Vol.2021-MUS-130 No.32
Vol.2021-EC-59 No.32
2021/3/17

アンチエイリアシング機構を導入したサンプリング周波数非依存畳み込み層を用いた音源分離

齋藤 弘一[1]　中村 友彦[1]　矢田部 浩平[2]　小泉 悠馬[3]　猿渡 洋[1]

概要：我々はこれまでに，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シングルチャネルの deep neural network（DNN）音源分離モデルを実現するため，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てきた．SFI 畳み込み層では，畳み込み層の重みがデジタルフィルタとみなせることに着眼し，アナログフィルタからのデジタルフィルタ設計手法の 1 つであるインパルス不変法を用いて重みを生成する．これにより，学習に用いなかったサンプリング周波数に対する畳み込み層の重みが生成できる．SFI 畳み込み層を用いた DNN 音源分離モデルでは，学習したサンプリング周波数と同一，またはより高いサンプリング周波数に対しては一貫した性能を示すものの，低いサンプリング周波数に対しては分離性能が低下することが実験的に確認されている．本稿では，SFI 畳み込み層の重み生成過程においてエイリアシングシングを引き起こしうるアナログフィルタに対応する重みを用いないことで，低いサンプリング周波数での分離性能低下が軽減できることを示す．楽音分離実験により，学習後に提案手法を SFI 畳み込み層に導入するだけでも，低いサンプリング周波数の音響信号に対して分離性能が向上することを確認した．

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，エイリアシング，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複数の音源信号が混合された観測信号から，各音源の個別の信号を抽出する技術である．近年，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]．音源分離は楽曲のリミックスや自動採譜，音声認識など様々なアプリケーションの前処理として利用されることが多い．そのため，様々なアプリケーションで対象とされる音響信号に対して適用可能であることが望ましい．

サンプリング周波数は音響信号の重要な構成要素の 1 つであり，アプリケーションに応じて定まる．例えば，楽曲のリミックスや編集では人間の可聴域をカバーする 44.1 kHz や 48 kHz が通常使用される [7, 8]．これらのアプリケーションは人間が聴くことを想定しているからである．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全ての可聴域の情報が必要ではない．ビートトラッキングでは 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]．汎用的な前処理として音源分離を用いるためには，これら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 音源分離モデルを構築する必要がある．

従来の DNN 音源分離モデルは，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いるため，学習データに含まれないサンプリング周波数の信号に対して適用できる保証はない．これは，従来の DNN 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 を構成する層は複数のサンプリング周波数の音響信号を処理できるように設計されていないためである．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 を実現するためには，サンプリング周波数を陽に表現した層を設計する必要がある．

これに対して，我々は単一の DNN で任意のサンプリング周波数の音響信号を扱うことができる，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]．畳み込み層は複数のデジタルフィ

[1] 東京大学
Hongo, Bunkyo, Tokyo 113-8654, Japan
[2] 早稲田大学
Ohkubo, Shinjyuku, Tokyo 169-8555, Japan
[3] 日本電信電話株式会社
Midori-Cho, Musashino-Shi, Tokyo 180-8585, Japan

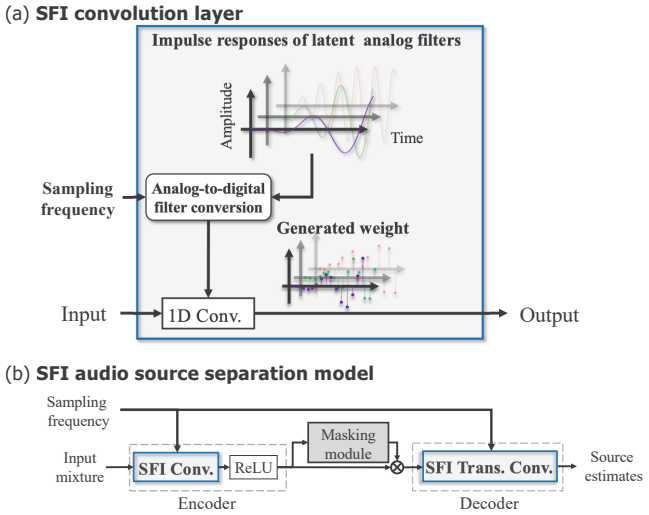


図 1: （a） SFI 畳み込み層の構造 （b） SFI 音源分離モデルの構造 [15].

Fig. 1 (a) Schematic illustration of SFI convolution layer, and (b) architectures of SFI audio source separation model [15].

