

周波数領域でのフィルタ設計に基づく サンプリング周波数非依存畳み込み層を用いた DNN 音源分離

齋藤 弘一¹ 中村 友彦¹ 矢田部 浩平² 猿渡 洋¹

概要: 音源分離は、様々なアプリケーションの前処理として利用される。そのため、前処理として汎用的に利用できる音源分離の実現には、各アプリケーションで要請される様々な音響条件下で一貫して動作する音源分離モデルが必要である。これに対し、我々はこれまでに、代表的な音響条件の1つであるサンプリング周波数に非依存な畳み込み層（サンプリング周波数非依存畳み込み層）を提案し、それを用いた深層ニューラルネットワーク（deep neural network: DNN）に基づく音源分離モデルを構築した。サンプリング周波数非依存畳み込み層では、畳み込み層をデジタルフィルタとみなし、アナログフィルタから畳み込み層の重みを生成する仮定をデジタルフィルタ設計と捉えることで、サンプリング周波数に不変な構造を導入する。この層では、これまでアナログフィルタをサンプリング周期でサンプルすることでデジタルフィルタを設計していた。しかし、このフィルタ設計手法では、Nyquist 周波数よりも高い成分をもつアナログフィルタに関してエイリアシングが起これ、これが主たる要因となり、学習したサンプリング周波数よりも低いサンプリング周波数において分離性能が低下した。そこで、本研究ではフィルタ設計におけるエイリアシングを回避するため、最小二乗法に基づく周波数領域でのフィルタ設計手法を導入する。提案法をサンプリング周波数非依存畳み込み層に導入することにより、学習データとは異なるサンプリング周波数においても学習したサンプリング周波数と同程度の分離性能が得られることを楽音分離実験により示す。

キーワード: シングルチャネル音源分離, サンプリング周波数非依存畳み込み層, 最小二乗近似, ディープニューラルネットワーク

1. はじめに

シングルチャネル音源分離はモノラル音響信号から、各音源の個別の信号を抽出する技術である。近年、深層ニューラルネットワーク（deep neural network: DNN）を用いたシングルチャネル音源分離モデルが多数提案され、高い分離性能を示している [1–7]。音源分離は様々なアプリケーションの前処理として用いられるため、DNN を用いることで高性能な分離性能を活かしつつ、様々なアプリケーションで要請される音響条件で動作することが望ましい。

代表的な音響条件の1つがサンプリング周波数である。アプリケーションに応じて必要なサンプリング周波数は異なる。例えば、分離音が人間が聴取するアプリケーションにおいては、多くの場合、可聴域を網羅するため 44.1 kHz や 48 kHz が用いられる [7, 8]。一方、音響信号に含ま

れる内容の認識を目的としたアプリケーションでは、必ずしも可聴域をカバーしなくともよい。例えばビートトラックキングでは 16 kHz [9]、自動採譜では 11.025 kHz や 22.05 kHz [10, 11]、音声認識では 8 kHz や 16 kHz が使用されている [12–14]。様々なアプリケーションの前処理として、汎用的に利用可能な音源分離モデルを構築するには、これら様々なサンプリング周波数の音響信号に対して一貫して動作する DNN を用いた音源分離モデル（DNN 音源分離モデル）を構築する必要がある。

しかし、従来の DNN 音源分離モデルは、サンプリング周波数毎に別々のデータとして扱ってしまっているため、学習データのサンプリング周波数に特化して学習されてしまい、学習データ以外のサンプリング周波数の音響信号に対して動作する保証はない。また、従来の DNN ではサンプリング周波数をパラメータとして扱っておらず、DNN を構成する層が複数のサンプリング周波数に対応できるよう設計されていない。そのため、任意のサンプリング周波数の音響信号に対して一貫して動作する DNN の実現には、

