卒業研究論文

論 文 題 目

深層学習を用いたドラムセット収録時の被り音抑圧

提出年月日	令和8年　2月　19日
学　　科	電 気 情 報 工 学 科
氏　　名	片山 碧人　　印
指導教員(主査)	北村 大地 准教授　　印
副　　査	柿元 健 准教授　　印
学　科　長	漆原 史朗 教授　　印

香川高等専門学校

Bleeding-sound reduction in drum set recordings using deep learning

Aoto Katayama

Department of Electrical and Computer Engineering
National Institute of Technology, Kagawa College

Abstract

In multitrack drum recording, close microphones capture each instrument but inevitably contain bleeding sound from other sources. This bleeding sound degrades mixing quality because common processing such as equalization and compression affects both the target and non-target components. To address this problem, I focus on bleeding-sound reduction rather than full drum source separation, where I consider three drum sources, kick drum (KD), snare drum (SD), and hi-hat cymbal (HH), for simplicity. In this thesis, I propose a time-domain, end-to-end approach based on Conv-TasNet that takes a waveform as input and outputs an estimated target waveform. The target close-microphone signal is used as the main input, while other close-microphone signals are exploited as auxiliary information. The auxiliary signals are converted into mel-spectrogram features and injected into the separator via feature-wise linear modulation (FiLM), enabling conditional modulation of intermediate representations. A training dataset is created from StemGMD by simulating close-microphone signals with controlled mixing gains. Experiments compare a public pre-trained LarsNet, LarsNet retrained on the created dataset, Conv-TasNet without FiLM, and the proposed method. Spectrogram inspection and SDR evaluation confirm improvement over the input mixture for KD, SD, and HH. The Conv-TasNet architecture is effective for drum bleeding-sound reduction, and FiLM further improves performance. This motivates future work on improved conditioning strategies and artifact reduction to better leverage auxiliary signals.

Keywords: drum set recording, bleeding-sound reduction, Conv-TasNet, Feature-wise Linear Modulation

（和訳）

ドラムセットのマルチトラック録音では，各音源に近接させてマイクロホンを配置するが，他音源の混入 (被り音) を避けることができない．被り音が混入した信号に対して，イコライザやコンプレッサなどの音源個別の処理を適用するミキシング処理を施すと，目的音源だけでなく被り音にも同様の処理が適用され，音源間のバランスが崩れる．本論文は，この被り音問題に対してドラムの全音源を並列に分離するのではなく被り音を抑圧することに焦点を当てる．ただし，キックドラム (KD)，スネアドラム (SD)，およびハイハット (HH) の 3 音源のみを議論の対象とする．提案手法は，音響信号の波形そのものを入出力とする end-to-end な DNN モデルである Conv-TasNet を基本アーキテクチャとして使用した．DNN モデルの主入力は目的音源の近接マイクロホン信号であり，補助情報として他音源の近接マイクロホン信号を使用する．補助情報はメルスペクトログラムに変換し，Feature-wise Linear Modulation (FiLM) によりセパレータの中間特徴へ条件付けを行う．データセットは StemGMD を基に作成し，混合係数を変更して近接マイクロホン信号を模擬した．実験では，公開されている学習済み LarsNet，作成したデータセットで学習した LarsNet，FiLM を使用しない Conv-TasNet，および提案手法を比較した．スペクトログラムによる比較では，入力信号に含まれる被り音成分が推定信号で抑圧されることを確認した．定量評価として SDR 評価で比較した結果，いずれの学習モデルも入力信号より改善し，特に Conv-TasNet 系の手法が高い性能を示した．さらに，FiLM による条件付けは追加の性能向上に寄与することを確認した．今後は，補助情報をより有効に活用するための条件付け設計の改良と，推定信号の音質改善が課題である．

目次

第 1 章　緒言 1

1.1　本論文の背景 1

1.2　本論文の目的 2

1.3　本論文の構成 3

第 2 章　基本理論および従来手法 5

2.1　はじめに 5

2.2　短時間フーリエ変換 5

2.3　メルスペクトログラム 6

2.4　DNN とその構成要素 8

2.5　Conv-TasNet 10

2.6　Feature-wise linear modulation 11

2.7　ドラムセットの音源分離に関する従来手法 13

2.7.1　StemGMD 13

2.7.2　LarsNet 14

2.8　本章のまとめ 15

第 3 章　提案手法 16

3.1　はじめに 16

3.2　提案手法の動機と概要 16

3.3　データセットの作成 17

3.4　DNN の入出力 20

3.5　DNN の構造 21

3.6　DNN の損失関数と学習 22

3.7　本章のまとめ 23

第 4 章　被り音抑圧実験 24

4.1　はじめに 24

4.2　実験条件 24

4.3　実験結果と比較 27

4.4　本章のまとめ 32

第 5 章　結言 34

謝辞 36

参考文献 36

付録 A　追加実験 40
A.1　提案手法をサンプリング周波数 44.1 kHz のデータセットで学習 40
A.2　結果と比較 40

第 1 章

緒言

1.1 本論文の背景

音楽ライブや楽器のレコーディングでは，マイクロホンを用いて人の声や楽器音を音響信号として収録する．収録した信号には，音量を調整するヘッドアンプ（head amplifier: HA），周波数特性を調整するイコライザ，ダイナミクスを制御するコンプレッサなどの信号個別の処理を施した後に，ミキシング処理を施す．マイクロホンを用いた収録方法の一つに，複数のマイクロホンで各音源を個別に録音するマルチトラック録音がある．本論文で焦点を当てるドラムセットの録音においても，マルチトラック録音が広く用いられる．

Fig. 1.1 のようにドラムセットは，単一音源ではなく，キックドラム（kick drum: KD），スネアドラム（snare drum: SD），ハイハットシンバル（hi-hat cymbal: HH），タムタム（tom-tom: TT），シンバル（cymbal: CY）など複数の音源から構成される．録音時には Fig. 1.2 のように各音源にマイクロホンを近接配置し，各音源を個別に収録する．しかし，近接マイクロホンにより得られる信号には，近接させた目的音源だけでなく他の音源成分も混入する．Fig. 1.3 にその模式図を示す．この目的音源以外の音源成分は被り音と呼ばれる．被り音が混入した信号に対して，Fig. 1.4 に示すような音源個別の処理を適用するミキシング処理を施すと，目的音源だけでなく被り音にも同様の処理が適用され，音源間のバランスが崩れる．その結果，ミキシング品質が低下する問題がある．

前述の問題に対処するため，録音信号から目的音源成分を抽出する音源分離や被り音抑圧が提案・利用されてきた．これまで，音源分離を行う手法として，非負値行列因子分解 [1] を時間方向に拡張した畳み込み非負値行列因子分解 [2] などが提案され，ドラムセットを含む様々な楽器の音源分離が検討されてきた．しかし，これらの手法によって得られる分離音の品質は，音楽制作現場における実用に足るとは言えず，実環境への適用性の観点で課題が残されている．そのため，ドラムセットを対象とする研究の多くは，分離音の高音質化よりも自動採譜を主目的とする方向で進められてきた [3].

近年，計算機性能の向上と深層学習技術の発展により，深層ニューラルネットワーク（deep neural network: DNN）を用いた音源分離・被り音抑圧が盛んに研究されており，ドラムセットに対しても DNN を用いた音源分離の研究が登場した [4]．DNN は大量データから特徴を学

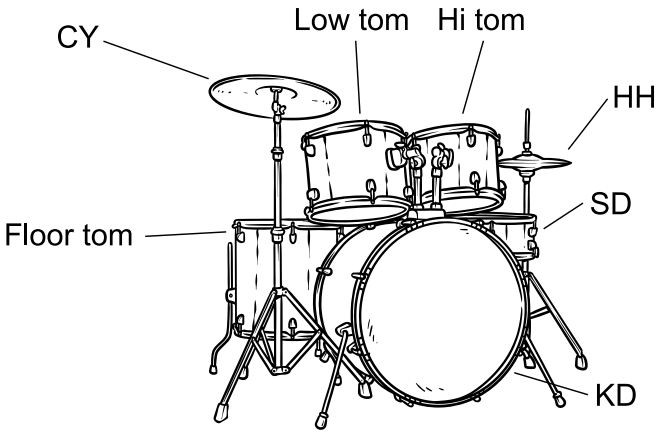


Fig. 1.1 Sound source components of drum set.

(a)　　(b)　　(c)

Fig. 1.2 Typical close-microphone setup in drums recording for (a) KD, (b) SD, and (c) HH.

習できる一方で，教師データの整備が課題となる．特に音源分離では，音源ごとに分離された信号を大量に用意する必要があり，データ収集および整備の困難さが指摘されてきた．しかし近年は，2023 年に公開された StemGMD [4]（関連論文は 2024 年）というドラムセットの大規模データセットやデータ生成手法の整備により，ドラムセットの音源分離へ DNN を適用する環境が整いつつある．一方で，文献 [4] で提案された DNN 手法では，入力をシングルチャネル信号に限定しており，マルチトラック録音を前提とする実運用の条件と整合しない．さらに，StemGMD は公開から日が浅いこともあり，ドラムセットの収録で得られるマルチチャネルの信号を活用した音源分離の設計と検証は十分に行われていない．

1.2　本論文の目的

前節で述べたように，既存のドラムセットを対象とした音源分離 [4] では，StemGMD に含まれるドラムセットの各音源の個別収録された信号を加算して得られる信号（シングルチャネル信号）を入力とし，全音源を同時に推定する音源分離モデルを構築している．しかし，音楽ライブ演奏や音楽制作におけるドラムセット収録では，Fig. 1.2 のように各音源に近接マイクロホンを配置するマルチトラック録音のマルチチャネル観測信号が得られる．本論文では，マルチトラック録音の各近接マイクロホンに混入する被り音の高精度な抑圧を目的とし，文献

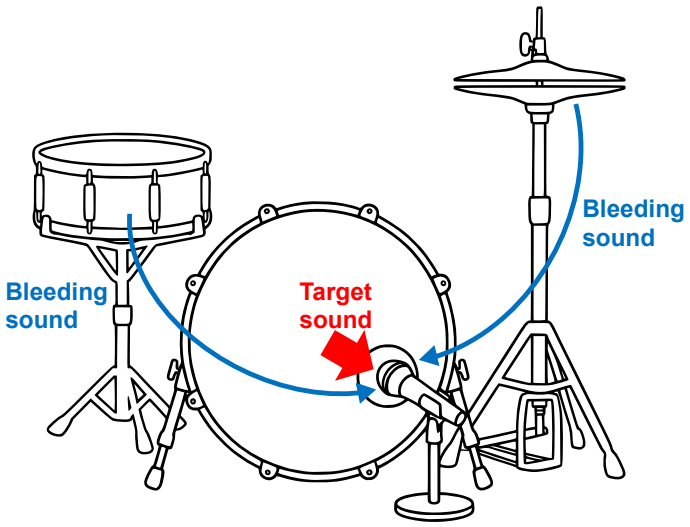


Fig. 1.3 A target sound and bleeding sounds during drum set recording.

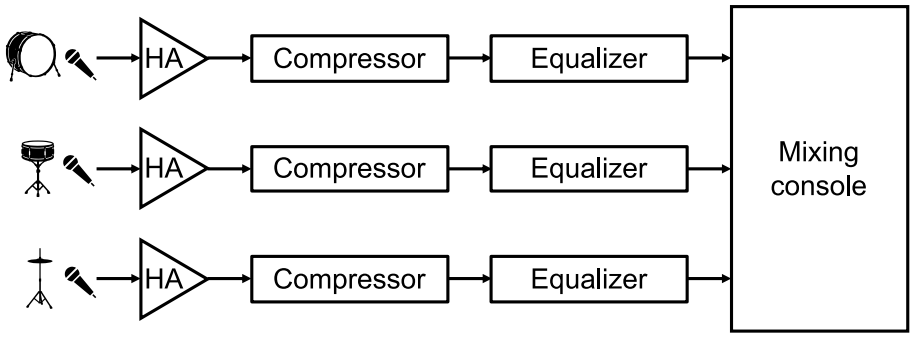


Fig. 1.4 Block diagram of mixing system.

[4] で提案された手法のように全音源を同時推定する音源分離モデルではなく，各音源の近接マイクロホン信号に混入する被り音を抑圧するモデルを音源ごとに構築するアプローチをとる．これはすなわち，KD の近接マイクロホン専用の被り音抑圧 DNN や，SD の近接マイクロホン専用の被り音抑圧 DNN モデルなどを学習することに相当する．またこれらの DNN モデルは，特定の音源の近接マイクロホン信号をシングルチャネル信号として入力するだけでなく，その信号に混入する被り音の音源の近接マイクロホン信号を補助情報として追加で入力する構造を持たせる．この補助情報は，被り音となる音源の情報を多く含んでいるため，提案 DNN モデルの中で被り音抑圧における重要なヒントとして活用されることを期待している．なお，本論文では被り音抑圧の基礎検討段階であることから，以降 KD，SD，および HH の 3 音源のみの観測信号および被り音抑圧問題を議論の対象とする．

1.3 本論文の構成

本論文は以下のように構成される．2 章では，音響信号処理および DNN の基礎理論と従来手法について説明する．基礎理論として短時間フーリエ変換（short-time Fourier transform:

STFT）[5]，DNN の畳み込み層，およびエンコーダ/デコーダ構造を述べ，提案手法で用いる DNN アーキテクチャについて説明する．また，従来の DNN 手法に基づく音源分離手法についても整理する．3 章では，本論文で提案する被り音抑圧手法の動機と詳細について記述する．特にデータセットの作成，DNN の入出力設計，ネットワーク構造など，従来手法との相違点を明確化する．4 章では，提案手法の有効性を検証するための実験を行い，従来手法による実験結果と比較する．5 章では，本論文で得られた知見を総括し，今後の課題と改善案について述べる．

第 2 章

基本理論および従来手法

2.1 はじめに

前章で述べたように，本論文ではマルチトラック録音で得られる近接マイクロホン信号を入力とする DNN モデルを提案するため，補助情報をどのような特徴量として表現し，その情報をネットワーク内部へどのように反映するかを明確化する必要がある．本章ではまず，2.2 節および 2.3 節で信号の表現として用いる STFT およびメルスペクトログラムについて説明する．次に，2.4 節および 2.5 節で DNN の要素として，畳み込み層，エンコーダ・デコーダ構造を説明し，提案手法の基盤となる Conv-TasNet について説明する．さらに，2.6 節では補助情報を DNN の中間層へ反映する手段を説明し，2.7 節では従来手法の DNN 構造，入出力，および学習設定を整理しすることで，提案手法の比較対象を明確にする．最後に，2.8 節で本章をまとめる．

2.2 短時間フーリエ変換

STFT は，Fig. 2.1 に示すように，時間変化する音響信号を時間周波数表現であるスペクトログラムに変換するための手法である [5]．STFT は，時間領域の信号を短時間区間ごとに切り出し，それぞれに窓関数を乗算したうえで周波数表現へと変換する．

信号長 L の信号 $\tilde{x}[l]$ の STFT を考える．ここで，$l = 1, 2, \cdots, L$ は離散時間サンプルのインデクスである．STFT において，時間領域から周波数領域への変換時の窓の長さを H，シフト長を ζ とする（Fig. 2.1 参照）．STFT では，各時間フレームの周波数成分は入力信号の有限長区間を窓関数で切り出して計算されるため，窓長 H は各時間フレームで参照される時間範囲を規定している．このように，ある出力を得るために参照される入力信号の範囲は一般に受容野と呼ばれる．STFT で得られる時間フレーム数 J は次式で求められる．

$$J = \frac{L}{\zeta} \tag{2.1}$$

ここで，時間フレーム数 J が整数となるように信号 $\tilde{x}[l]$ の終端にゼロを挿入する処理（ゼロパディング）が施されている（ゼロパディング後の信号長を L と解釈する）．周波数ビン数 I

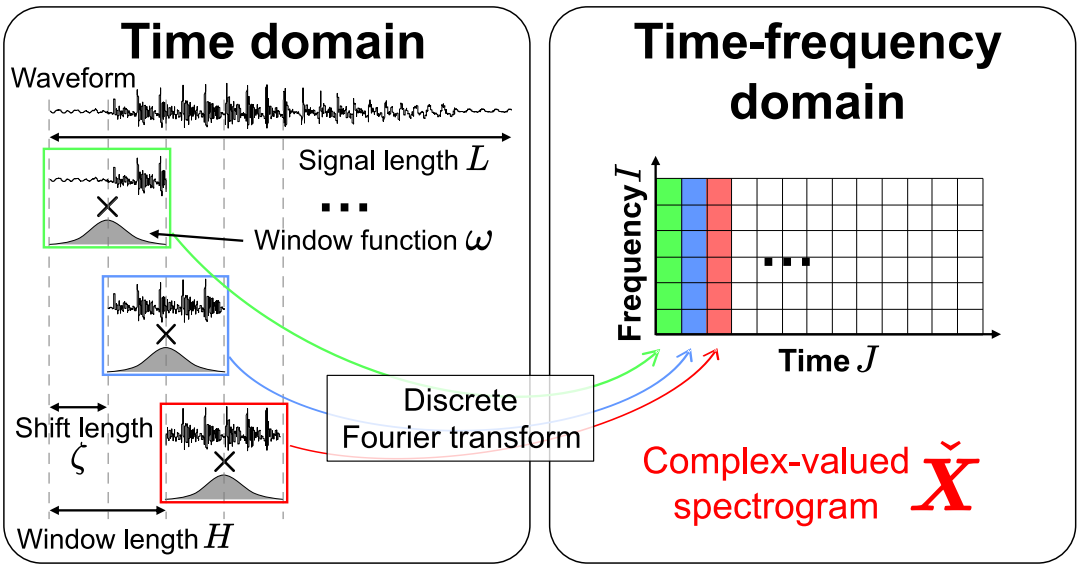


Fig. 2.1 Mechanism of STFT.