ルタからなるフィルタバンクとして解釈できる．デジタルフィルタが定義される離散時間領域は，サンプリング周波数とは不可分である．そのため，サンプリング周波数を用いずとも定義できるアナログフィルタから当該デジタルフィルタが設計する．すなわち，当該デジタルフィルタに対して，アナログフィルタからのフィルタ生成過程を導入する．これにより，サンプリング周波数に依存しない潜在的なフィルタ表現を実現する．この表現を畳み込み層に組み込むことで，サンプリング周波数に非依存な畳み込み層が実現できる．

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．一方，学習データよりも低いサンプリング周波数の信号に対しては，サンプリング周波数が低くなるほど分離性能が低下した [15].

本稿では，この問題を解決するための初期検討として，学習した SFI 畳み込み層に対し，対象音源の Nyquist 周波数未満の中心周波数を持つアナログフィルタのみを用いる手法を提案する．また，楽音分離実験により提案手法の有効性を検証する．

2. SFI 音源分離モデル

2.1 SFI 畳み込み層の概要

SFI 畳み込み層は，単一の DNN で任意のサンプリング周波数の音響信号を扱うために提案された畳み込み層である．層の中では入力特徴量と重みの相互相関が出力される．そのため，層の重みを，時間反転したインパルス応答としたデジタルフィルタと解釈できる．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に本質的に依存しており，この領域のままサンプリング周波数に非依存な構造を構築することは難しい．そこで，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，デジタルフィルタとして重みが生成される過程を導入した畳み込み層として，我々は SFI 畳み込み層を提案した．アナログフィルタは，畳み込み層の重みの潜在的なフィルタ表現として機能するため，本稿では潜在アナログフィルタ表現と呼ぶ．

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 $M^{(\mathrm{in})} \times M^{(\mathrm{out})}$ 個のアナログフィルタのインパルス応答から生成される．ここで $M^{(\mathrm{in})}$ と $M^{(\mathrm{out})}$ はそれぞれ入力と出力のチャネルサイズである．SFI 畳み込み層は次の 3 ステップで生成される．

（i） 入力信号のサンプリング周波数に応じて，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，$M^{(\mathrm{in})} \times M^{(\mathrm{out})} \times L$ 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する．

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表現を用いることで，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畳み込み層の重み生成

我々は以前，フィルタ設計手法としてインパルス不変法を用いた SFI 畳み込み層を提案した [15]．サンプリング周期を T，離散時間のインデックスを $l = 1, \cdots, L$，連続時間を $t \in \mathbb{R}$ とする．インパルス不変法によると，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．

$$h[l] = g(lT) \tag{1}$$

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

ステップ（i）におけるアナログフィルタ $g(t)$ として，multi-phase gammatone filter（MP-GTF）[6] が用いられた．MP-GTF のインパルス応答は以下の式 (1) で表される．

$$g^{(\text{MP-GTF})}(t) = at^{p-1}e^{-2\pi bt}\cos(2\pi ft + \phi) \tag{2}$$

ここで a は振幅，p はフィルタ次数，b はバンド幅，f は中心周波数，ϕ は初期位相を表す．また，パラメータ b は equivalent rectangular bandwidth（ERB）スケール [16] を用いて $b = \mathrm{ERB}(f)/1.57$ で表される．ここで

$\mathrm{ERB}(f) = 24.7 + f/9.265$ である．中心周波数 f と位相 ϕ の組み合わせに対して，$g^{(\text{MP-GTF})}(t)$ を L 点でサンプリングすることで，長さ L の離散時間インパルス応答が得られる．

SFI 畳み込み層を Conv-TasNet [3] に組み込む場合（2.3 節参照），畳み込み層の後には活性化関数として rectified linear unit（ReLU）が用いられているため，MP-GTF の出力の負値が 0 となり，入力信号の情報が欠落しうる．この情報の欠落を避けるため，各フィルタに対して逆相，すなわち同一中心周波数かつ $\phi + \pi$ の位相を持つフィルタも用いる．

