IEEE Transactions on Audio, Speech, and Language Processing, vol.18, no.7, pp.1830–1840, 2010.

[3] 北村大地，角野隼斗，高宗典玄，高道慎之介，猿渡　洋，小野順貴，“独立深層学習行列分析に基づく多チャネル音源分離の実験的評価,” 信学技報，pp.13–20, 2018.

[4] A.A. Nugraha, A. Liutkus, and E. Vincent, “Multichannel music separation with deep neural networks,” 2016 24th European Signal Processing Conference (EUSIPCO), pp.1748–1752, 2016.

[5] S. Uhlich, F. Giron, and Y. Mitsufuji, “Deep neural network based instrument extraction from music,” 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2135–2139, 2015.

[6] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, “Improving music source separation based on deep neural networks through data augmentation and network blending,” 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.261–265, 2017.

[7] N. Takahashi and Y. Mitsufuji, “Multi-scale multi-band densenets for audio source separation,” 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp.21–25, 2017.

[8] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, “Densely connected convolutional networks,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2261–2269, 2017.

[9] N. Takahashi, N. Goswami, and Y. Mitsufuji, “Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation,”

2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), pp.106–110, 2018.

[10] Y. Liu, B. Thoshkahna, A. Milani, and T. Kristjansson, “Voice and accompaniment separation in music using self-attention convolutional neural network,” 2020. https://arxiv.org/abs/2003.08954

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems, eds. by I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol.30, pp.••–••, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[12] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.21–25, 2021.

[13] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020), pp.••–••, April 2020.

[14] Z. Rafii, A. Liutkus, F.-R. Stöter, S.I. Mimilakis, and R. Bittner, “The MUSDB18 corpus for music separation,” Dec. 2017. https://doi.org/10.5281/zenodo.1117372

[15] 田野崎蓮，杉田泰則，“楽曲情報を考慮した dnn ベース楽器音分離モデルに関する一考察,” 電気学会研究会資料，CT-21-036-043 制御研究会，pp.7–11，2021.