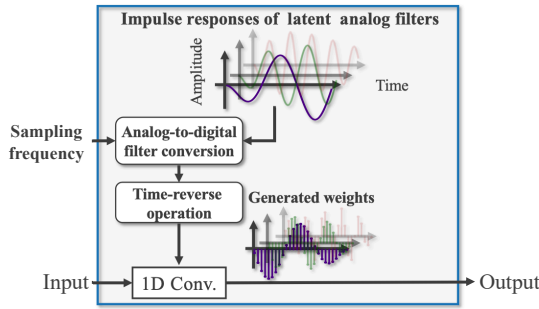
¹ 東京大学

Hongo, Bunkyo, Tokyo 113-8654, Japan

² 早稲田大学

Ohkubo, Shinjyuku, Tokyo 169-8555, Japan

(a) SFI convolution layer



(b) SFI audio source separation model

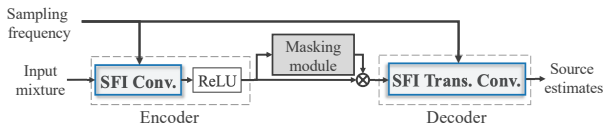


図 1: (a) SFI 畳み込み層と (b) SFI 音源分離モデルの構造

サンプリング周波数が明示的にモデル化された層を構築する必要がある。

この問題意識に基づき、我々は以前、サンプリング周波数非依存 (sampling-frequency-independent: SFI) 畳み込み層を提案した [15]。畳み込み層は入力特徴量と当該層のパラメータである重みの相互相関を出力する。そのため、信号処理の観点からみると、畳み込み層はデジタルフィルタ、重みはインパルス応答を時間反転したものと同みなせる。デジタルフィルタは離散時間領域で定義されるため、サンプリング周波数とは不可分である。そこで、当該デジタルフィルタに対して、サンプリング周波数に依存しない連続時間領域で定義されるアナログフィルタからのフィルタ生成過程を導入し、サンプリング周波数に非依存なフィルタ表現を実現した。当該アナログフィルタは重みに対する潜在的な変数として働くため、潜在アナログフィルタと呼ぶ。また、従来の音源分離用 DNN の 1 つである Conv-TasNet [3] に対して、SFI 畳み込み層を導入し、サンプリング周波数に対する分離性能の変化を調査した。

SFI 畳み込み層を用いた DNN 音源分離モデル (SFI 音源分離モデル) では、学習データよりも高いサンプリング周波数の信号を分離する際には、学習データと同じサンプリング周波数の信号を分離した際と同等の分離性能を示した [15]。一方、学習データよりも低いサンプリング周波数に関しては分離性能が大きく低下した。また、当該 Nyquist 周波数以上の周波数帯域に主要な成分をもつアナログフィルタにおいてエイリアシングが起こっていることを発見した。そこで、我々は学習済みの SFI 畳み込み層に対して事後的にエイリアシングを起こしたフィルタを取り除くことで、学習データよりも低いサンプリング周波数の分離性能が大きく向上することを確認した [15]。

これらの発見に基づき、本稿ではエイリアシングが本質的に起こり得ないフィルタ設計を用いた SFI 畳み込み層を

提案する。提案層では、周波数領域において、当該 Nyquist 周波数までのアナログフィルタの周波数応答を近似するようなデジタルフィルタを設計する。また、当該フィルタ設計手法を用いた場合でも、アナログフィルタのパラメータを誤差逆伝搬法を用いて DNN と同時に学習できることを示す。

2. SFI 畳み込み層

2.1 層構造

SFI 畳み込み層の構造を図 1 (a) に示す。SFI 畳み込み層は、連続時間領域で定義された $M^{(in)} \times M^{(out)}$ 個のアナログフィルタのインパルス応答をパラメータとして持つ。ここで、 $M^{(in)}$ と $M^{(out)}$ はそれぞれ入力と出力のチャンネルサイズである。SFI 畳み込み層は次のような手順で動作する。

- (i) デジタルフィルタ設計手法に従い、入力信号のサンプリング周波数に応じて、各アナログフィルタから長さ L の離散時間インパルス応答を生成する。
- (ii) 生成された離散時間インパルス応答を時間反転し、 $M^{(in)} \times M^{(out)} \times L$ のサイズの畳み込み層の重みを生成する。
- (iii) ステップ (ii) で生成した重みを通常の畳み込み層の重みとして使用し、入力特徴量を変換する。

このように、SFI 畳み込み層は重みをアナログフィルタをサンプリング周波数に応じて生成するため、異なるサンプリング周波数に関しても一貫した重みを生成できる。また、畳み込み層との差異は重みの生成部分だけであるため、転置畳み込み層に関しても、同様に潜在アナログフィルタを用いることで、SFI 転置畳み込み層に拡張できる。