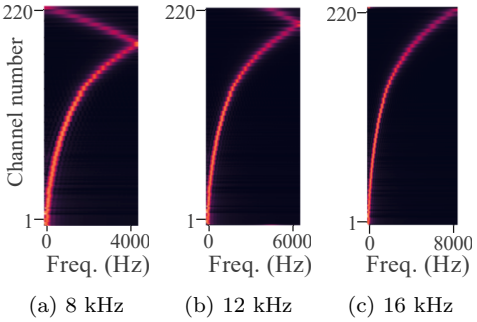
畳み込み層の各チャネル m によって異なる $g(t)$ を用いるため，$g^{(\text{MP-GTF})}(t)$ に対して，インデックス m を用いて，$g_m^{(\text{MP-GTF})}(t), a_m, p_m, b_m, f_m, \phi_m$ と表現する．文献 [15] と同様に，本稿でも f_m, ϕ_m を学習する．

2.3 SFI 畳み込み層の 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, 17] 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 W を持つ）と ReLU によって N チャネルの擬似的な時間周波数表現に変換する．その後，音源ごとに用意されたマスキングモジュールによって，擬似的な時間周波数表現に対してマスク処理が行われる．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．デコーダは，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 W 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 [17] 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 W を分離対象のサンプリング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 W 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 W 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，サンプリング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疑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．



(a) 8 kHz　(b) 12 kHz　(c) 16 kHz

図 2: 8 kHz，12 kHz，16 kHz のサンプリング周波数に対する SFI 畳み込み層のチャネル数 $m = 220$ までの周波数応答（(a) と (c) は [15] より引用）

Fig. 2 Frequency responses of first 220 filters of trained SFI convolution layer at sampling frequencies of 8, 12, and 16 kHz.

3. 提案手法

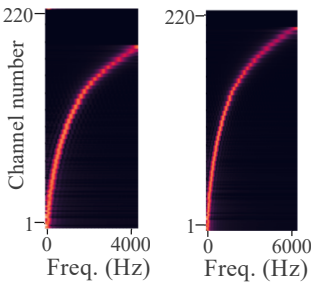
3.1 動機

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リング周期でサンプリングをすること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．そのため，この設計手法はエイリアシングシングを考慮しておらず，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，エイリアシングシングが起こりうる．図 2 は，学習によって得られた 8 kHz，12 kHz，16 kHz の信号を分離する際のエンコーダ内の SFI 畳み込み層の周波数応答である（実験条件については 4 節参照）．図 2（a），（b）に示す通り，Nyquist 周波数近辺やそれを越えた中心周波数を持つアナログフィルタに対応するデジタルフィルタでエイリアシングシングが起きている．このため，学習データよりも低いサンプリング周波数の信号を分離する場合には，マスキングモジュールは，折り返し雑音を含んだ特徴量を用いて分離を行うこととなる．

特徴量領域でのエイリアシングは，DNN を用いた音源分離 [5]，音素認識，音声感情認識 [18]，画像認識 [19] などのタスクにおいて性能低下を引き起こすことが知られている．そのため，我々はエイリアシングを低減する機構を SFI 音源分離モデルに導入し，学習データよりも低いサンプリング周波数での分離性能低下の原因を調査する．

3.2 アンチエイリアシング機構の導入

エイリアシングの低減のため，学習後の SFI 畳み込み層に対して，分離対象信号の Nyquist 周波数未満の中心周波数を持つアナログフィルタのみで分離を行う方法を提案す



(a) 8 kHz (b) 12 kHz

図 3: 8 kHz，12 kHz のサンプリング周波数における提案手法を用いた SFI 畳み込み層のチャネル数 $m = 220$ までの周波数応答

Fig. 3 Frequency responses of first 220 filters of SFI convolution layer with anti-aliasing mechanism at sampling frequencies of 8 kHz and 12 kHz.

表 1: 各比較手法の特徴
Table 1 Features of Conv-TasNet, SFI Conv., and Proposed

Method	$g_m(t)$	Samp. freq. adapt.	Anti-aliasing mech.
Conv-Tasnet [3]	-	No	-
SFI conv.	$g_m^{(\text{MP-GTF})}(t)$	**Yes**	No
Proposed	$g_m^{(\text{MP-GTF})}(t)$	**Yes**	**Yes**

る．提案手法では，モデルの推論時にエンコーダとデコーダの重みに関して，対応するアナログフィルタの中心周波数 f_m が分離対象である信号の Nyquist 周波数以上となるフィルタの重みを全てゼロにすることにより，エイリアシングによる影響を軽減する．

図 2 の（a）と（b）に対して，提案手法を導入した際のエンコーダ内の SFI 畳み込み層の周波数応答を図 3（a）と（b）に示す．

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 [20] を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 16 kHz に，テストデータは $8, 12, 16, \ldots, 32$ kHz にリサンプリングして用いた．またデータを DNN に入力する際，各曲平均 0，分散 1 の標準化を行った．評価指標には signal-to-distortion raito（SDR）[21] を用いた．SDR の計算には BSSEval v4 toolkit [22] 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 の平均と分散をその手法の SDR として評価した．