は $I = \lfloor H/2 \rfloor + 1$ を満たす整数（$\lfloor \cdot \rfloor$ は床関数）となる．STFT によって得られるスペクトログラム $\check{\mathbf{X}} \in \mathbb{C}^{I \times J}$ の各時間要素は次式で表される．

$$\check{\mathrm{x}}[i,j] = \sum_{h=1}^{H} \check{x}[h + \zeta(j-1)]\omega[h] \exp\left[-2\pi\iota \frac{(h-1)(i-1)}{H}\right] \tag{2.2}$$

ここで，ι は虚数単位，$\omega[h]$ は窓関数，$i = 1, 2, \cdots, I$ は周波数ビンのインデクス，$j = 1, 2, \cdots, J$ は時間フレームのインデクス，$h = 1, 2, \cdots, H$ は時間フレーム内のインデクスを示し，時間周波数領域での信号（複素数）は立体フォントとする．このように，時間領域の信号は一定幅の短時間ごとに窓関数を乗じて離散フーリエ変換を行うことで，列が時間フレーム，行が周波数ビンのスペクトログラムと呼ばれる複素行列 $\check{\mathbf{X}}$ で表すことができる．複素スペクトログラムは各時間周波数の振幅成分と位相成分を持っているが，音響信号処理では振幅成分のみを取り扱うことが多い．その場合は振幅スペクトログラム $\check{\mathbf{X}}$ の各要素に関して絶対値を取った振幅スペクトログラム $|\check{\mathbf{X}}| \in \mathbb{R}_{\geq 0}^{I \times J}$ や，絶対値の 2 乗を取ったパワースペクトログラム $|\check{\mathbf{X}}|^{\cdot 2} \in \mathbb{R}_{\geq 0}^{I \times J}$ を処理の対象とする．ここで，行列に対する絶対値記号およびドット付き指数乗はそれぞれ要素ごとの絶対値および要素ごとの指数乗を表す．

2.3　メルスペクトログラム

前節で述べた通り，音響信号は STFT によって時間周波数領域の信号（スペクトログラム）に変換することができる．さらにこの信号を，人間の聴覚特性に基づく周波数尺度に変換する手法として，メルフィルタバンクを用いた次元圧縮が用いられる [6]．Fig. 2.2 にバンドパスフィルタ数 $P = 16$，下限周波数 $f_{\min} = 0$ Hz，および上限周波数 $f_{\max} = 22050$ Hz の場合のメルフィルタバンクを示す．メルフィルタバンクとは，Fig. 2.2 に示す三角状のバンドパス

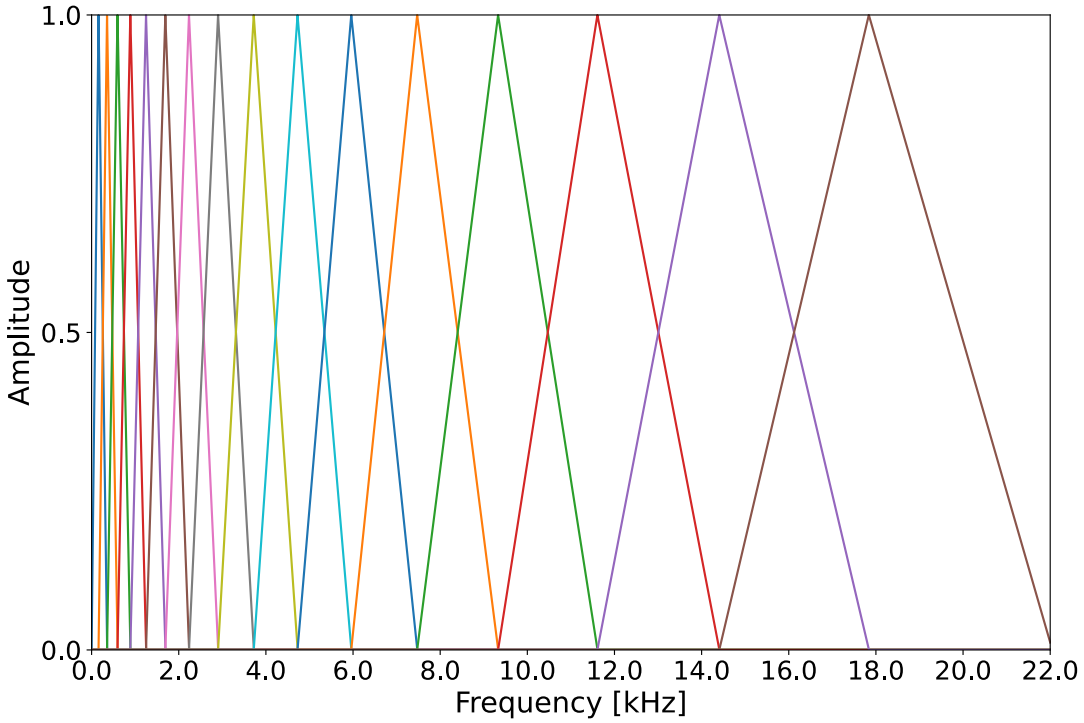
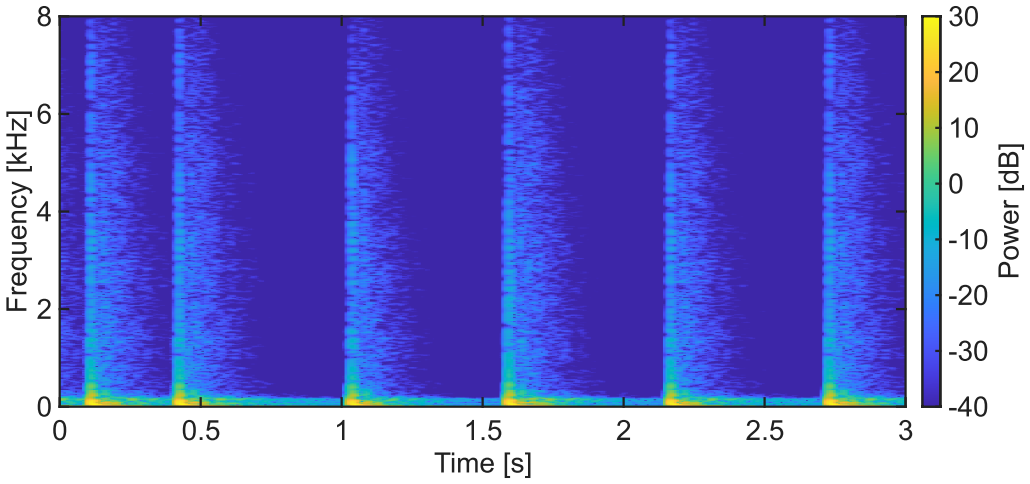


Fig. 2.2 Mel-filter bank ($f_{\min} = 0$ Hz, $f_{\max} = 22050$ Hz, $P = 16$).

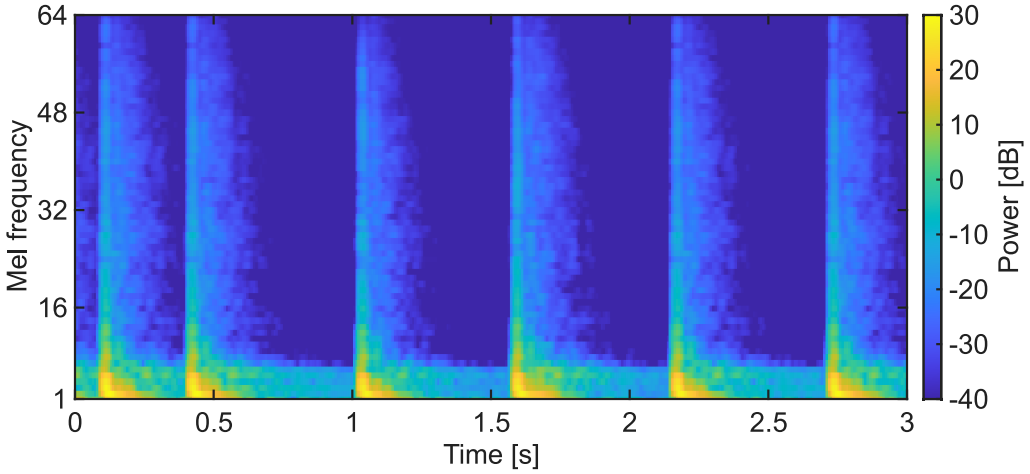
フィルタを用いて周波数軸を分割し，各帯域のエネルギーを抽出して特徴量化する手法である．STFT により得られたパワースペクトログラム $|\check{\mathbf{X}}|^{\cdot 2} \in \mathbb{R}_{\geq 0}^{I \times J}$ に対して，メルフィルタバンク行列 $\mathbf{W}_{\text{mel}} \in \mathbb{R}_{\geq 0}^{P \times I}$ を適用することで，メルスペクトログラム $\check{\mathbf{X}}_{\text{mel}} \in \mathbb{R}_{\geq 0}^{P \times J}$ を得る．

$$\check{\mathbf{X}}_{\text{mel}} = \mathbf{W}_{\text{mel}} |\check{\mathbf{X}}|^{\cdot 2} \tag{2.3}$$

ここで，$\mathbf{W}_{\text{mel}}$ は，STFT の I 次元の周波数軸を，聴覚特性に基づく P 次元のメル周波数ビンに次元圧縮する行列として定義される．Fig. 2.3 (a) に KD の演奏を近接マイクロホンで観測した際の音響信号のパワースペクトログラム，Fig. 2.3 (b) に同一信号のメルスペクトログラムを示す．ただし，窓関数はハン窓，STFT の窓長は 1024 点（64 ms），シフト長は 128 点（8 ms），メルスペクトログラムの次元 P は 64 である．パワースペクトログラムは，STFT で得られる周波数ビンのパワーを線形周波数軸上にそのまま表示する．一方，メルスペクトログラムは STFT のパワースペクトルをメルフィルタバンクで帯域ごとに集約した表現であり，周波数方向の次元数が削減されることで細部が平滑化される．また，メルフィルタバンクは高周波ほどバンド幅が広くなるため，高周波帯域ではエネルギーがより広く平均化さる特徴を持つ．パワースペクトログラムとメルスペクトログラムを比較すると，メルスペクトログラムはパワースペクトログラムに比べ縦軸方向が荒くなっていることが分かる．



(a)



(b)

Fig. 2.3 Example of (a) power and (b) mel spectrograms for KD source observed by close microphone.

2.4　DNN とその構成要素

DNN は機械学習の手法のひとつであり，生物の脳を構成するニューロンという神経細胞を模倣することで様々なタスクを解くことができる．タスクの種類には回帰問題，他クラス分類問題などがある．また，回帰問題は予想する値が無限に存在するので無限個のクラスがある分類問題と解釈することができる．音源分離や被り音抑圧は，複数の音源が混合された信号を入力とし，特定の音源の音響信号を推定するため回帰問題に分類される．

DNN には層のつなぎ方やデータの処理方法によって，いくつかの代表的な種類に分けることができる．前の層の全てのニューロンが次の層の全てのニューロンと結合される全結合型ニューラルネットワーク，畳み込み層とプーリング層を持つ畳み込みニューラルネットワーク

(convolutional neural network: CNN) [7]，過去の情報を次の入力に使用する再帰型ニューラルネットワーク [8] などがある．

本節では，次章で述べる提案手法で用いる畳み込み層について詳しく説明する．畳み込み層は，入力データから局所的な特徴を畳み込み演算により抽出するために用いられる．本論文では，以後畳み込み層を Conv 層と呼び，畳み込み演算を行う次元数に応じて Conv1D 層，Conv2D 層，Conv3D 層と表記する．本節では Conv1D 層を例に挙げて畳み込み演算の説明を行う．

Conv1D 層は，入力信号の 1 つの次元に対して畳み込み演算を行う層である．従って，音響信号のような時系列信号に対して適用されることが多い．信号長が L でチャネル数が C の入力を $\boldsymbol{X} \in \mathbb{R}^{C\times L}$ と定義し，その要素を $x_c[l]$ と表す．ここで，$c = 1, 2, \cdots, C$ および $l = 1, 2, \cdots, L$ はそれぞれチャネルのインデクスおよび離散時間インデクスを表す．このとき，Conv1D 層の入出力関係は次式で表せる．

$$\begin{aligned} y_\delta[\xi] &= \epsilon_\delta + \sum_c \boldsymbol{w}_{\delta,c} * \boldsymbol{x}_c[\xi] \\ &= \epsilon_\delta + \sum_c \sum_{v=1}^{V} w_{\delta,c}[v] x_c[\kappa(\xi-1) + (v-1) + 1] \end{aligned} \tag{2.4}$$

ここで，$\boldsymbol{w}_{\delta,c}$ および $\boldsymbol{x}_c[\xi]$ はそれぞれ学習可能フィルタ（カーネル）および入力信号の局所時間ベクトルであり，次式で定義される．

$$\boldsymbol{w}_{\delta,c} = [w_{\delta,c}[1], w_{\delta,c}[2], \cdots, w_{\delta,c}[v], \cdots, w_{\delta,c}[V]]^{\mathrm{T}} \in \mathbb{R}^V \tag{2.5}$$

$$\boldsymbol{x}_c[\xi] = [x_c[\kappa(\xi-1)+1], x_c[\kappa(\xi-1)+2], \cdots, x_c[\kappa(\xi-1)+V]]^{\mathrm{T}} \in \mathbb{R}^V \tag{2.6}$$

また，$\delta = 1, 2, \cdots, \Delta$ は出力のチャネルインデクス，$\xi = 1, 2, \cdots, \Xi$ は出力の時間フレームインデクス，κ はストライド長，V はカーネル長，$y_\delta[\xi]$ は δ 番目のチャネルの時間フレーム ξ に対応する出力，ϵ_δ は δ 番目の出力チャネルのバイアス，演算子 $*$ はベクトル間の畳み込み演算（ただし，厳密にはベクトル間の相互相関演算）である．Fig. 2.4 は式 (2.4) を図で示したものである．ただし，入力のチャネル数 $C = 1$，出力のチャネル数 $\Delta = 2$，信号長 $L = 15$，フィルタ長 $V = 3$，ストライド長 $\kappa = 3$ の例を図示しており，さらに式 (2.4) 中のバイアス ϵ_δ の加算は省略している．畳み込み層では入力の局所時間信号とフィルタを Fig. 2.4 のように計算した結果を出力する．また，カーネルは出力チャネル数 Δ に応じた数だけ用意される．なお，Conv1D 層の逆演算を実現するものとして，ConvTranspose1D 層も存在する．特に，次に述べるエンコーダ・デコーダモデルにおいてエンコーダに Conv1D 層を用いた場合は，デコーダに ConvTranspose1D 層を用いることが一般的である．