2.2 インパルス不変法を用いた SFI 畳み込み層

我々は以前、アナログフィルタのインパルス応答をサンプリング周期に応じて直接サンプリングするフィルタ設計手法を用いた SFI 畳み込み層を提案した。当該フィルタ設計手法では、アナログフィルタが有限長のインパルス応答をもつフィルタ (FIR フィルタ) であり、当該インパルス応答長が、得られるデジタル FIR フィルタのインパルス応答長と一致する場合、サンプルされた時刻において両者のインパルス応答を一致させられる。このようにサンプルされた時刻で両者のインパルス応答が不変であることから、本稿ではインパルス不変法と呼ぶ。また、アナログフィルタが無限長のインパルス応答を持つフィルタ (IIR フィルタ) の場合には、デジタル IIR フィルタを用いなければインパルス不変とはならないが、本稿ではデジタル FIR フィルタが定義されている有限の範囲でアナログフィルタを打ち切って、デジタル FIR フィルタで近似する。

離散時間のインデックスを $l = 1, \dots, L$, 連続時間を $t \in \mathbb{R}$, サンプル周期を T とする. インパルス不変法により, 離散時間インパルス応答 $h[l]$ はアナログフィルタ $g(t)$ から以下の式 (1) に従って生成される.

$$h[l] = Tg(lT) \quad (1)$$

サンプル周期 T を変えることで, 異なるサンプル周期の離散時間インパルス応答が得られる.

2.3 Conv-TasNet への適用

音声, 楽音に対して高い分離性能を示している DNN 音源分離モデルである Conv-TasNet [3, 16] に SFI 畳み込み層を導入することで, SFI 音源分離モデルへと拡張できる [15]. Conv-TasNet はエンコーダ, デコーダ, マスキングモジュールから成る. エンコーダとデコーダは, 従来の時間周波数変換とその逆変換を模したものと解釈できる. エンコーダは入力信号を 1 次元畳み込み層 (カーネルサイズ L とストライド F を持つ) と rectified linear unit (ReLU) によって N チャンネルの擬似的な時間周波数表現に変換する. マスキングモジュールは畳み込み層で構成されるブロック R 個で構成され, 各ブロックは指数関数的に増加する dilation 係数をもつ X 個の 1 次元 dilated 畳み込み層で構成される. 畳み込み層のブロックの詳細は文献 [3] を参照されたい. マスキングモジュールでは, 分離を行うためのマスクが推定される. デコーダは, マスキングモジュールで推定されたマスクを用いたマスク処理により得られたターゲットの擬似的な時間周波数表現を, カーネルサイズ L とストライド F をもつ 1 次元転置畳み込み層によって, 時間信号に変換する.

Conv-TasNet を SFI 音源分離モデルに拡張するためには, エンコーダの畳み込み層とデコーダの転置畳み込み層を, それぞれ SFI 畳み込み層と SFI 転置畳み込み層に置き換えればよい (図 1 (b) 参照). マスキングモジュールは, 文献 [16] で用いられた Conv-TasNet と同一である.

SFI 音源分離モデルでは, カーネルサイズ L とストライド F を分離対象のサンプル周期周波数に応じて変更できる. エンコーダとデコーダは時間周波数変換に対応するため, L と F はそれぞれフレーム長とフレームシフトとみなせる. したがって, 推論時に L と F を連続時間領域で学習と同一となるように調節することで, サンプル周期周波数が変わってもマスキングモジュールに入力される擬似的な時間周波数表現のフレーム長とフレームシフトを同一にできる.