比較手法として，提案手法（Proposed）と，文献 [15] にて提案されたアンチエイリアシング機構を導入していない

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
Table 2 Hyperparameters of masking modules used in experiments

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

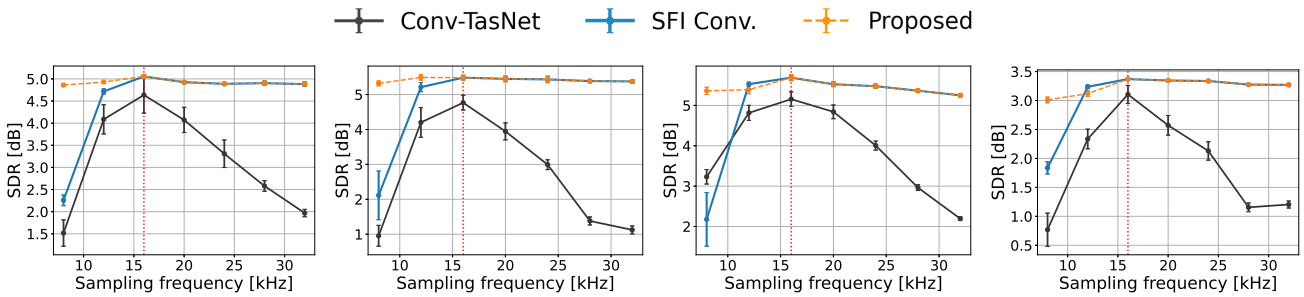
SFI 音源分離モデル（SFI Conv.）と Conv-TasNet [3] を用いた．Proposed と SFI Conv. の相違点は，Proposed では推論時にアンチエイリアシング機構を適用する点のみであることに注意されたい．推論時には全ての手法に対して，分離対象の信号を学習したサンプリング周波数にリサンプルを行わず，未知のサンプリング周波数の信号としたまま入力した．比較 3 手法の特徴を表 1 に示す．

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，パワーをランダムに $[0.75, 1.25]$ 倍した信号を用いた．また，ステレオ音源の左右のチャ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，16 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 $L = 80$，$W = 40$ 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，表 2 の様に定めた．

$g_m^{(\text{MP-GTF})}(t)$ の f_m と ϕ_m $(m = 1, \ldots, 220)$ もネットワークと同時に学習した．また，2.2 節で述べたように，全ての m に関して $f_{m+220} = f_m$, $\phi_{m+220} = \phi_m + \pi$ とし，常に逆相のフィルタも存在するように設計した．f_m と ϕ_m の学習について，50 Hz から 8000 Hz にかけて ERB スケール [16] において 48 等分になるように $f_i^{(\text{c})}$ $(i = 1, \cdots, 48)$ を用いて，$f_m = f^{(\text{c})}_{\lfloor m/K \rfloor + 1}$ となるように初期化した．ここで，i において $f_i^{(\text{c})} < f_{i+1}^{(\text{c})}$ であり，$m = 1, \ldots, 140$（$m = 141, \ldots, 220$）に関して，$K = 5$（$K = 4$）である．また，ϕ_m は同じ $f_i^{(\text{c})}$ に対して $[0, \pi)$ の範囲で均等になるように初期化し，他のパラメータは $a_m = 1$, $p_m = 2$ と設定した．

学習時の最適化は，文献 [15] 同様，RAdam [23] と lookahead optimizer [24] 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



(a) *vocals* (b) *bass* (c) *drums* (d) *other*

図 4: 様々なサンプリング周波数のテストデータに対する Conv-TasNet，SFI Conv.，Proposed の SDR

Fig. 4 SDRs of Conv-TasNet, SFI Conv., and proposed model for test data at various sampling frequencies.

warm restarts [25] 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 raito（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 を左チャネルと右チャ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 $i = 1, \ldots, 4$ 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 [17]．ここで α は $\mathbf{s} \in \mathbb{R}^{T'}$ を分離対象の混合音，$\bar{\mathbf{s}}_i \in \mathbb{R}^{T'}$ を分離音とすると，以下の式 (3) で計算される．T' は分離信号の信号長を表す．

$$\alpha = \arg\min_{\alpha} (\mathbf{s} - \Sigma_i \alpha_i \bar{\mathbf{s}}_i)^2 \qquad (3)$$