最後に，DNN を用いたネットワークアーキテクチャによく含まれるエンコーダとデコーダについて説明する．エンコーダは，入力データを低次元な潜在表現へと非線形写像し，逆にデコーダは潜在表現からデータへの復元処理を行うように学習する．この一連のプロセスは従来の音響信号処理で用いられる STFT および逆 STFT（inverse STFT: ISTFT）による処理系と類似している．しかし，STFT があらかじめ定義された正弦波を基底関数として用いるの

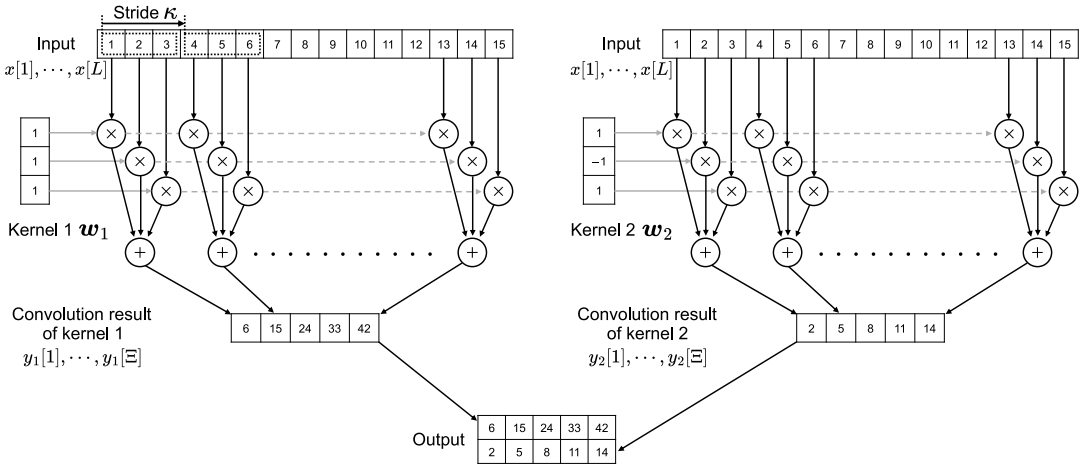


Fig. 2.4 Conceptual diagram of Conv1D, where $L = 15$, $C = 1$, $\delta = 2$, $V = 3$, and $\kappa = 3$. Note that the addition of bias parameters (ϵ_1 and ϵ_2) is omitted.

に対し，エンコーダおよびデコーダは学習可能なフィルタ（カーネル）で構成される点が異なる．これにより，音源分離などの特定のタスクに最適化された特徴量抽出が可能となる利点がある．

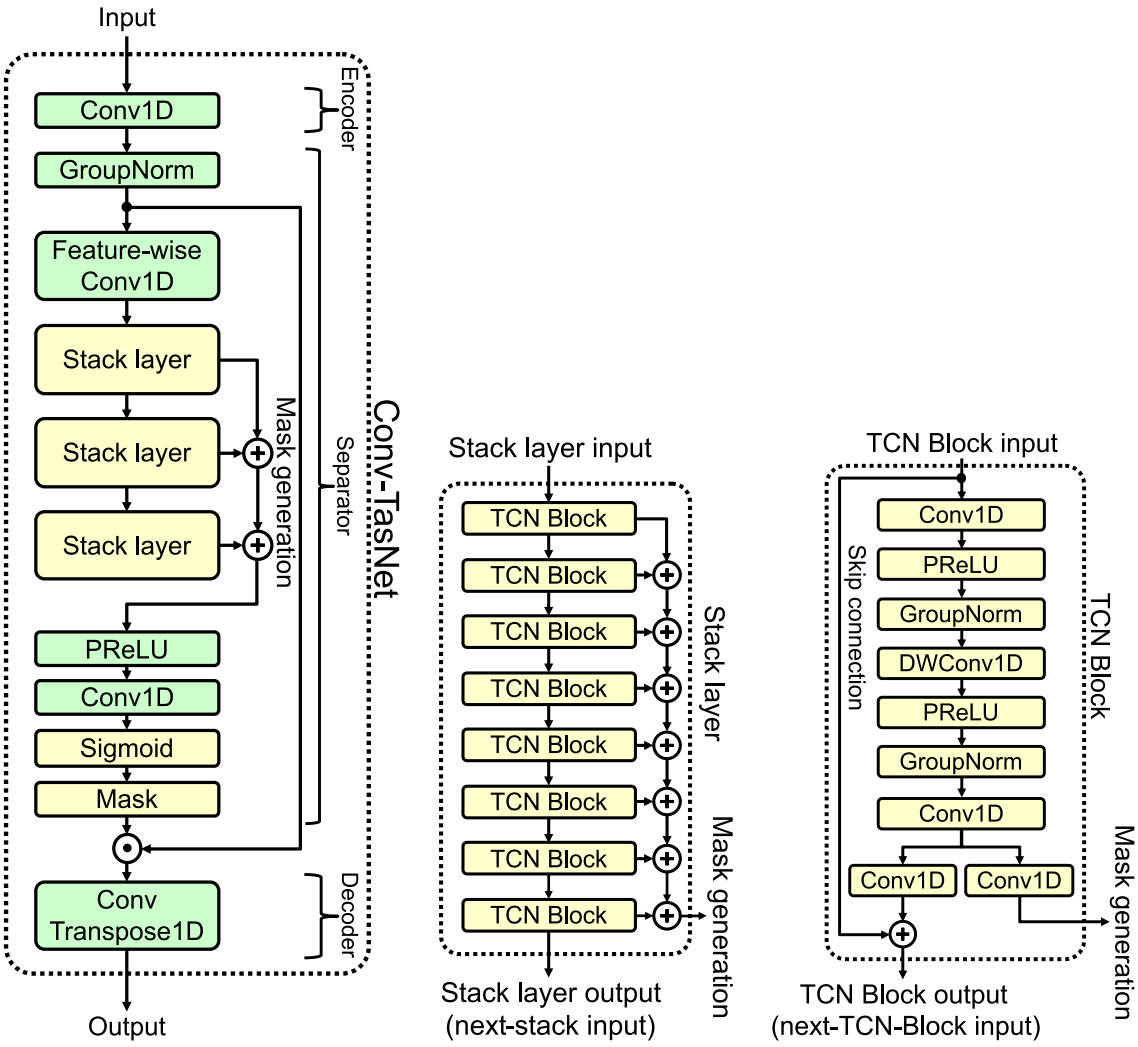
2.5　Conv-TasNet

音響信号の波形そのものを入出力とする end-to-end な音源分離 DNN モデルとして提案された手法に，完全畳み込み時間領域音声分離ネットワーク（fully convolutional time-domain audio separation network: Conv-TasNet）[9] がある．スペクトログラムやメルスペクトログラムなどにエンコードされた時間周波数表現を必要とせず，学習可能な解析・合成処理を内部に持つ点が特徴である．本論文では，Conv-TasNet の強力な音源分離性能を最大限活用するため，提案手法の基本アーキテクチャに Conv-TasNet を用いる．

Conv-TasNet は，Fig. 2.5 に示すように，大きくエンコーダ，セパレータ，デコーダの 3 要素から構成される．エンコーダで得た潜在表現 $\boldsymbol{E}$ に対してセパレータがマスク $\boldsymbol{M}$ を推定し，そのマスクを適用した潜在表現 $\hat{\boldsymbol{E}}$ を次式のように得る．

$$\hat{\boldsymbol{E}} = \boldsymbol{M} \odot \boldsymbol{E} \tag{2.7}$$

ここで，$\odot$ は要素ごとの積を表す．最後に推定された $\hat{\boldsymbol{E}}$ をデコーダで波形へと復元する．なお，セパレータの内部構造は時系列畳み込みネットワーク（temporal convolutional network: TCN）と呼ばれている．また，Fig. 2.5 中の GroupNorm はグループ正規化層を表し，学習安定化のために特徴量の特定のグループに関して正規化を行う．また，PReLU と Sigmoid はそれぞれアクティベーション関数としてよく用いられる非線形関数のパラメトリック正流線形関数（parametric rectified linear unit: PReLU）およびシグモイド関数である．さらに，DWConv は depth-wise 畳み込み層であり，通常の畳み込み層と比較して，カーネルによる入力チャネル間の線形結合がされないという違いがある．式 (2.4) に示した Conv1D の計算を



(a) (b) (c)

Fig. 2.5 Architecture of Conv-TasNet: (a) overall network comprising encoder, separator, and decoder with mask estimation, (b) expanded view of Stack layer in (a), and (c) expanded view of a TCN Block in (b).

depth-wise Conv1D に置き換えた場合，次式となる．

$$y_c[\xi] = \epsilon_c + \sum_{v=1}^{V} w_c[v] x_c[\kappa(\xi - 1) + (v - 1) + 1] \tag{2.8}$$

2.6 Feature-wise linear modulation

次章で述べる提案手法では，Conv-TasNet に対して特徴量的線形変調（feature-wise linear modulation: FiLM）[10] と呼ばれるネットワークを融合している．この FiLM は，Fig. 2.6 に示すように中間層の特徴量に対してチャネルごとのスケーリングおよびシフトを適用する

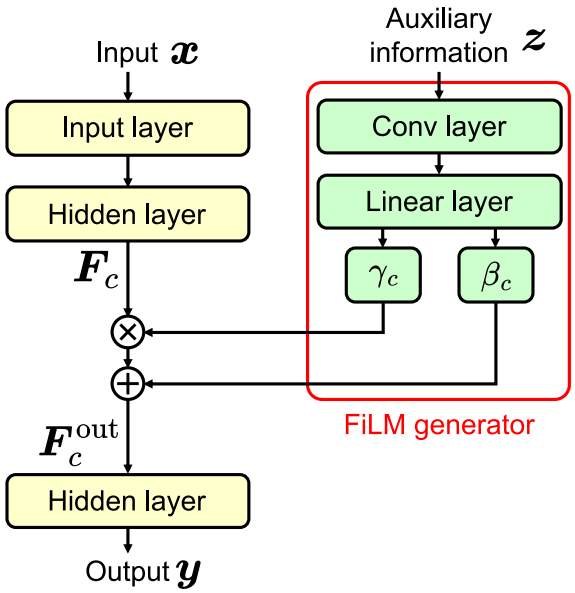


Fig. 2.6 DNN overview with FiLM applied to intermediate features.

ことで，FiLM の入力（補助情報）に応じた表現の変調を実現する一般的な条件付け層である [10]．FiLM を適用する中間層の特徴量を F とする．Conv2D 層の出力を例にすると，特徴量は $\mathsf{F} \in \mathbb{R}^{C\times V\times W}$ と表せる．ここで C はチャネル数，V と W は特徴量の次元であり，c 番目のチャネルの特徴量を $\boldsymbol{F}_c \in \mathbb{R}^{V\times W}$ と表す．FiLM は各チャネル c に対するスケーリング係数 $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \cdots, \gamma_C]^{\mathrm{T}} \in \mathbb{R}^C$ とバイアス係数 $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_C]^{\mathrm{T}} \in \mathbb{R}^C$ を用いて特徴量 $\boldsymbol{F}_c$ を変調し，出力特徴量 $\boldsymbol{F}_c^{\text{out}} \in \mathbb{R}^{V\times W}$ を次式で与える．

$$\boldsymbol{F}_c^{\text{out}} = \gamma_c \cdot \boldsymbol{F}_c + \beta_c \tag{2.9}$$

ただし，式 (2.9) の $\gamma_c \cdot \boldsymbol{F}_c$ と β_c の和は特徴量 $\boldsymbol{F}_c$ の次元（$V \times W$）に拡張して計算される．また，$\boldsymbol{F}_c^{\text{out}}$ を全てのチャネルについて含むテンソルを $\mathsf{F}^{\text{out}} \in \mathbb{R}^{C\times V\times W}$ とする．すなわち FiLM は Fig. 2.7 に表されるように，補助情報に応じて中間特徴の各チャネルを線形変調することで，現在の損失関数をより小さくすることができるように中間表現を調整する．

FiLM における $\boldsymbol{\gamma}$ および $\boldsymbol{\beta}$ は固定値ではなく，補助情報および損失関数に基づいて学習される係数である．本論文では，FiLM に入力する補助情報を $\boldsymbol{z}$ と定義する．$\boldsymbol{z}$ から $\boldsymbol{\gamma}$ および $\boldsymbol{\beta}$ を生成するネットワークは FiLM ジェネレータと呼ばれる（Fig. 2.6 参照）．FiLM ジェネレータを $f_\theta(\cdot)$ とおくと，$\boldsymbol{\gamma}$ および $\boldsymbol{\beta}$ は次式で表される．

$$(\boldsymbol{\gamma}, \boldsymbol{\beta}) = f_\theta(\boldsymbol{z}) \tag{2.10}$$

ただし，θ は FiLM ジェネレータが内包する学習可能なパラメータの集合である．FiLM は，入力段で特徴を $[\boldsymbol{x}, \boldsymbol{z}]$ のように単純に連結して入力する方法と異なり，$\boldsymbol{x}$ のみから得られる中間表現の各チャネルを補助情報 $\boldsymbol{z}$ に応じて直接変調できる点に特徴がある．

FiLM は，視覚質問応答タスクにおける画像と言語を用いた視覚推論 [10] において有効性が示されて以来，画像と音声，音声とテキストといったマルチモーダル学習への応用 [11, 12] や，

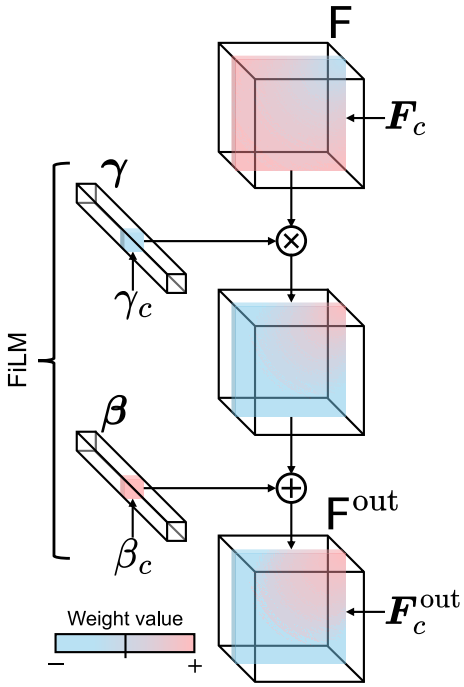


Fig. 2.7 A single FiLM layer for a Conv2D output, where each two-dimensional feature is scaled and shifted by γ and β, respectively.