3. 提案手法

3.1 動機

インパルス不変法は, 式 (1) に示す様に, アナログフィルタを所望のサンプル周期でサンプル周期をすること

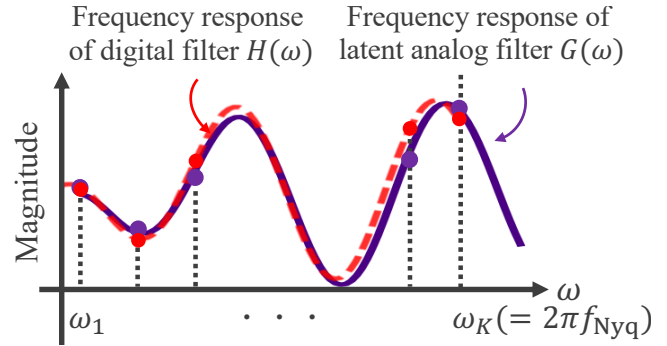


図 2: 提案法で用いるデジタルフィルタ設計手法の概要

で対応するデジタルフィルタを得るフィルタ設計手法である. この手法では, 入力信号の Nyquist 周波数よりも高い周波数成分をもつアナログフィルタからデジタルフィルタを作成する際には, エイリアシングが起こる. このとき, 学習データよりも低いサンプル周期周波数のデータを分離する際に分離性能が低下することを実験により確認した [15]. そこで, 分離性能の低下要因がエイリアシングであるかを調査するため, 学習済みの SFI 畳み込み層の重み生成時に, 観測信号の Nyquist 周波数よりも高い中心周波数をもつアナログフィルタに対応する重みを全て 0 とした. このアンチエイリアシング機構を導入することで, 学習データよりも低いサンプル周期周波数での分離性能が大幅に向上し, フィルタ設計時のエイリアシングが性能低下の主要因であることを発見した.

そこで, 本稿では重みの生成段階からエイリアシングを回避できるフィルタ設計手法を用いた SFI 畳み込み層を構築する.

3.2 周波数領域でのフィルタ設計に基づく SFI 畳み込み層

本節では, 提案層におけるアナログフィルタからのデジタルフィルタ設計について述べる. また, 簡単のため, 1 つのアナログフィルタに関するフィルタ設計処理について述べる.

所望の Nyquist 周波数を f_{Nyq} , 角周波数を ω とし, 潜在的なアナログフィルタの周波数応答を $G(\omega)$ で表す. 提案層では, エイリアシングを回避できるように, f_{Nyq} までの周波数帯域の $G(\omega)$ を近似するデジタルフィルタを設計する (図 2 参照). アナログフィルタの周波数応答を, 入力信号の Nyquist 周波数まで K 点でサンプル周期し, それらを $\mathbf{g} = [G(\omega_1), \dots, G(\omega_K)]^T$ とする. ここで, T は転置記号である. インパルス応答長が $2Q + 1$ であるデジタルフィルタのインパルス応答を $\tilde{\mathbf{b}} = [\tilde{b}_{-Q}, \dots, \tilde{b}_Q]^T$ とする. インパルス応答 $\tilde{\mathbf{b}}$ を持つデジタルフィルタの周波数応答 $H(\omega)$ は,

表 1: 各比較手法の特徴

Method	$g_n(t)$	Samp. freq. adaptation	Anti-aliasing mechanism
Conv-Tasnet [3]	-	-	-
SFI ImpInv [15]	$g_n^{(\text{Gauss})}(t)$	Yes	No
SFI ImpInv+ [15]	$g_n^{(\text{Gauss})}(t)$	Yes	Yes
Proposed	$g_n^{(\text{Gauss})}(t)$	Yes	Yes

$$H(\omega) = \sum_{q=-Q}^Q \tilde{b}_q e^{-jq\omega T} = [e^{jQ\omega T}, \dots, e^{-jQ\omega T}] \tilde{\mathbf{b}} \quad (2)$$

となる。また、サンプリング点 ω_1 から ω_K までの周波数応答を $\mathbf{h}$ とすると

$$\mathbf{h} = \begin{bmatrix} H(e^{j\omega_1 T}) \\ \vdots \\ H(e^{j\omega_K T}) \end{bmatrix} = \underbrace{\begin{bmatrix} e^{jQ\omega_1 T} & \dots & e^{-jQ\omega_1 T} \\ \vdots & \ddots & \vdots \\ e^{jQ\omega_K T} & \dots & e^{-jQ\omega_K T} \end{bmatrix}}_W \tilde{\mathbf{b}} \quad (3)$$

と表せる。ここで、式 (3) の右辺の行列を W で表す。所与のサンプリング周波数に対してアナログフィルタ $G(\omega)$ からデジタルフィルタ $H(\omega)$ を設計する問題は、次のような誤差 $E(\tilde{\mathbf{b}})$ を最小化するインパルス応答 $\tilde{\mathbf{b}}$ を求める最小二乗問題として定式化できる。

$$E(\tilde{\mathbf{b}}) = \|\mathbf{g} - \mathbf{h}\|_2^2 = \|\mathbf{g} - W\tilde{\mathbf{b}}\|_2^2 \quad (4)$$