4.2 実験結果

図 4 にテストデータに対する分離性能を示す．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（16 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 16 kHz から大きく，もしくは小さくなるにつれて SDR が大きく減少した．一方 SFI Conv. と Proposed では 16 kHz の信号のみで学習したにもかかわらず，16 kHz から 32 kHz の信号に対して，16 kHz の信号を分離した際と同程度の SDR を示した．この結果から，文献 [15] での報告同様，入力信号のサンプリング周波数に応じてエンコーダとデコーダの SFI 畳み込み層の重みを変更することによって，入力信号のサンプリング周波数を変えても一貫した分離性能を得られることを確認した．

しかし SFI Conv. では，入力信号のサンプリング周波数が 16 kHz から小さくなるにつれて SDR が減少，特に 8 kHz の信号を入力した際は大幅に減少した．一方，Proposed では，SFI Conv. と比較して，（c）*drums* や（d）*other* の 12 kHz の場合を除いて，8 kHz や 12 kHz の信号を分離する際，より一貫した SDR を示した．学習後の f_m の最大値が 8 kHz 未満であったため，提案手法は 8 kHz と 12 kHz のサンプリング周波数の信号に対する分離にのみ用いた．この結果は，アンチエイリアシング機構を導入することにより，学習データよりも低いサンプリング周波数のデータを分離する際の分離性能の低下を低減できることを示している．したがって，提案手法を導入したインパルス不変法を用いた SFI 畳み込み層により，学習したサンプリング周波数以外でも一貫して動作する DNN が実現できる．

図 4 の（c）*drums* と（d）*other* から，12 kHz の信号を入力した際，Proposed が SFI Conv. よりも低い SDR を示している．この原因として，*drums* と *other* の分離においては，中心周波数 f_m が Nyquist 周波数近辺のフィルタが分離性能に大きく寄与しており，アンチエイリアシング機構の導入により，当該フィルタが用いられなくなったため，分離性能が低下したと考えられる．

5. 結論

本稿では，SFI 畳み込み層を用いた SFI 音源分離モデルにおいて，学習データよりも低いサンプリング周波数の信号を分離する際の分離性能の低下を解決するため，SFI 畳み込み層へのアンチエイリアシング機構の導入を提案した．提案手法は，推論時に学習後の SFI 畳み込み層に対して，分離対象信号の Nyquist 周波数未満の中心周波数を持つアナログフィルタのみで分離を行う．楽音分離実験により，学習後の SFI 畳み込み層に対して提案手法を導入することで，SFI 音源分離モデルが学習したサンプリング周波数以外の信号で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 Inter-*

national 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. and Saruwatari, H.: Time-Domain Audio Source Separation Based on Wave-U-Net Combined with Discrete Wavelet Transform, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 386–390 (2020).

[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 Subband 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., Strohman, T. and Bacchiani, M.: Toward Domain-Invariant Speech Recognition via Large Scale Training, *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] 齋藤弘一，中村友彦，矢田部浩平，小泉悠馬，　猿渡洋：潜在アナログフィルタ表現に基づく畳み込み層を用いたサンプリング周波数非依存な DNN 音源分離，日本音響学会春季研究発表会講演論文集 (2021).

[16] Hohmann, V.: Frequency analysis and synthesis using a Gammatone filterbank, *Acta Acustica united with Acustica*, Vol. 88, No. 03, pp. 433–442 (2002).

[17] 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).

[18] Gong, Y. and Poellabauer, C.: Impact of Aliasing on Deep CNN-Based End-to-End Acoustic Models, *Proceedings of INTERSPEECH*, pp. 2698–2702 (2018).

[19] Zeiler, M. and Fergus, R.: Visualizing and Understanding Convolutional Networks, *Proceedings of European Conference on Computer Vision*, pp. 818–833 (2014).

[20] Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I. and Bittner, R.: MUSDB18-HQ - an uncompressed version of MUSDB18 (2019).

[21] et al., E. V.: Performance measurement in blind audio source separation, *IEEE TASLP*, Vol. 14, No. 4, pp. 1462–1469 (online), DOI: 10.1109/TSA.2005.858005 (2006).

[22] 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).

[23] 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).

[24] 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).

[25] Loshchilov, I. and Hutter, F.: SGDR: Stochastic Gradient Descent with Warm Restarts, *Proceedings of International Conference on Learning Representations* (2017).