音声信号やその他の時系列データに対する条件付き変換 [13] など，幅広く利用されてきた．さらに，解剖学的画像のノイズ除去タスクにおける空間適応的変調 [14] や，グラフ構造データへの適用 [15] も報告されており，応用範囲は非常に広範である．これらの文献では，FiLM が多様な条件付き変調に有効であることを示している．

2.7　ドラムセットの音源分離に関する従来手法

2.7.1　StemGMD

教師ありの DNN 学習には，入力データとそれに対応する正解データを大量に用意する必要がある．ドラムセットの音源分離や被り音抑圧においても，同様に複数の音源が混合した観測信号と各音源の混合前の信号が必要となる．しかし，ドラムセットを対象としたデータセットは数が少なく，規模も限られていた．そこで Mezza らは，ステムを含むドラムセットの大規模なデータセットとして StemGMD を構築した [4]．ここで，ステムとは楽曲を構成する各要素を個別にした音源信号を指す．例えば，ドラムセットでは，KD，SD，HH などをそれぞれ単独で含む信号をステムと呼ぶ．StemGMD は，Google Magenta というチームにより公開された Groove MIDI dataset（GMD）[16] を基に構築されたデータセットである．GMD は 13.6 時間のドラムトラックからなる大規模コーパスであり，各演奏について MIDI ファイルと対応するフルキットの混合信号が含まれている．ここで MIDI は音響信号そのものではなく，どの打撃をいつ，どの程度の強さで演奏したかといった演奏情報（叩き方）を記録したデータ

である．MIDI ファイルをソフトウェア音源に入力することで，対応する打撃音を合成し，音響信号として出力できる．しかし，GMD にはドラムセットのステムが含まれていなかった．StemGMD は，GMD にはなかったドラムセットのステムを含み，さらに元の 22 チャネルを 9 つのチャネル（KD，SD，ハイタム，ロータム，フロアタム，オープン HH，クローズド HH，クラッシュ CY，およびライド CY）に整理したデータセットである．GMD に含まれる MIDI ファイルをドラムセットのソフトウェア音源に入力し，各チャネルのステムを合成することで，総再生時間 1224 時間に達する大規模なデータセットへと拡張した．各チャネルのステムはソフトウェア音源により合成されており，それらを加算することで混合信号が得られる．そのため，混合信号とステムが対応付けられた教師あり学習用データとして利用できる．また，各チャネルのステムの合成にはソフトウェア Logic Pro X に含まれる 10 種類の異なるドラムキットを用いており，幅広い音色をカバーしている．

2.7.2　LarsNet

StemGMD が提案された文献 [4] において，ドラムセット音源分離を実現するベースラインネットワークモデルとして LarsNet が提案されている．LarsNet は，ステレオのドラムミックスから 5 つのステムを推定する深層学習モデルである．LarsNet の構造を Fig. 2.8 に示す．推定対象は KD，SD，HH，TT，および CY である．ただし，TT はハイタム，ロータム，およびフロアタムを統合，HH はオープン HH およびクローズド HH を統合，CY はクラッシュシンバルおよびライドシンバルを統合したステムを対象としている．LarsNet の内部では，時間周波数領域の信号を入力する U-Net [17] を 5 本並列に配置した構成で，各 U-Net がそれぞれ 1 つのステムに対応する時間周波数領域のソフトマスクを推定する．U-Net への入力はステレオの混合信号から計算した 2 チャネルの振幅スペクトログラム $\mathsf{X} \in \mathbb{R}_{\geq 0}^{2\times I\times J}$ であり，出力は入力と同サイズのマスク $\mathsf{M}_u \in [0,1]^{2\times I\times J}$ である．ここで，$u = 1, 2, \cdots, 5$ は 5 つのステムのインデクスを表す．そして，推定マスクを観測信号の振幅スペクトログラムに適用し，混合信号の位相スペクトログラムを付与して ISTFT により時間波形を復元する．LarsNet の損失関数は正解となるステム信号の振幅スペクトログラム $\mathsf{X}_u \in \mathbb{R}_{\geq 0}^{2\times I\times J}$ と予測された各ステムの振幅スペクトログラム間の L_1 ノルムとし，次式のように定義される．

$$\text{Minimize} \ \left\|\mathsf{X}_u - \mathsf{M}_u \odot \mathsf{X}\right\|_1 \tag{2.11}$$

ここで $\|\cdot\|_1$ は L_1 ノルム，$\odot$ は要素ごとの積を表す．

LarsNet における学習データの準備では，9 つのステムを単純に時間領域で加算して観測信号を作成している．そして，この観測信号を入力とし，KD，SD，HH，TT，および CY の 5 つのステムを並列に推定する一つの DNN モデルを学習している．従って，入力はステレオの混合信号のみであり，出力は 5 つのステレオのステム（音源信号）である．

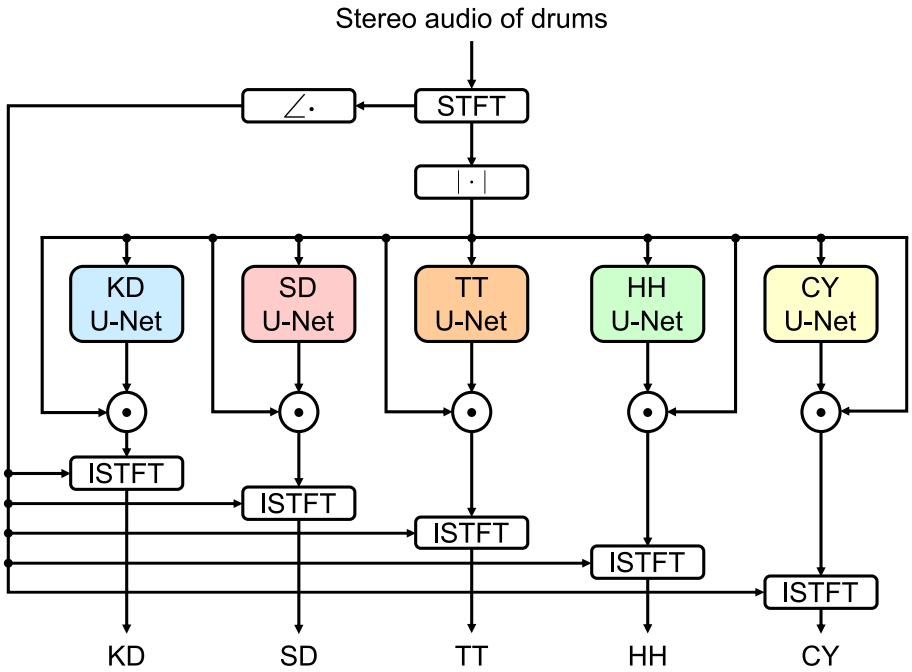


Fig. 2.8 Architecture of LarsNet, where $\angle\cdot$ and $|\cdot|$ show phase and amplitude operations for complex-valued spectrogram, respectively. LarsNet contains stem-specific U-Nets that predict time-frequency masks.

2.8 本章のまとめ

本章では，被り音抑圧を扱うための信号表現（STFT およびメルスペクトログラム）および提案手法で用いる DNN の基本要素を明確化した．また，提案手法で用いる Conv-TasNet と FiLM の仕組みを説明し，モデルで扱う構成要素を明確化した．さらに，従来手法として LarsNet を取り上げ DNN 構造，入出力，および学習設定について整理した．次章では，これらの整理を踏まえ，データセット作成と入出力設計を含む提案手法の具体的構成を述べる．

第 3 章

提案手法

3.1 はじめに

本章では，マルチトラック録音で得られる近接マイクロホン信号を活用し，被り音を抑圧する DNN モデルを提案する．2 章で説明した Conv-TasNet を基本アーキテクチャとし，補助情報を FiLM を介してセパレータ内部の中間特徴へ条件付けを行うことで被り音抑圧を行う DNN モデルを説明する．まず，3.2 節で提案手法で近接マイクロホン信号を使用するに至った動機および提案手法の概要を説明する．3.3 節では提案手法の学習および評価で使用する近接マイクロホン信号を模擬的に生成した方法を説明する．提案手法の DNN モデルについては，3.4 節で入出力，3.5 節で構造，3.6 節で損失関数と学習の順で説明する．最後に，3.7 節で本章をまとめる．

3.2 提案手法の動機と概要

ドラムセットのマルチトラック録音における近接マイクロホン信号には，近接させた目的音源だけでなく他の音源成分も混入する被り音問題が存在する．Fig. 1.4 のように被り音が混入した信号に対してミキシング処理を施すと，目的音源だけでなく被り音にも同様の処理が適用され，音源間のバランスが崩れることになる．従って，近接マイクロホン信号から被り音成分を抑圧する必要がある．

2.7.2 項で述べた既存のドラムセット音源分離手法である LarsNet は，被り音抑圧問題にそのまま適用できる可能性が高い．しかしながら，音楽信号における音源分離は，芸術性を失わないほどの高精度な処理が求められる難しさがある．その観点で既存手法の分離精度は必ずしも十分とはいえず，歪みや他音源の残留が問題として残る．LarsNet はドラムセット全体（全音源）を含むステレオの観測信号を入力としており，一つのモデルで全ての音源の分離信号を出力する DNN モデルである．この構成はドラムセット音源分離を包括的に表現したものであるが，難しいタスクの達成を目指した最適化であることから，学習すべき特徴量が膨大となり，予測精度の劣化につながっている可能性がある．

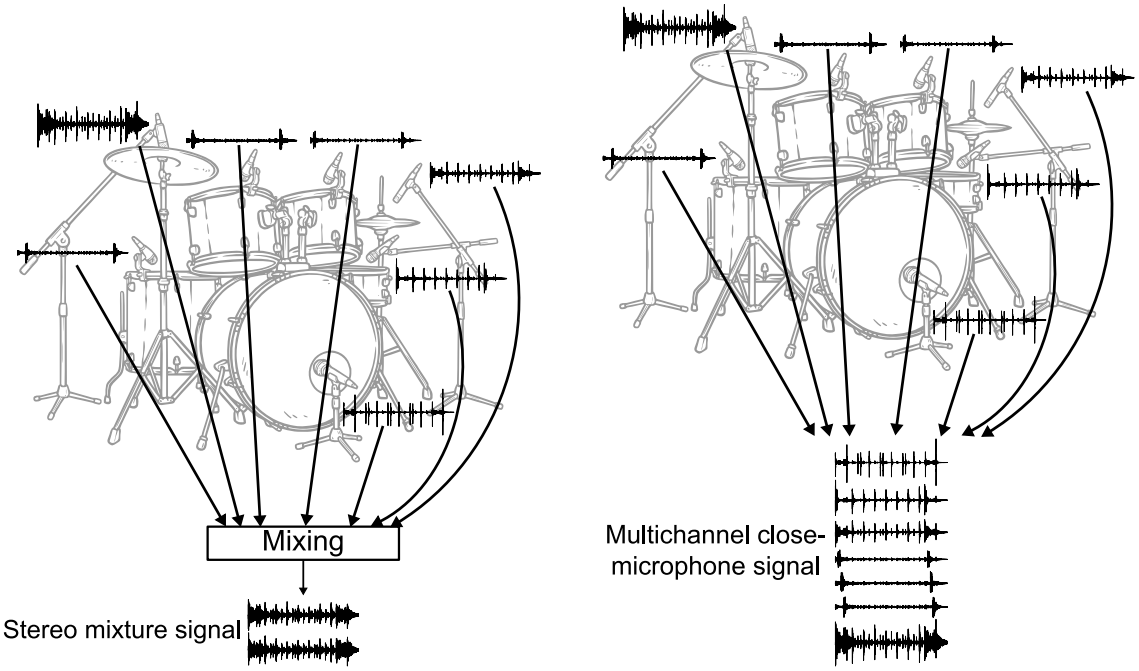
Fig. 3.1 に，LarsNet および提案手法が想定する観測信号の違いを示す．本論文で焦点を当てる被り音抑圧では，LarsNet の想定する観測信号（Fig. 3.1 (a)）とは異なり，マルチトラック録音されたマルチチャネルの観測信号（Fig. 3.1 (b)）を入力として与えることができる．従って，各音源の近接マイクロホンの録音信号を用いることができ，この観測信号に特化した被り音抑圧のアプローチや DNN モデルを考えることで，LarsNet を直接適用する場合よりも高精度な被り音抑圧を期待できる．マルチトラック録音における近接マイクロホン信号には，近接させた目的音源成分が他チャネル（他の近接マイクロホン信号）より相対的に大きな音量で含まれている．例えば Fig. 3.2 のように，KD に近接させたマイクロホン信号には KD 成分が強く含まれるため，KD 成分の抽出は相対的に容易になる可能性が高い．従って，このような観測信号に対しては，Fig. 3.3 のように特定の音源の近接マイクロホン信号に特化した（その音源専用の）DNN モデルを全ての音源分用意する方が学習に必要なデータ量，学習の収束時間，および予測精度の観点から望ましいと思われる．さらに，現在対象としている近接マイクロホン信号に含まれている被り音は，他の音源の近接マイクロホンで大きな音量で観測されている．例えば，KD の近接マイクロホン信号には SD や HH が被り音として含まれるが，その成分は SD の近接マイクロホンや HH の近接マイクロホンで強い成分として観測している．これらの「被り音となる音源の近接マイクロホン信号」を Fig. 3.3 のように補助情報として DNN に入力することで，主入力の観測信号に含まれる被り音を効果的に抑圧する DNN モデルを構築できる可能性が高い．

以上の観点から，本論文では，マルチトラック録音で得られる各音源の近接マイクロホン信号を対象とした被り音抑圧 DNN モデルを提案する．この提案手法は Fig. 3.3 に示す通り，各音源に特化した DNN モデルであり，目的音源の近接マイクロホン信号を主入力，非目的音源（被り音音源）の近接マイクロホン信号を補助情報とする構造を持っている．

3.3 データセットの作成

本節では，ドラムセットの被り音抑圧 DNN モデルを学習するためのデータセットの作成方法について述べる．本論文で使用するデータセットは StemGMD [4] を基に作成する．StemGMD にはドラムセットの各音源のステムが含まれている．従来手法では，StemGMD に含まれる 9 チャネルのステムを単純に加算した信号を DNN の入力信号としていた．しかし，提案手法では，マルチトラック録音で得られる近接マイクロホン信号を DNN の入力として使用するため，各ステムを単純に加算するだけでは入力信号を作成することができない．例えば，KD の近接マイクロホン信号に含まれる KD 成分，SD の近接マイクロホン信号に含まれる KD 成分，および HH の近接マイクロホン信号に含まれる KD 成分は，近接条件の違いにより大きさが全く異なる．そのため，近接マイクロホン信号を作成するときは，目的音源成分と被り音成分の相対的なレベルを模擬する必要がある．この点は SD の近接マイクロホン信号および HH の近接マイクロホン信号の模擬についても同様である．

本論文では近接マイクロホン信号中の目的音源信号を基準とし，被り音となる音源の信



(a)　　(b)

Fig. 3.1 Observed signals assumed in (a) LarsNet (stereo mixture signal) and (b) proposed method (multichannel close-microphone signals).

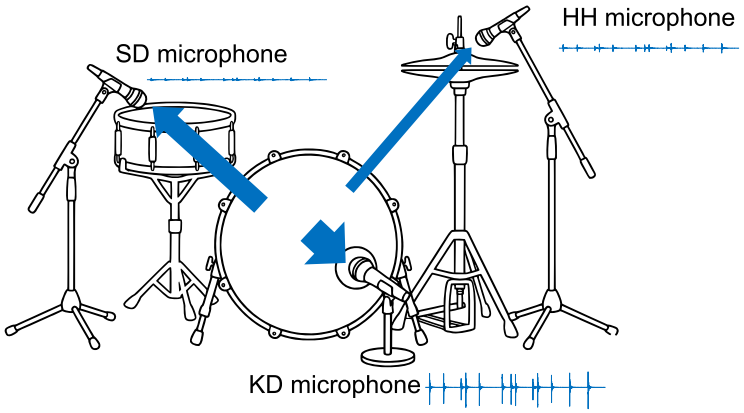
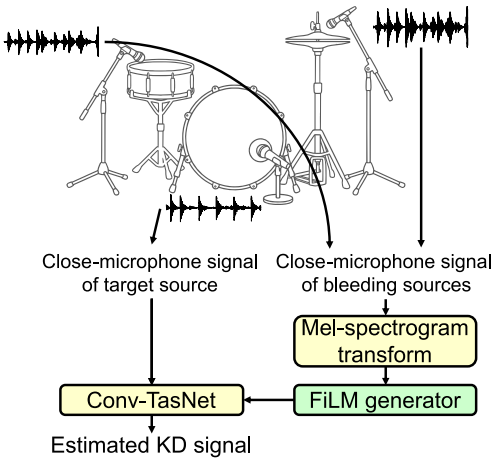
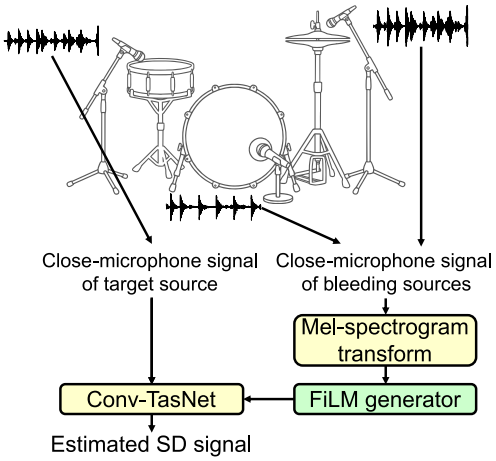


Fig. 3.2 KD signal captured by each microphone in multi-track drum set recording.