誤差 $E(\tilde{\mathbf{b}})$ の最小化解は、 W の Moore–Penrose の擬似逆行列 $W^\dagger$ を用いて以下のように表せる。

$$\mathbf{b}^* = W^\dagger \mathbf{g} \quad (5)$$

この過程によって得られた $\mathbf{b}^*$ を時間反転したものが SFI 畳み込み層のあるチャンネルの重みとなる。この設計手法では、入力信号の Nyquist 周波数までアナログフィルタを近似することになるため、インパルス不変法とは異なり、重み生成段階からエイリアシングを回避できる。

適切にアナログフィルタを設計することで、アナログフィルタのパラメータも DNN と同時に誤差逆伝播法を用いて学習できる。誤差逆伝播法では、ロス関数から各学習パラメータに対する勾配を計算する必要がある。提案層のフィルタ設計手法は、式 (5) のように $\mathbf{g}$ と $\mathbf{b}^*$ が線形に結ばれているため、畳み込み層の重みである $\mathbf{b}^*$ に対する勾配を $W^\dagger$ を用いて $\mathbf{g}$ に対する勾配へと変換できる。そのため、 $\mathbf{g}$ がアナログフィルタの学習したいパラメータに関して微分可能であれば、当該パラメータに関する勾配を計算できる。

4. 実験的評価

4.1 実験条件

楽音分離用のデータセットである MUSDB18-HQ [17] を

表 2: 実験で用いたマスキングモジュールのハイパーパラメータ

Symbol	Description	Value
N	# of channels of latent representation	440
X	# of convolutional blocks in each repeat	6
R	# of repeats	2
B	# of channels in bottleneck and residual paths' 1×1 convolution blocks	160
Sc	# of channels in skip-connection paths' 1×1 convolution blocks	160
H	# of channels in convolution blocks	160
P	Kernel size in convolution blocks	3

用いて実験的評価を行った。MUSDB18-HQ は 86 曲の学習データ、14 曲の検証データ、50 曲のテストデータから成り、各曲は *vocals*, *bass*, *drums*, *other* の 4 つの楽器から構成されている。学習データと検証データは 32 kHz に、テストデータは 8, 12, 16, ..., 48 kHz にリサンプリングして用いた。学習データとテストデータ共に各曲の観測音響信号に対して、平均 0、分散 1 となるようにデータ標準化を行った。評価指標には signal-to-distortion ratio (SDR) [18] を用いた。SDR の計算には BSSEval v4 toolkit [19] を用いた。実験に用いた全手法に対して、4 種類の乱数シードで実験を行い、4 シードでの SDR の平均と標準誤差をその手法の SDR として評価した。

比較手法として、Conv-TasNet [3]、インパルス不変法で設計した SFI 畳み込み層を用いた SFI 音源分離モデル (SFI ImpInv) [15]、SFI ImpInv にアンチエイリアシング機構を導入した手法 (SFI ImpInv+) [15]、提案手法 (Proposed) の 4 手法を用いた。全手法推論時、入力信号を学習したサンプリング周波数へのリサンプルをせず、未知のサンプリング周波数の信号としたまま入力した。比較した 4 手法の特徴を表 1 に示す。なお、SFI ImpInv+ と Proposed は同じ特徴を有しているが、Proposed では周波数領域における最小二乗近似でエイリアシングを生じないフィルタを直接設計している点が異なっている。

学習では、文献 [15] と同一のデータ拡張手法を用いた。ミニバッチには、各曲からランダムに 8 秒切り取り、ゲインを [0.75, 1.25] の範囲からランダムに選び楽器毎に適用した。また、ステレオ音源の左右のチャンネルのうちランダムに選定されたモノラル音源を用い、ミニバッチの半分において曲間で楽器をランダムにシャッフルした。

各手法の学習では、32 kHz の信号に対してフレーム長が 5 ms、フレームシフトが 2.5 ms となるように、つまり $L = 160$, $F = 80$ となるようにエンコーダとデコーダのパラメータを設定した。マスキングモジュールのパラメータは、文献 [3] の Table 1 と同様のパラメータ表記を用いて、文献 [15] と同様、表 2 の様に定めた。

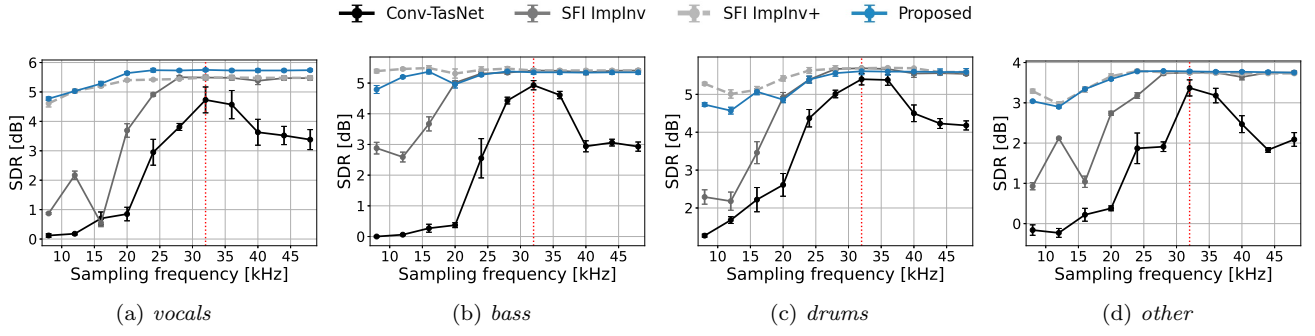


図 3: 様々なサンプリング周波数のテストデータに対する従来手法と提案手法の SDR

潜在的なアナログフィルタとして、本稿では次式のように、時間領域で $g^{(\text{Gauss})}(t)$ 、周波数領域で $G^{(\text{Gauss})}(\omega)$ となるフィルタを用いた。

$$g^{(\text{Gauss})}(t) = 2\sqrt{2\sigma^2\pi} \exp\left(-\frac{1}{2}\sigma^2 t^2\right) \cos(\omega_c t + \phi) \quad (6)$$

$$G^{(\text{Gauss})}(\omega) = \exp\left[-\frac{(\omega - \omega_c)^2}{2\sigma^2} + j\phi\right] + \exp\left[-\frac{(\omega + \omega_c)^2}{2\sigma^2} - j\phi\right] \quad (7)$$

ここで、 ω_c は中心角周波数、 σ はバンド幅、 ϕ は位相を表す。畳み込み層の各チャンネル $n = 1, \dots, N$ によって異なるフィルタを用いるため、 $g^{(\text{Gauss})}(t)$ 、 $G^{(\text{Gauss})}(\omega)$ とパラメータ ω_c, σ, ϕ を、インデックス n を用いて、 $g_n^{(\text{Gauss})}(t)$ 、 $G_n^{(\text{Gauss})}(\omega)$ 、 $\omega_{c,n}, \sigma_n, \phi_n$ と表現する。また、 $\omega_{c,n}, \sigma_n, \phi_n$ を学習可能なパラメータとし、ネットワークと同時に学習した。各パラメータの学習における初期化についてだが、 $\omega_{c,n}$ は、50 Hz から 16 kHz にかけて ERB スケール [20] において $N = 440$ 等分になるように初期化した。 σ_n は 20π を初期値として用い、学習によってバンド幅が狭くなりすぎないように σ_n が 2π 以上になるように制限を加えた。 ϕ_n は 0 から π の範囲で一様分布からランダムで初期値を与えた。

提案手法におけるサンプル数 K は学習時、推論時で変更可能であるが、今回は学習時、推論時共に 320 点とした。

学習時の最適化は、文献 [15] 同様、RAdam [21] と look-ahead optimizer [22] を用いた。また、勾配クリッピングを用い、勾配の L_2 ノルムの値が 5.0 以下となるようにした。学習率のスケジューラには stochastic gradient descent with warm restarts [23] を用いた。最適化手法と学習率のスケジューラのパラメータも文献 [15] で用いられた値を用いた。各手法の学習において、バッチサイズを 12、イタレーションを 250 とし、損失関数は scale-invariant source-to-noise ratio (SI-SNR) を用いた。