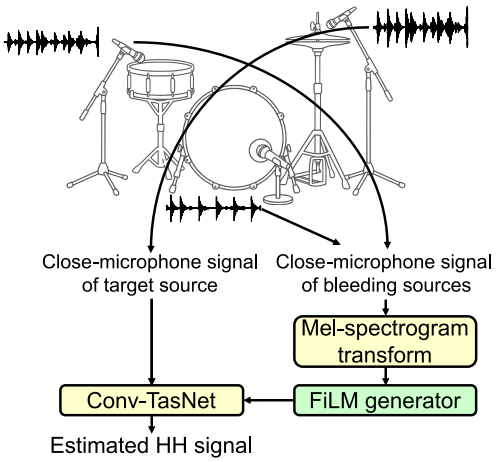
号には係数を乗じて加算することで，各近接マイクロホン信号を模擬する．例として，KD の近接マイクロホン信号を作成するときは，KD のステムはそのままとし，SD および HH のステムにそれぞれ係数を乗じて加算する．$\lambda = 1, 2, \cdots, \Lambda$ をマイクロホンのインデクス，$\psi = 1, 2, \cdots, \Psi$ を音源のインデクスとする．いま，λ 番目のマイクロホンは $\psi = \lambda$ 番目の音源に近接させていると定義し，$m_\lambda[l]$ および $n_\psi[l]$ をそれぞれ λ 番目の近接マイクロホン信号



(a)



(b)



(c)

Fig. 3.3 Target-dependent bleeding-sound reduction with three separate models trained for different close-microphone inputs: (a) KD, (b) SD, and (c) HH.

および ψ 番目の音源のステムとすると，$m_\lambda[l]$ を次式のように模擬する．

$$m_\lambda = \sum_\psi p_{\lambda,\psi} n_\psi[l] \tag{3.1}$$

$$p_{\lambda,\psi} = \begin{cases} 1 & (\lambda = \psi) \\ \rho_{\lambda,\psi} & (\lambda \neq \psi) \end{cases} \tag{3.2}$$

ここで，$p_{\lambda,\psi}$ は各ステムに乗じられる係数である．λ 番目のマイクロホンに対して $\psi \neq \lambda$ 番目の音源は被り音となるため，$\psi \neq \lambda$ 番目のステムに対しては ρ という値が乗じられて加算される．従って，ρ を 1 よりもある程度小さい乱数で定めれば，観測信号 $m_\lambda[l]$ は $\psi = \lambda$ 番目の音源を強く含み，$\psi \neq \lambda$ 番目の音源を小さい音量で含む信号となる．本論文においては KD, SD，および HH の 3 音源のみを対象とするため，これらを順番に $\psi = 1, 2, 3$ と定義すれば，Fig. 3.4 に示すように，KD の近接マイクロホン信号は

$$m_1[l] = \sum_\psi p_{1,\psi} n_\psi[l] \tag{3.3}$$

$$= p_{1,1} n_1[l] + p_{1,2} n_2[l] + p_{1,3} n_3[l] \tag{3.4}$$

$$= n_1[l] + \rho_{1,2} n_2[l] + \rho_{1,3} n_3[l] \tag{3.5}$$

となる．同様に SD および HH の近接マイクロホン信号の模擬は

$$m_2[l] = \rho_{2,1} n_1[l] + n_2[l] + \rho_{2,3} n_3[l] \tag{3.6}$$

$$m_3[l] = \rho_{3,1} n_1[l] + \rho_{3,2} n_2[l] + n_3[l] \tag{3.7}$$

と表せる．なお，$\rho_{\lambda,\psi}$ は乱数で毎回生成するため，定数ではない．ただし，ステム $n_\psi[l]$ は正規化された信号とする．なお，StemGMD には 10 種類の異なる音色のドラムセットで生成されたステムかつ複数の曲が存在する．しかし，式 (3.5) から (3.7) による近接マイクロホン信号の模擬には常に同一の音色のドラムセットかつ同一の曲に属するステム同士で行い，異なる曲のステムは混合しないものとした．

3.4　DNN の入出力

本節では，提案手法の DNN モデルの入力および出力について述べる．3.2 節で述べた通り，提案手法は実際のマルチトラック録音で得られる近接マイクロホン信号の入力を想定している．すなわち，特定の音源の近接マイクロホン信号を主入力とし，残りの音源の近接マイクロホン信号を補助情報として入力する設計とする．このとき，Fig. 3.3 に示す通り，主入力は波形のまま Conv-TasNet へと接続され，補助情報はメルスペクトログラムに変換されて FiLM ジェネレータに接続される．例えば，KD の近接マイクロホン信号の被り音抑圧の場合は，KD の近接マイクロホン信号が波形のまま Conv-TasNet に入力され，SD および HH の近接マイクロホン信号がそれぞれメルスペクトログラムに変換されて FiLM ジェネレータに入力される．KD の近接マイクロホン信号に混入している被り音は SD と HH であるため，FiLM

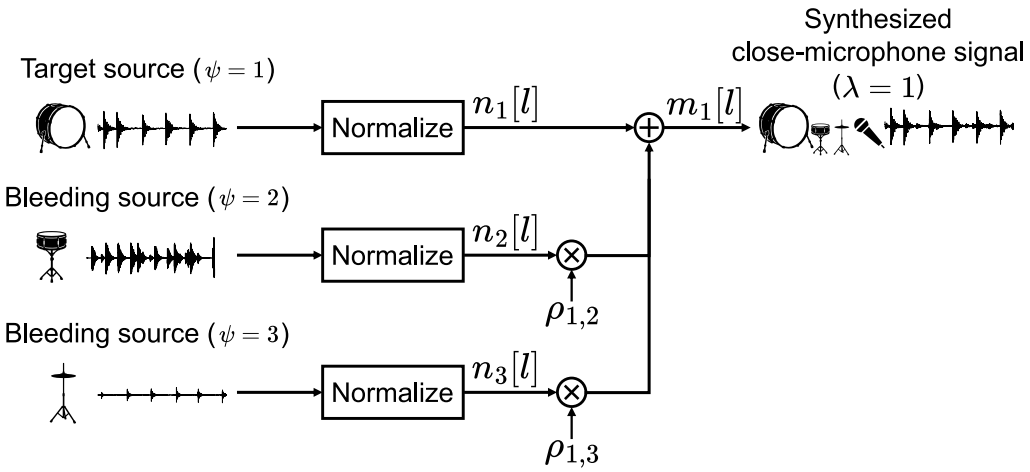


Fig. 3.4 Synthesis of a target close-microphone signal for KD ($\lambda = 1$).

に入力された補助情報がこれらの被り音抑圧にとって有効な情報として活用されることを意図している．

提案手法の出力は主入力として与えられた観測信号中の近接音源信号である．すなわち，主入力に含まれる被り音のみが抑圧され，近接している音源の成分のみとなって出力されることを想定している．例えば，主入力が KD の近接マイクロホン信号である場合，そこに含まれる被り音は抑圧され，KD の成分のみが得られるように DNN モデルを学習する．

なお，3.2 節で述べた通り，提案手法の DNN モデルは音源ごとに学習・構築される．具体的には，KD の近接マイクロホン信号の被り音を抑圧するモデル，SD の近接マイクロホン信号の被り音を抑圧するモデル，および HH の近接マイクロホン信号の被り音を抑圧するモデルをそれぞれ独立に構築する．このようにモデルを独立させることで，各音源に対する学習を個別に行うことが可能となり，より高精度な被り音抑圧が期待できる．

3.5 DNN の構造

本節では, 提案手法の DNN モデルについて説明する．Fig. 3.5 に提案手法の DNN モデルを示す．本モデルは, エンコーダ, セパレータ (TCN ブロック), およびデコーダからなる Conv-TasNet を基本アーキテクチャとする．また，補助情報から $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \cdots, \boldsymbol{\gamma}_t, \cdots, \boldsymbol{\gamma}_T] \in \mathbb{R}^{C \times T}$ および $\boldsymbol{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_t, \cdots, \boldsymbol{\beta}_T] \in \mathbb{R}^{C \times T}$ を生成する FiLM ジェネレータを併用し，各 TCN ブロックの中間特徴量を変調する．ここで，$t = 1, 2, \cdots, T$ は TCN ブロック数，C はセパレータの中間層のチャネル数であり，$\boldsymbol{\gamma}_t \in \mathbb{R}^C$ および $\boldsymbol{\beta}_t \in \mathbb{R}^C$ は t 番目の TCN ブロックの係数ベクトルを示している．いま，Fig. 3.5 に示す通り，Stack layer は 3 個，1 つの Stack layer 内の TCN ブロックは 8 個なので $T = 24$ であり，セパレータの中間層のチャネル数は $C = 32$ である．$\boldsymbol{\gamma}_t$ および $\boldsymbol{\beta}_t$ は，TCN ブロック内の複数の畳み込み層および活性化関数を通した後のデータに対して適用される．また，Fig. 2.5 に示す基本的な Conv-TasNet との違いとして，Fig. 3.5 に示す提案手法の DNN モデルは入力信号のチャネルを拡張する

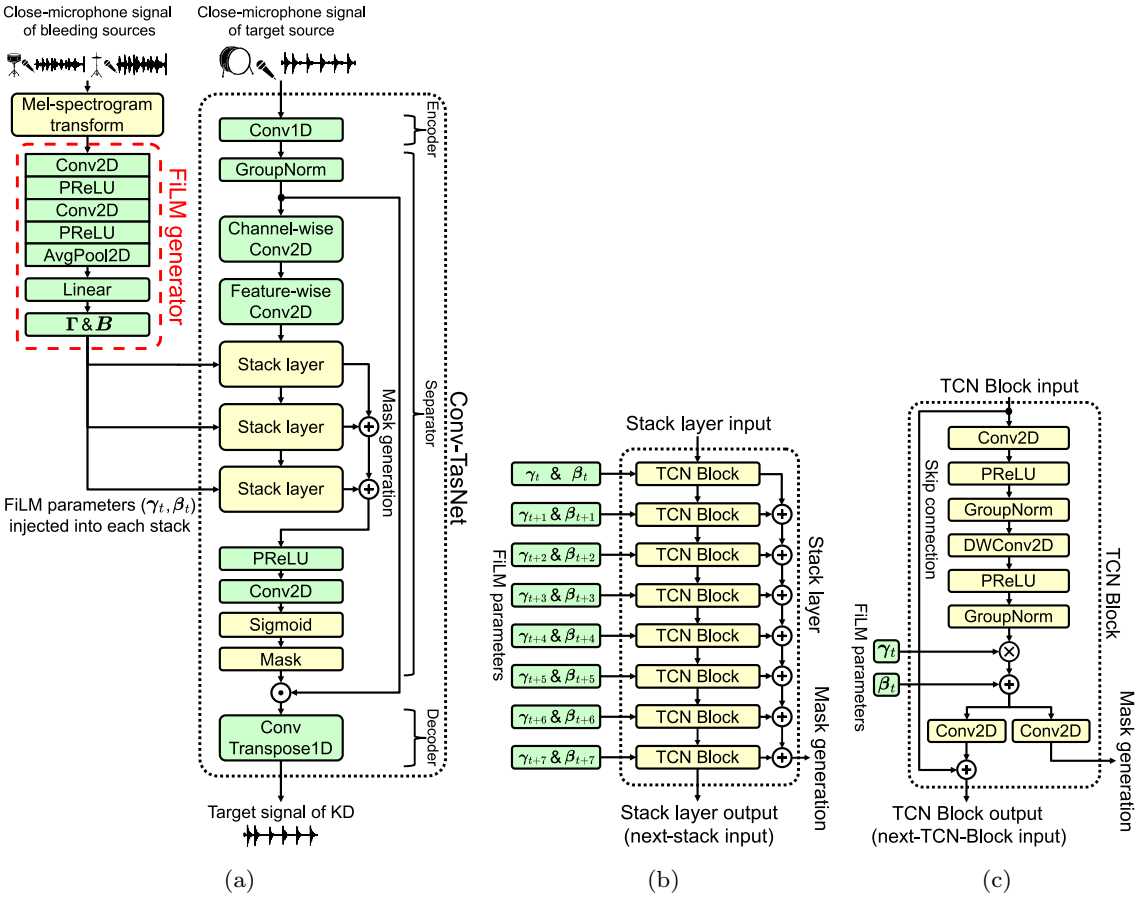


Fig. 3.5 Architecture of proposed model, (a) Conv-TasNet with FiLM generator, (b) expanded view of Stack layer in (a) with FiLM parameters, and (c) expanded view of a TCN Block in (b) with FiLM parameters.

Channel-wise Conv2D 層を追加した．さらに，セパレータの畳み込み層を Conv1D 層から Conv2D 層に変更し，時間方向とエンコーダで抽出される特徴方向（Conv1D 層の出力チャネルによる拡張次元）の畳み込みを行うようにした．これらは，時間と特徴の次元の他にチャネルという更なる次元を追加することで被り音抑圧に有効な特徴をより多く得ることを目的としている．

3.6　DNN の損失関数と学習

提案手法の DNN で使用する損失関数について説明する．提案手法では尺度不変信号対雑音比（scale-invariant signal-to-noise ratio: SI-SNR）[9] を用い，負の SI-SNR が小さくなるように DNN を学習する．

一般的に，信号対雑音比（signal-to-noise ratio: SNR）は，信号 $\sigma[l]$ のエネルギーが雑音

$\eta[l]$ のエネルギーに対してどれだけ強いかを比率で数値化したものであり，次式で計算される．

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_l |\sigma[l]|^2}{\sum_l |\eta[l]|^2} \quad [\mathrm{dB}] \tag{3.8}$$

この SNR を音源分離に適した形に変更し，音源分離における推定信号と正解信号のスケール差の影響を排除した指標が SI-SNR である．SI-SNR では，推定信号 $\hat{s}[l]$ を正解信号 $s[l]$ に直交射影し，スケール差を吸収する．直交射影によるスカラー α は以下の式で表される．

$$\alpha = \frac{\sum_l \hat{s}[l]s[l]}{\sum_l |s[l]|^2} \tag{3.9}$$

この α を用いて，SI-SNR は次式で計算される．

$$\text{SI-SNR} = 10\log_{10}\frac{\sum_l |\alpha s[l]|^2}{\sum_l |\hat{s}[l] - \alpha s[l]|^2} \quad [\mathrm{dB}] \tag{3.10}$$

式 (3.10) の分子は推定信号に含まれる正解信号成分のエネルギー，分母は推定信号に含まれる雑音成分のエネルギーである．

SI-SNR は推定信号のスケールに対して不変となる評価値であるため，学習時には DNN の出力信号の波形の振幅が不定となる．その出力信号を推定信号としてそのまま定義してしまうと，被り音抑圧における正解信号に対して推定信号の振幅が極端に大きいまたは小さい場合がある．そこで提案手法では，DNN の入力信号 $d[l]$ を DNN の出力信号 $\hat{d}[l]$ に射影し，得られた係数で出力信号をスケーリングして波形の振幅を入力信号に揃え，その信号を推定信号 $\hat{s}[l]$ と定義する．直交射影によるスカラー ν およびスケーリングされた出力信号（推定信号 $\hat{s}[l]$）は次式で計算される．

$$\nu = \frac{\sum_l d[l]\hat{d}[l]}{\sum_l |\hat{d}[l]|^2} \tag{3.11}$$

$$\hat{s}[l] = \nu\hat{d}[l] \tag{3.12}$$

以上より，提案手法の DNN モデルの損失関数は負の SI-SNR と定義し，これを小さくする DNN の出力信号が得られるように，DNN モデル内の全てのパラメータを更新する．なお，提案手法では KD，SD，および HH それぞれに対して専用の DNN モデルを学習するため，各モデルは単一音源の推定に対応した損失関数を直接最小化することができ，従来手法である LarsNet と比較して学習目標が明確となっている．

3.7 本章のまとめ

本章では，目的音源の近接マイクロホン信号を主入力とし，被り音源の近接マイクロホン信号を補助情報とする被り音抑圧手法を提案した．DNN モデルは Conv-TasNet を基盤とし，補助情報を FiLM を介してセパレータ内部の中間特徴へ条件付けを行う構造とした．また，StemGMD を基に混合係数を制御して近接マイクロホン信号を疑似的に生成し，提案手法を学習・評価可能なデータセット作成手順を示した．次章では，従来手法と提案手法で実験を行い，被り音抑圧性能に比較・検証をする．

第 4 章

被り音抑圧実験

4.1 はじめに

本章では，前章で構築したデータセットおよび DNN を用いて被り音抑圧実験を行い，提案手法の有効性を検証する．4.2 節では，提案手法を用いた被り音抑圧実験の実験条件を説明する．また，比較対象として，公開モデル LarsNet，再学習モデル LarsNet，FiLM なし Conv-TasNet を用意する．比較は 4.3 節で行い，音源対歪み比（source-to-distortion ratio: SDR）による定量評価を用いて DNN アーキテクチャおよび補助情報による条件付けがどの程度結果に寄与するのかを明らかにする．最後に，4.4 節で本章をまとめる．

4.2 実験条件

本節では，比較実験で定めるデータセットや各手法の実験条件について詳細を述べる．まず，データセットについて述べる．StemGMD のデータに対して Table 4.1 に示す変更を加えたものを本実験における学習データとして用いる．StemGMD に含まれる信号はステレオであるため，全てのステムに対して左右チャネルを平均化してモノラル信号に変換した．サンプリング周波数は 16 kHz になるようにリサンプリングを全てのステムに適用した．元の 44.1 kHz では学習に多くの時間がかかり，計算機資源の要求も大きくなるため，16 kHz に統一した．なお，リサンプリングを行わない場合の実験についても検討しており付録 A に示す．また，各ステムは最大振幅が 0.3 となるように正規化を施した．StemGMD に含まれるステムの信号長は統一されていない．そこで，作成するデータセットの信号長は 3 秒と定義し，StemGMD の各ステムを 3 秒ごとに切り出した．3 秒に満たない区間はデータセットに含めないこととした．3.3 節で述べた，データセット作成時の混合係数 $\rho_{\lambda\psi}$（$\lambda \neq \psi$）は，Table 4.2 に示す値を中心値としその値の 0.9 倍から 1.1 倍の範囲の一様分布に従う乱数を用いた．この乱数値を用いて，式 (3.5) から式 (3.7) で混合することで，被り音を含む近接マイクロホン信号を模擬した入力データおよびその正解データを作成した．なお，Table 4.2 の値は，実際にマルチトラック録音されたドラムセットの近接マイクロホン信号から求めた実測に基づく振幅

Table 4.1 Dataset parameters for StemGMD and created dataset

Parameter	StemGMD	Created dataset
Signal length [s]	Various length	3
Sampling frequency [kHz]	44.1	16
Channels	2 ch. (stereo)	1 ch. (monaural)

Table 4.2 Mixing gains for simulating close-microphone signal with bleeding sounds, where the target source is (a) KD, (b) SD, or (c) HH

(a)

Source	Mixing gain
KD	1.00000
SD	0.15181
HH	0.00378

(b)

Source	Mixing gain
KD	0.11886
SD	1.00000
HH	0.29651

(c)

Source	Mixing gain
KD	0.46051
SD	1.67126
HH	1.00000

比である．

次に，本実験の学習条件について述べる．本実験の学習条件を Table 4.3 にまとめる．作成したデータセットの総時間はおよそ 125 時間であったが，計算資源の制約があるため，学習データとして 7.5 時間分（27,000 秒）のデータを使用した．学習中の評価に使用するデータ（検証データ）およびモデルの評価に使用するデータ（テストデータ）はそれぞれ 3000 秒とした．信号長を 3 秒に固定しているため，これらの学習，検証，およびテストに用いる観測信号の総数はそれぞれ 9,000，1,000，および 1,000 である．DNN の最適化には Adam [18] を用いた．学習における最大エポック数は 300 回，バッチサイズは 8 に設定した．また，20 エポックの間損失関数値が改善しない場合，早期終了をするように設定した．Adam の学習率はウォームアップ付きコサインアニーリング [19] という手法を用いて変動させた．この手法は学習初期に学習率を徐々に上げ，その後コサイン曲線で滑らかに減少させる方法であり，初期の学習を安定させる効果がある．学習率の初期値は 5×10^{-6} とし，10 エポック目で最大学習率 5×10^{-5} となるように線形的に上昇させ，その後コサイン曲線で最小学習率 2.5×10^{-6} まで減少させた．ただし，早期終了をした場合，学習率は最小学習率まで下がることはない．

最後に，比較手法について述べる．提案手法を含め，次に示す 4 つの手法で比較を行う．1 つ目の手法は，学習済み公開モデルを使用した LarsNet である．LarsNet の DNN モデルのソースコードと学習済みモデルは公開されている．これを使用し，音源分離を行った結果を評価する．ただし，DNN への入力は本論文で作成したデータセットを用いるため，信号長，サンプリング周波数，チャネル数，および音量バランスが学習時のデータと異なる．LarsNet で要求される入力は，信号長が 11.85 秒，サンプリング周波数が 44.1 kHz，チャネル数が 2（ステレオ）である．そのため，信号長は足りない部分をゼロ詰めし，サンプリング周波数は観測信号をアップサンプリングしてそろえた．また，チャネルはシングルチャネルの信号を複製

Table 4.3 Dataset split and training hyperparameters

Parameters	**Value**
Total training audio duration [s]	27,000
Total validation audio duration [s]	3,000
Total test audio duration [s]	3,000
Maximum number of epochs	300
Batch size	8
Early-stopping patience	20
Learning-rate schedule	Cosine annealing with warm-up
Warm-up start learning rate	5×10^{-6}
Peak learning rate	5×10^{-5}
Minimum learning rate	2.5×10^{-6}
Warm-up epochs	10

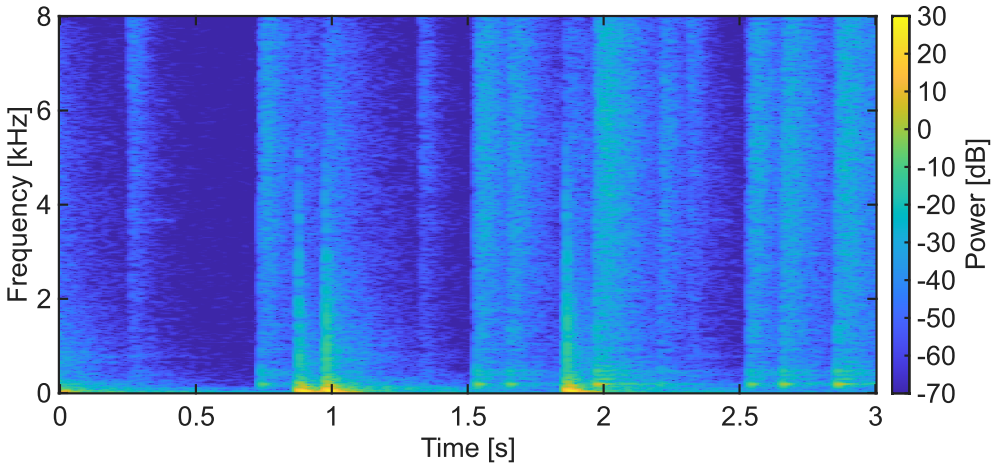
して 2 チャネルの信号として入力した．よって，この手法では学習データとテストデータの形式が大きく異なるため，この手法の結果は参考値として扱う．また，LarsNet は KD，SD，HH，TT，および CY の 5 つのステムを並列に同時推定するモデルである．例えば，KD の近接マイクロホン信号をこの LarsNet に入力した場合も，5 つのステムの推定信号を出力するため，そのうちの KD の推定信号を被り音抑圧の結果とみなして評価に用いている．そのほかの SD の近接マイクロホン信号や HH の近接マイクロホン信号の入力に対しても同様に，目的音源の推定信号を評価に用いている．LarsNet で用いている STFT は，窓長が 4096 点（約 92.9 ms），シフト長は 1024 点（約 23.2 ms），窓関数はハン窓である．2 つ目の手法は，本論文で作成したデータセットを用いて学習した LarsNet である．1 つ目の手法は，学習データとテストデータの形式が一致していないため LarsNet にとって不利な条件といえる．そこで，本論文で作成したデータセットを用いて LarsNet を初期状態から学習した．LarsNet の入力層は本論文で作成したデータセットに合わせ，信号長を 3 秒，サンプリング周波数を 16 kHz，チャネルをシングルチャネルに変更した．サンプリング周波数の変更に伴い，LarsNet 内部の受容野が大きく変わらないよう，STFT の窓長を 4096 点（約 92.9 ms）から 2048 点（128 ms），シフト長を 1024 点（約 23.2 ms）から 512 点（32 ms）に変更した．また，推定するステムを KD，SD，および HH の 3 つにするように DNN モデルを変更した．損失関数は，2.7.2 項で述べた L_1 ノルムを使用し，U-Net ごとの損失の平均を求めてモデル全体の損失の最小化を行った．2 つ目の手法の推定信号の取り扱いは 1 つ目の手法と同様の方法とした．3 つ目の手法は，提案手法で用いるモデルの FiLM を除去したモデルである．従って，入力は特定の音源の近接マイクロホンの観測信号（時間波形）のみであり，その他の音源の近接マイクロホン信号を補助情報として入力する機構はない．4 つ目の手法は，3 章で述べた提案手法である．補助情報に対して行う STFT の窓長は 2048 点（128 ms），シフト長は 1024 点（64 ms）であ

り，メルスペクトログラムに変換するときのメルフィルタバンクの次元数は $P = 64$ である．以上の 4 つの手法で実験および評価を行い，比較する．

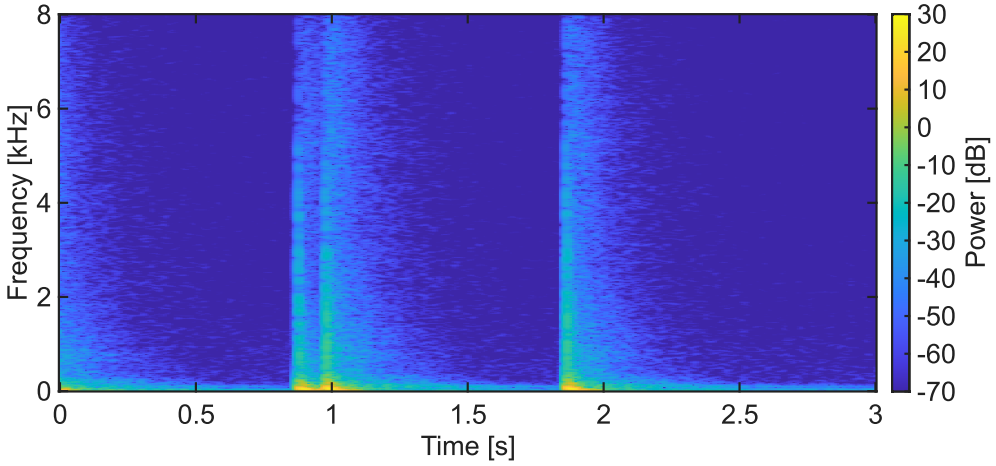
4.3 実験結果と比較

まず，提案手法の実験結果について述べる．Fig. 4.1 (a) から (c) にそれぞれ KD の入力信号（KD の近接マイクロホンの観測信号），正解信号，および推定信号のスペクトログラムを示す．このスペクトログラムは，被り音を抑圧が確認でき，かつ推定精度がテストデータ全体の平均に近い信号を例として示したものである．また，窓関数はハン窓，STFT の窓長を 1024 点（64 ms），シフト長を 128 点（8 ms）とし，スペクトログラムの振幅を dB に変換して表示し，カラーバーの範囲は $[-70, 30]$ dB に固定した．Fig. 4.1 (a) のスペクトログラムにあって Fig. 4.1 (b) のスペクトログラムにない成分が，被り音に対応する成分である．Fig. 4.1 (c) のスペクトログラムでは，これらの成分が抑圧されていることが分かる．しかし，Fig. 4.1 (c) のスペクトログラムを Fig. 4.1 (b) と比較すると，Fig. 4.1 (b) には存在しない縦縞状のアーティファクトや残響成分の劣化が観測される．被り音抑圧は一定程度達成されている一方で，正解信号の完全な再現には至っていないことが示唆される．同様に Fig. 4.2 および Fig. 4.3 に SD および HH の同様の図を掲載する．SD および HH にも KD と同様に被り音成分の抑圧および縦縞状のアーティファクトや残響成分の劣化を確認することができる．また，Fig. 4.4 に Fig. 4.3 とは異なる HH の入力信号（HH の近接マイクロホンの観測信号），正解信号，および推定信号のスペクトログラムを示す．Fig. 4.4 (a) および Fig. 4.4 (b) を見ると，Fig. 4.3 (a) の信号と比較してこの HH の入力信号は被り音が多く含まれ，HH と被り音源（KD および SD）の発音に時間的な被りが生じていることが分かる．この信号では Fig. 4.4 (c) に示されるように被り音成分の完全な抑圧ができていないことが分かる．このような被り音を多く含み，目的音源と被り音源の発音に時間的な被りが生じる信号は，HH で多数確認され，その信号では同様に推定信号において被り音を抑圧することができていない信号が多かった．

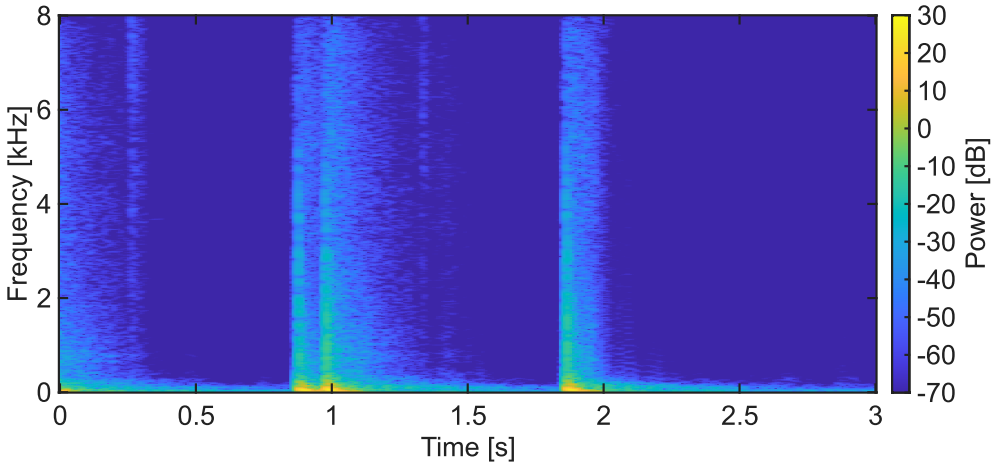
次は，定量的に従来手法と提案手法の比較を行う．評価指標は，被り音の抑圧度合いと目的音源の歪みの少なさの両方を考慮した全体的な分離精度を表す音源対歪み比（source-to-distortion ratio: SDR）[20] を用いた．また，基準として観測信号の SDR も算出し，4 手法の結果と併せて比較・評価する．Table 4.4 に，テストデータに対する平均 SDR を示す．Table 4.4 より，いずれの音源においても入力信号に比べて各手法の SDR が大きく，学習モデルにより被り音抑圧が達成されていることが確認できる．公開済み学習モデルの LarsNet は，入力データが学習時のと異なるため参考値として扱うが，それでも KD，SD，および HH でそれぞれ 30.57 dB，16.71 dB，および -5.29 dB と入力信号から改善している．本論文で作成したデータセットに合わせて学習した LarsNet は，KD, SD, および HH で 34.09 dB, 21.66 dB, および -1.23 dB となり，公開モデルより高い SDR が得られた．さらに，Conv-TasNet は，全音源で LarsNet を上回り，KD，SD，および HH で 36.14 dB，23.06 dB，および 2.22 dB を示した．これより，本論文の設定においては，Conv-TasNet 系の構成が有効であるといえ