テストデータにはステレオ音源を用い、モノラル音源で学習した DNN を左チャンネルと右チャンネルに別々に適用した。SI-SNR はスケール不変であり、BSSEval v4 toolkit で算出される SDR はスケール依存であるため、各楽器 $i = 1, \dots, 4$ の分離音に対してスケール α_i を用いて音量補

正を行った [16]。時間長 Y の混合音を $\mathbf{s} \in \mathbb{R}^Y$ 、 i 番目の分離音を $\hat{\mathbf{s}}_i \in \mathbb{R}^Y$ とすると、 $\alpha = [\alpha_1, \dots, \alpha_4]^T$ は以下の式 (8) で計算される。

$$\alpha = \underset{\alpha}{\operatorname{argmin}} \left(\mathbf{s} - \sum_i \alpha_i \hat{\mathbf{s}}_i \right)^2 \quad (8)$$

4.2 実験結果

図 3 はテストデータに対する分離性能を示した図である。各 SDR は 4 種類の乱数シード値で学習したモデルの SDR の平均であり、エラーバーは標準誤差を表す。赤破線は学習データのサンプリング周波数 (32 kHz) を示している。Conv-TasNet では、入力信号のサンプリング周波数が 32 kHz から離れるにつれて SDR が大きく減少した。SFI ImpInv では、32 kHz の信号のみで学習したにもかかわらず、32 kHz から 48 kHz の信号に対して、32 kHz の信号を分離した際と同程度の SDR を示した。一方で、入力信号のサンプリング周波数が 32 kHz より低くなるにつれて、SDR が大きく減少した。SFI ImpInv+, Proposed では、32 kHz の信号のみで学習したにもかかわらず、全てのテストデータである 8 kHz から 48 kHz の信号に対して 32 kHz の信号を分離した際と同等の SDR を示した。この結果から、エイリアシングを回避した SFI ImpInv+ と Proposed のどちらも、入力信号のサンプリング周波数に対して適応することで、学習データ以外のサンプリング周波数の信号に対しても一貫した分離性能が得られることを確認した。特に、周波数領域において最小二乗近似でフィルタを設計する手法を用いることにより、エイリアシングの影響が本質的に回避され、文献 [15] で導入したアンチエイリアシング機構を導入しなくても、32 kHz より低いサンプリング周波数の信号に関しても一貫した分離性能が得られることが確認できた。

SFI ImpInv+ と Proposed の SDR を比較すると、*vocals* では 20 kHz から 48 kHz にかけて Proposed の方が若干高く、*bass*、*drums* では 8 kHz から 24 kHz にかけて SFI ImpInv+ の方が若干高かった。しかし、SDR に差はあるものの、著者の 1 人が聴取した限り同程度の分離品質であった。この結果は、提案法がフィルタ設計の段階から系

統的にエイリアシングを回避しつつ、サンプリング周波数によらず一貫した分離性能を達成できることを示している。また、用いた2つのフィルタ設計手法の差異よりも、エイリアシングの有無が分離性能に影響することが確認できた。今後は、より詳細な分離品質の評価のため主観評価実験を行う予定である。

5. 結論

本稿では、インパルス不変法を用いたSFI畳み込み層におけるエイリアシングの問題を、重み生成段階から本質的に解決するために、周波数領域において最小二乗近似でフィルタを設計する手法を用いたSFI畳み込み層を提案した。楽音分離実験により、提案手法が、インパルス不変法を用いたSFI畳み込み層にアンチエイリアシング機構を導入せずとも、フィルタ設計の段階から系統的にエイリアシングを回避しつつ、学習データ以外のサンプリング周波数を持つ信号に対しても一貫して動作することを確認した。