(a)

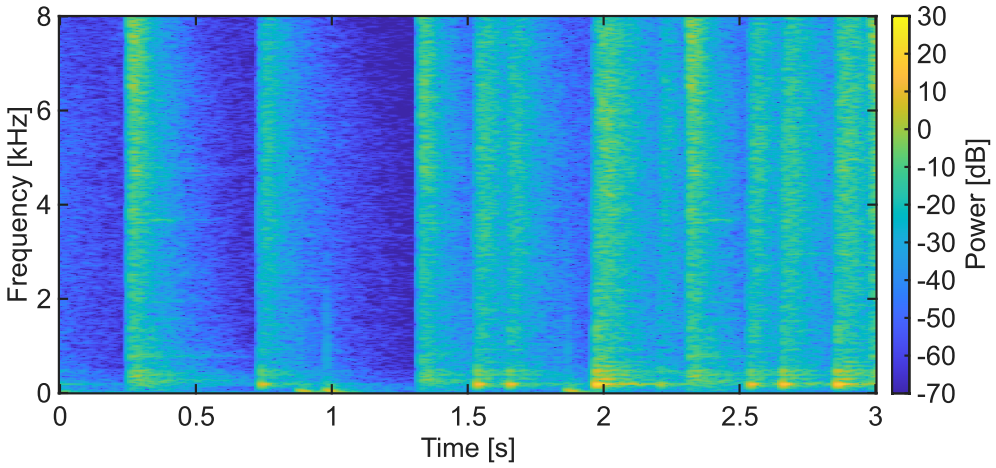


(b)

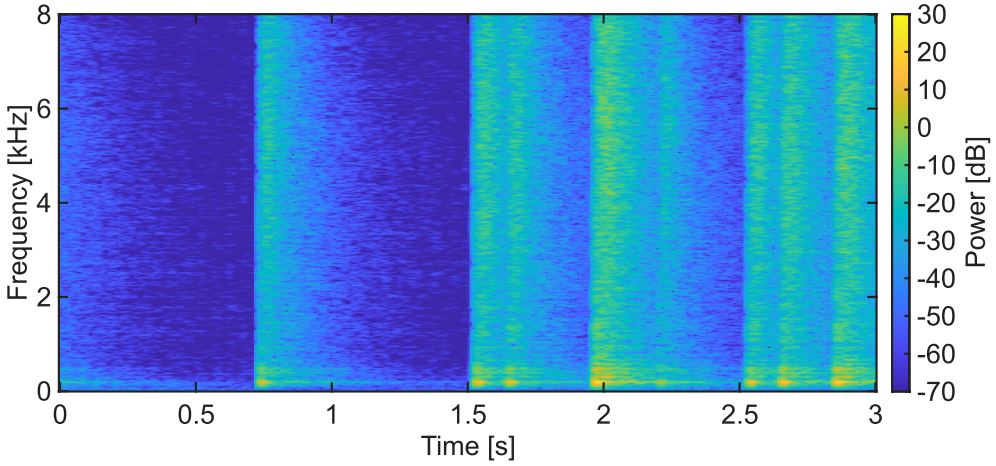


(c)

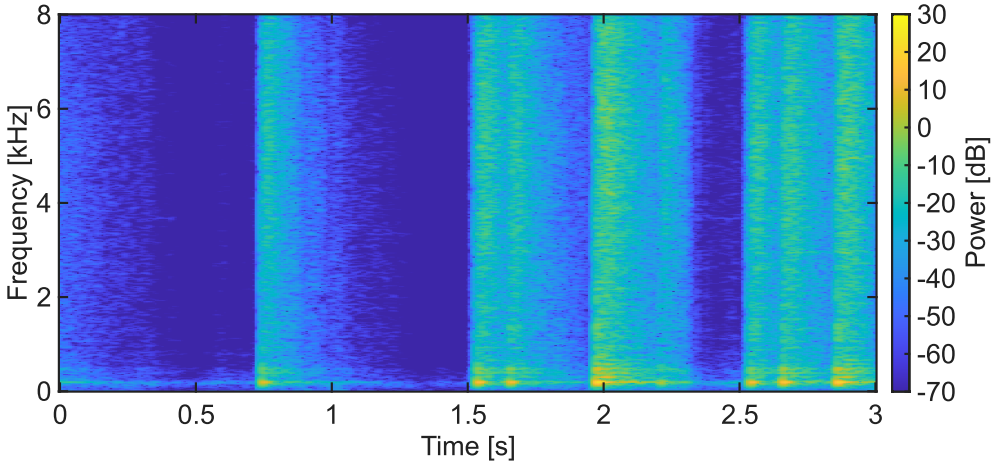
Fig. 4.1 Spectrograms of (a) KD close-microphone signal (with bleeding sound), (b) oracle KD stem, and (c) estimated KD stem.



(a)

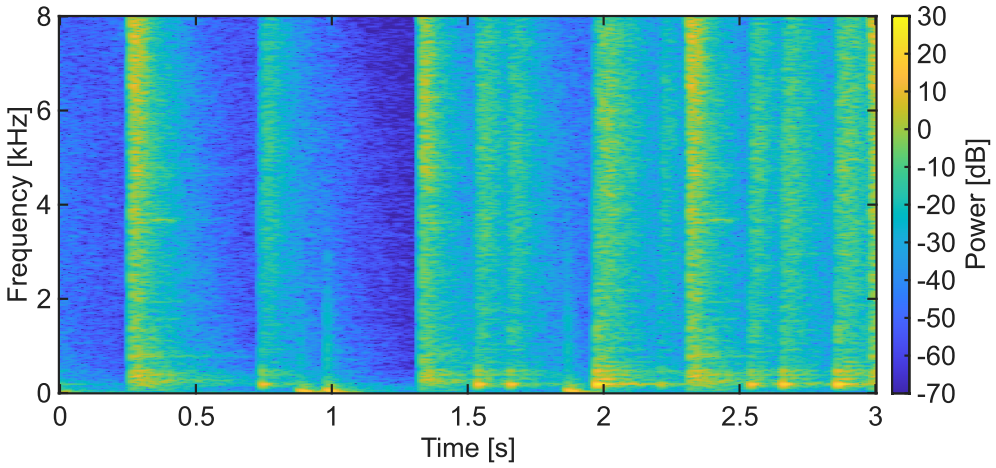


(b)

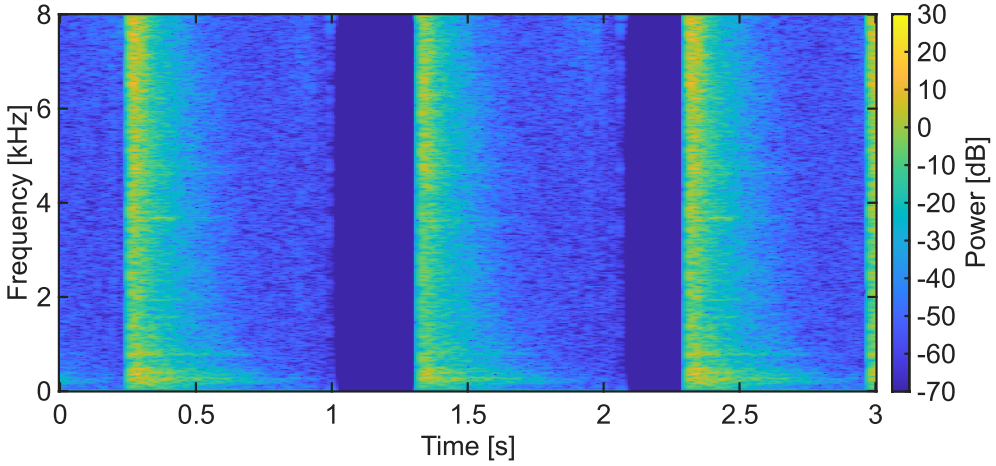


(c)

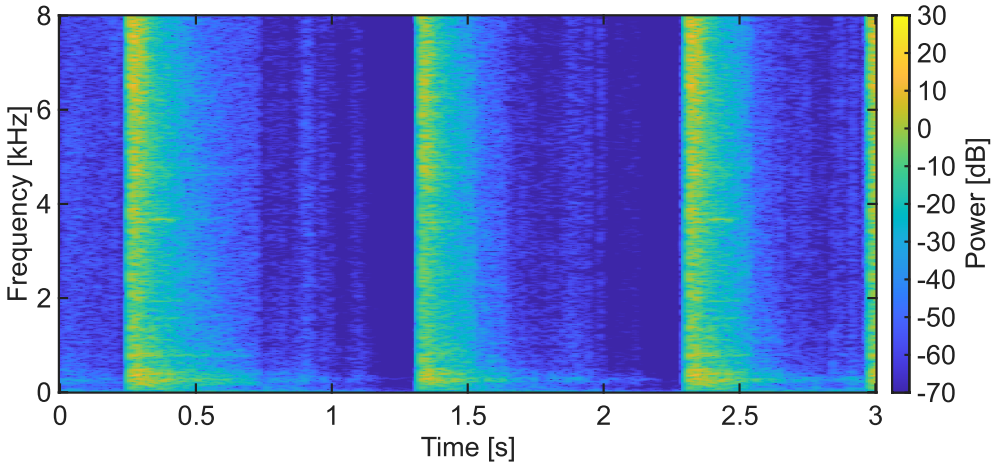
Fig. 4.2 Spectrograms of (a) SD close-microphone signal (with bleeding sound), (b) oracle SD stem, and (c) estimated SD stem.



(a)

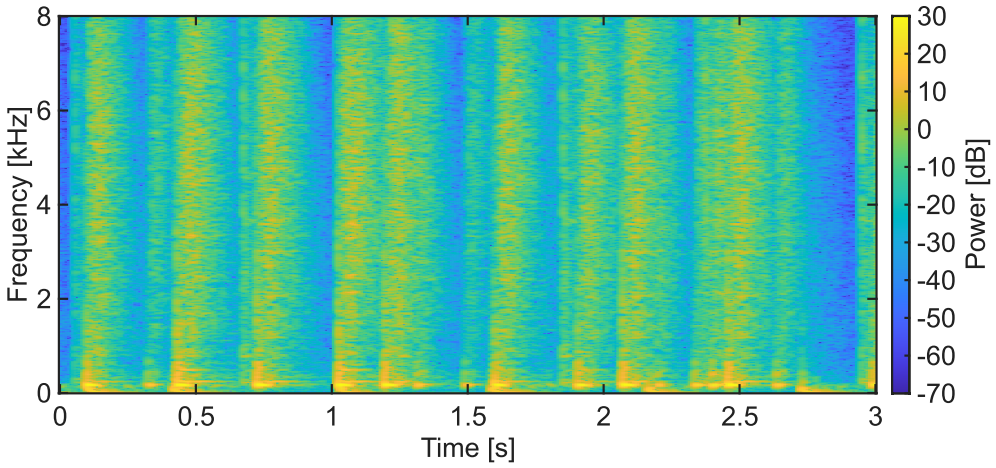


(b)

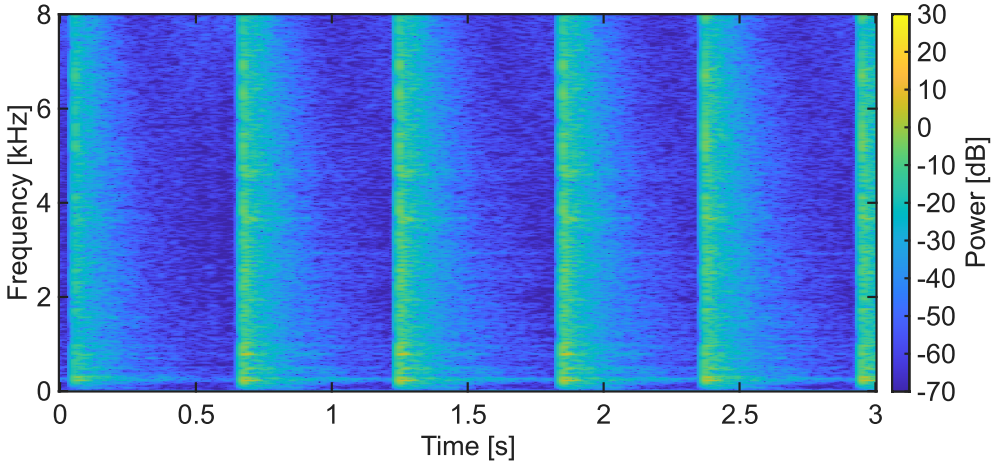


(c)

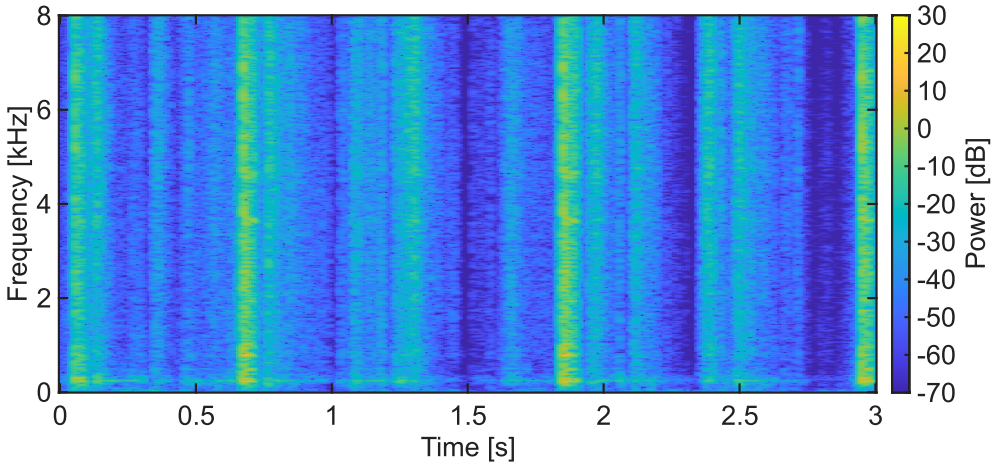
Fig. 4.3 Spectrograms of (a) HH close-microphone signal (with bleeding sound), (b) oracle HH stem, and (c) estimated HH stem.



(a)



(b)



(c)

Fig. 4.4 Spectrograms of different signals, (a) HH close-microphone signal (with bleeding sound), (b) oracle HH stem, and (c) estimated HH stem.