謝辞 本研究はJSPS 科研費JP20K19818の助成を受けた。

参考文献

- [1] Hershey, J. R., Chen, Z., Le Roux, J. and Watanabe, S.: Deep Clustering: Discriminative Embeddings for Segmentation and Separation, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 31–35 (2016).
- [2] Yu, D., Kolbæk, M., Tan, Z. and Jensen, J.: Permutation Invariant Training of Deep Models for Speaker-Independent Multi-talker Speech Separation, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 241–245 (2017).
- [3] Luo, Y. and Mesgarani, N.: Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 8, pp. 1256–1266 (2019).
- [4] Stoller, D., Ewert, S. and Dixon, S.: Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation, *Proceedings of International Society for Music Information Retrieval Conference*, pp. 334–340 (2018).
- [5] Nakamura, T., Kozuka, S. and Saruwatari, H.: Time-Domain Audio Source Separation with Neural Networks Based on Multiresolution Analysis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2021). to appear.
- [6] Ditter, D. and Gerkmann, T.: A Multi-Phase Gammatone Filterbank for Speech Separation Via Tasnet, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 36–40 (2020).
- [7] Liu, H., Xie, L., Wu, J. and Yang, G.: Channel-wise Sub-band Input for Better Voice and Accompaniment Separation on High Resolution Music, *Proceedings of INTERSPEECH* (2020).
- [8] Défossez, A., Usunier, N., Bottou, L. and Bach, F.: Music Source Separation in the Waveform Domain, *arXiv preprint arXiv:1911.13254* (2019).
- [9] Krebs, F., Böck, S., Dorfer, M. and Widmer, G.: Downbeat Tracking Using Beat Synchronous Features with Recurrent Neural Networks, *Proceedings of International Society for Music Information Retrieval Conference*, pp. 129–135 (2016).
- [10] Sigtia, S., Benetos, E. and Dixon, S.: An End-to-End Neural Network for Polyphonic Piano Music Transcription, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 5, pp. 927–939 (2016).
- [11] Pedersoli, F., Tzanetakis, G. and Yi, K. M.: Improving Music Transcription by Pre-Stacking A U-Net, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 506–510 (2020).
- [12] Yu, D., Seltzer, M., Li, J., Huang, J. and Seide, F.: Feature Learning in Deep Neural Networks - Studies on Speech Recognition, *Proceedings of International Conference on Learning Representations* (2013).
- [13] Narayanan, A., Misra, A., Sim, K. C., Pundak, G., Tripathi, A., Elfeky, M., Haghani, P., Strohmaier, T. and Bacchiani, M.: Toward Domain-Invariant Speech Recognition via Large Scale Training, *Proceedings of IEEE Spoken Language Technology Workshop*, pp. 441–447 (2018).
- [14] Gao, J., Du, J. and Chen, E.: Mixed-Bandwidth Cross-Channel Speech Recognition via Joint Optimization of DNN-Based Bandwidth Expansion and Acoustic Modeling, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 3, pp. 559–571 (2019).
- [15] Saito, K., Nakamura, T., Yatabe, K., Koizumi, Y. and Saruwatari, H.: Sampling-Frequency-Independent Audio Source Separation Using Convolution Layer Based on Impulse Invariant Method, *Proceedings of European Signal Processing Conference* (arXiv preprint arXiv:2105.04079) (2021).
- [16] Samuel, D., Ganeshan, A. and Naradowsky, J.: Meta-Learning Extractors for Music Source Separation, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 816–820 (2020).
- [17] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimitakis, S. I. and Bittner, R.: MUSDB18-HQ - an uncompressed version of MUSDB18, <https://doi.org/10.5281/zenodo.3338373> (2019).
- [18] Vincent, E., Gribonval, R. and Fevotte, C.: Performance Measurement in Blind Audio Source Separation, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 14, No. 4, pp. 1462–1469 (2006).
- [19] Stöter, F.-R., Liutkus, A. and Ito, N.: The 2018 Signal Separation Evaluation Campaign, *Proceedings of International Conference on Latent Variable Analysis and Signal Separation*, pp. 293–305 (2018).
- [20] Hohmann, V.: Frequency Analysis and Synthesis Using a Gammatone Filterbank, *Acta Acustica united with Acustica*, Vol. 88, No. 03, pp. 433–442 (2002).
- [21] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. and Han, J.: On the Variance of the Adaptive Learning Rate and Beyond, *Proceedings of International Conference on Learning Representations* (2020).
- [22] Zhang, M., Lucas, J., Ba, J. and Hinton, G.: Lookahead Optimizer: k Steps Forward, 1 Step Back, *Proceedings of Advances in Neural Information Processing Systems*, pp. 9597–9608 (2019).
- [23] Loshchilov, I. and Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts, *Proceedings of International Conference on Learning Representations* (2017).