Table 4.4 Average SDR [dB] on test data

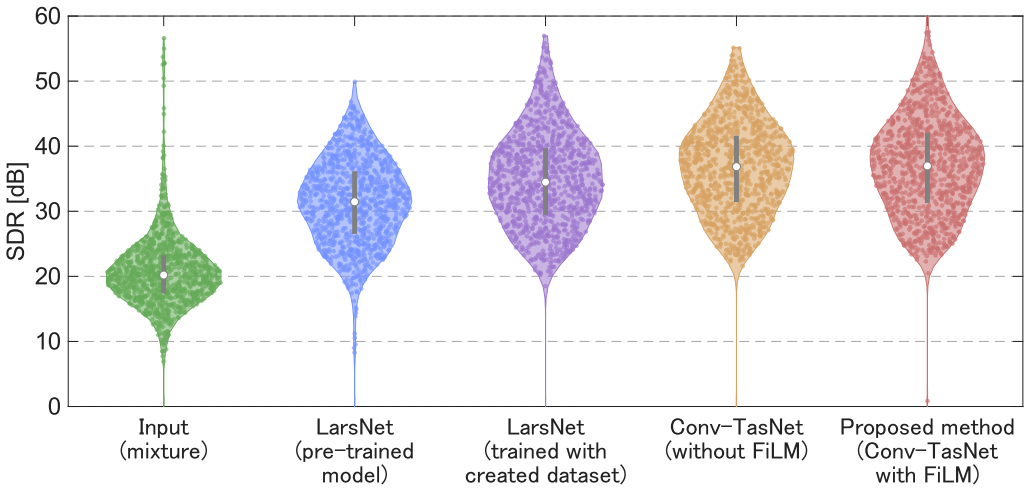
Target source	Input (mixture)	LarsNet (pre-trained model)	LarsNet (trained with created dataset)	Conv-TasNet (without FiLM)	Proposed method (Conv-TasNet with FiLM)
KD	19.75	30.57	34.09	36.14	**36.37**
SD	10.63	16.71	21.66	**23.06**	22.87
HH	−22.32	−5.29	−1.23	2.22	**2.25**

る．提案手法では，KD，SD，および HH で 36.37 dB，22.87 dB，および 2.25 dB となり，KD と HH において Conv-TasNet をわずかに上回った．これより，被り音の近接マイクロホンの信号を補助情報として FiLM に入力し，Conv-TasNet の条件付けとして活用することは，被り音抑圧の改善に寄与する可能性が高い．しかしながら，得られた改善が微量であり，劇的な改善を得るにはより効果的なネットワーク構造を模索する必要があることが分かる．得られた改善が微量に留まった要因として以下が考えられる．本論文が対象とするマルチトラック録音で得られる近接マイクロホン信号は目的音源成分の割合が大きく，目的音源と被り音源の SN 比が比較的高い．このような条件では，Conv-TasNet 単体でも被り音抑圧を一定程度達成できるため，FiLM による条件付けの寄与が相対的に小さかった可能性がある．

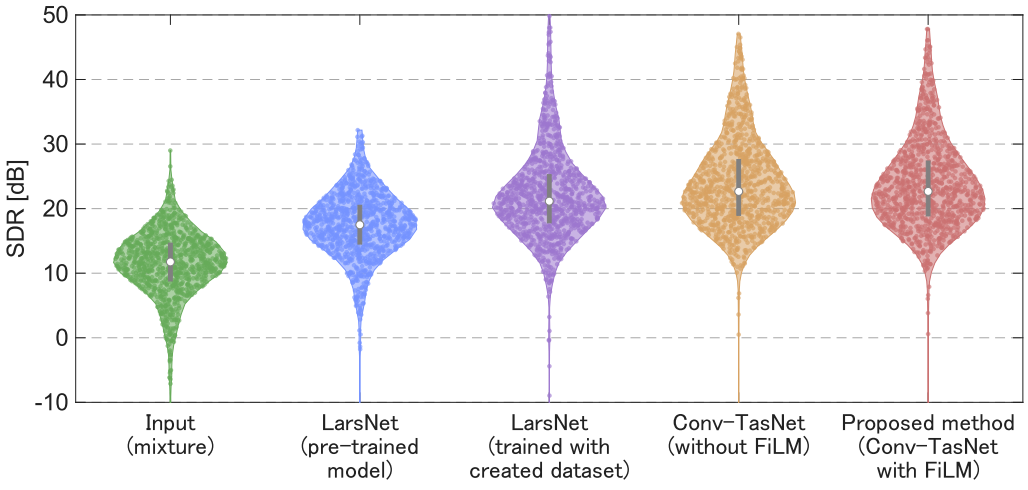
Fig. 4.5 には，各手法のテストデータに対する SDR 分布をバイオリンプロットとして示す．白点は中央値，太い縦棒は四分位範囲（25–75%）である．Fig. 4.5 からも，入力信号に比べて各学習モデルの中央値が高く，特に Conv-TasNet および提案手法では高い中央値を示すことが確認できる．また，音源ごとに分布の広がりや裾の長さが異なる．例えば，KD および SD では各手法で分布全体が高 SDR 側にシフトする一方，HH では分布のばらつきが大きく，推定後も低 SDR に留まる結果があることが確認できる．このため，平均値だけでなく分布形状からも，被り音抑圧精度の差やそのばらつきを確認できる．

4.4　本章のまとめ

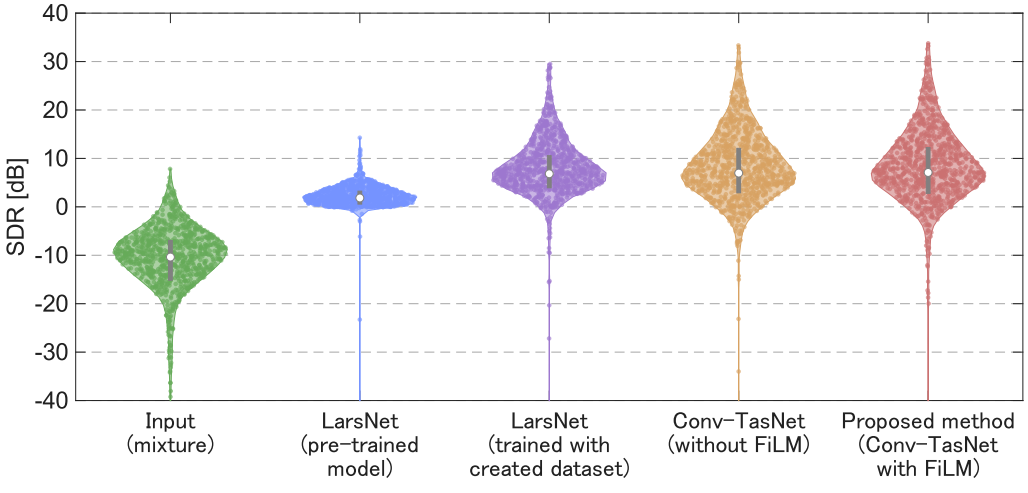
本章では，被り音抑圧実験により提案手法の有効性を検証し，KD，SD，および HH のいずれにおいても入力信号に比べて SDR が改善することを確認した．また，LarsNet と比較して Conv-TasNet の DNN モデルが有効である傾向が示され，時間領域モデルの適性が確認された．一方で，FiLM による条件付けの有無の差は小さく，補助情報の効果は限定的であった．スペクトログラムによる評価では，被り音成分の抑圧は確認できた一方で，縦縞状の成分や残響の粗さなどの歪みが残ることが確認でき，音質の改善が課題として残った．次章では，これらの結果を総括し，補助情報をさらに活用するための条件付け設計や推定信号の音質改善を中心に，今後の課題および改善方針を述べる．



(a)



(b)



(c)

Fig. 4.5 Violin plots of SDR for (a) KD, (b) SD, and (c) HH among five methods.

第 5 章

結言

本論文では，ドラムセット収録における被り音の問題に着目し，マルチトラック録音で得られるマルチチャネル信号を活用した被り音抑圧手法を提案した．提案手法では，目的音源の近接マイクロホン信号を主入力とし，被り音となる音源の近接マイクロホン信号を補助情報として FiLM により条件付けすることで，目的音源の推定精度向上を狙った．また，時間領域 end-to-end モデルである Conv-TasNet を基本アーキテクチャとして採用し，KD，SD，および HH の 3 音源を対象に被り音抑圧を行った．

1 章では，ドラムセットのマルチトラック録音で生じる被り音がミキシング品質を低下させる問題を整理し，本研究の目的を示した．既存手法はシングルチャネル信号を入力としており，実運用の収録条件と整合しない点を研究ギャップとして位置づけ，KD，SD，および HH の 3 音源を対象とした被り音抑圧に焦点を当てた．2 章では，STFT，DNN，Conv-TasNet，および FiLM の基礎を整理し，提案手法の理解に必要な枠組みを示した．従来手法として LarsNet を取り上げ，データセットの特徴と，U-Net による分離の入出力および損失をまとめ，本研究との比較軸を明確化した．3 章では，目的音源の近接マイクロホン信号を主入力とし，被り音の近接マイクロホン信号を FiLM による条件付けとして入力する被り音抑圧手法を提案した．4 章では，提案手法の有効性を検証するため，公開モデル LarsNet，再学習モデル LarsNet，FiLM なし Conv-TasNet，提案手法を同一条件で評価し，被り音抑圧性能を比較した．スペクトログラムの比較により，被り音の成分が推定信号で抑圧されていることを確認した．一方で，推定に伴う歪みが残る場合があった．SDR を用いた比較では，いずれの手法も入力信号の SDR を上回る結果となった．また，Conv-TasNet を用いた構成が有効であることが分かった．FiLM については，SDR の改善が小さく，優位性が限定的であった．

最後に今後の課題と改善方針について述べる．本研究では被り音の抑圧は確認できた一方で，推定に伴う歪みが残る場合がある．そのため，音質の更なる改善が必要であり，芸術性を損なわないレベルには達していない．特に HH は他の音源に比べて SDR が低い傾向がみられ，HH に対する抑圧性能と音質の両立が今後の課題である．また，FiLM による条件付けの有無による SDR の差は小さく，補助情報の利用効果は限定的であった．この要因として，近接マイクロホン信号を主入力とする DNN では，単一入力のみでも被り音抑圧が一定程度達成

でき，補助情報が十分に活用されなかった可能性が考えられる．今後は，FiLM の効果をより引き出すために，FiLM および FiLM ジェネレータへの入力特徴量の再検討，追加損失の導入による歪み低減，実録マルチトラックデータによる検証，を進める必要がある．さらに，推定結果を補助情報として再利用するような学習（一次推定結果を用いた再推定）も有効である可能性があり，補助情報の与え方を含めて検討する必要がある．以上より，マルチトラック録音で得られる情報を活用した被り音抑圧の有効性を示すとともに，条件付け手法の設計最適化と音質向上に向けた課題を明確化した．

謝辞

本論文は，香川高等専門学校電気情報工学科北村研究室にて行われた研究に基づくものです．

まず，本研究を進めるにあたり，ご多忙のところ熱心にご指導くださいました指導教員の北村大地准教授に心より感謝申し上げます．北村大地准教授には，論文執筆や研究に関する議論など，細部にわたるまで丁寧にご指導いただきました．図のセンスは絶望的で書く文章は言葉足らずな私の指導をこれからもよろしくお願いします．

本論の副査である柿元健准教授には，論文の構成や記述に関して大変有益な助言を頂き，大変お世話になりました．ここに厚く御礼申し上げます．

北村研究室の先輩である専攻科生の加藤大輝氏，鈴木慶氏，和気佑弥氏，小川遼氏，谷野宮蒼士氏には，に関するアドバイス等をはじめ，数々のご支援をいただきました．特に，加藤大輝氏は本研究および論文執筆のメンターとして多くの学びとアドバイスをいただきました．ここに改めて深く感謝申し上げます．また，北村研究室同期の大喜多景元氏，森末結氏とは，この１年間多くの時間を共有し，研究室生活を過ごしました．何気ない話をしたり，一緒にご飯を食べたり，ボードゲームをしたりしとても楽しい１年となりました．ここに感謝申し上げます．

最後になりますが，現在に至るまで私の学生生活を金銭的に支え，暖かく見守って下さった両親には感謝の念に堪えません．これまで本当にありがとうございました．

参考文献

[1] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 556–562, 2000.

[2] P. Smaragdis, “Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs,” in *Proc. Int. Conf. Indep. Compon. Anal. Blind Signal Separation*, vol. 3195, pp. 494–499, 2004.

[3] O. Gillet and G. Richard, “Transcription and separation of drum signals from polyphonic music,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 529–540, 2008.

[4] A. I. Mezza, R. Giampiccolo, A. Bernardini, and A. Sarti, “Toward deep drum source separation,” *Pattern Recognit. Lett.*, vol. 183, pp. 86–91, 2024.

[5] J. B. Allen and L. R. Rabiner, “A unified approach to short-time Fourier analysis and synthesis,” *Proc. IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977.

[6] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[7] A. Ajit, K. Acharya, and A. Samanta, “A review of convolutional neural networks,” in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng.*, pp. 1–5, 2020.

[8] I. D. Mienye, T. G. Swart, and G. Obaido, “Recurrent neural networks: A comprehensive review of architectures, variants, and applications,” *Information*, vol. 15, no. 9, Art. no. 517, 2024.

[9] Y. Luo and N. Mesgarani, “Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, pp. 3942–3951, 2018.

[11] S. Birnbaum, V. Kuleshov, S. Z. Enam, P. W. Koh, and S. Ermon, “Temporal FiLM: Capturing long-range sequence dependencies with feature-wise modulations,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, pp. 10287–10298, 2019.

[12] H. Sato, T. Moriya, M. Mimura, S. Horiguchi, T. Ochiai, T. Ashihara, A. Ando, K. Shinayama, and M. Delcroix, “SpeakerBeam-SS: Real-time target speaker extraction with lightweight Conv-TasNet and state space modeling,” in *Proc. Interspeech*, pp. 5033–5037, 2024.

[13] D. A. M. G. Wisnu, R. E. Zezario, S. Rini, F.-R. Li, Y.-T. Peng, H.-M. Wang, and Y. Tsao, “STSM-FiLM: A FiLM-conditioned neural architecture for time-scale modification of speech,” *arXiv preprint arXiv:2510.02672*, 2025.

[14] S. Alfattama and A. Vaish, “Anatomically adaptive feature-wise linear modulation for deep learning-based low-dose CT denoising,” *arXiv preprint arXiv:2501.19128*, 2025.

[15] M. Brockschmidt, “GNN-FiLM: Graph neural networks with feature-wise linear modulation,” in *Proc. Int. Conf. Mach. Learn.*, vol. 119, pp. 1144–1152, 2020.

[16] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman, “Learning to groove with inverse sequence transformations,” in *Proc. Int. Conf. Mach. Learn.*, vol. 97, pp. 2269–2279, 2019.

[17] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, pp. 234–241, 2015.

[18] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. Int. Conf. Learn. Represent.*, 2015.

[19] C. Zhang, Y. Shao, H. Sun, L. Xing, Q. Zhao, and L. Zhang, “The WuC-Adam algorithm based on joint improvement of warmup and cosine annealing algorithms,” *Math. Biosci. Eng.*, vol. 21, no. 1, pp. 1270–1285, 2024.

[20] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

発表文献一覧

国内学会

1. **片山碧人**, 北村大地, “ドラムセットのマルチトラック録音における被り音抑圧の検討,” 第 28 回日本音響学会関西支部 若手研究者交流研究発表会, p. 18, 2025.

受賞

1. 第 28 回日本音響学会関西支部 若手研究者交流研究発表会 奨励賞

付録 A

追加実験

A.1 提案手法をサンプリング周波数 44.1 kHz のデータセットで学習

本付録では追加で行った実験について述べる．3 章で作成したデータセットは，サンプリング周波数を 44.1 kHz から 16 kHz にダウンサンプリングしているため，高周波の情報が失われている．LarsNet の公開されているモデルは 44.1 kHz のデータセットを用いて学習を行っているため，高周波の情報を失ったデータを使用するのでは本来の分離性能を引き出すことができていない可能性がある．また，高周波の成分に分離性能を向上させるような情報がある可能性もある．そこで，サンプリング周波数を 44.1 kHz のままで混合した近接マイクロホン信号のデータセットを作成し，公開モデル LarsNet で評価および提案手法で学習・評価を行う．サンプリング周波数を 44.1 kHz にするにあたって提案手法のエンコーダとデコーダのカーネルサイズを 256 点（16 ms）から 512 点（約 11.6 ms），ストライドを 128 点（8 ms）から 256 点（約 5.8 ms）に変更した．また，FiLM ジェネレータへ入力するメルスペクトログラムの STFT 設定の窓長を 2048 点（128 ms）から 4096 点（約 92.9 ms），シフト長を 1024 点（64 ms）から 2048 点（約 46.4 ms）に変更した．

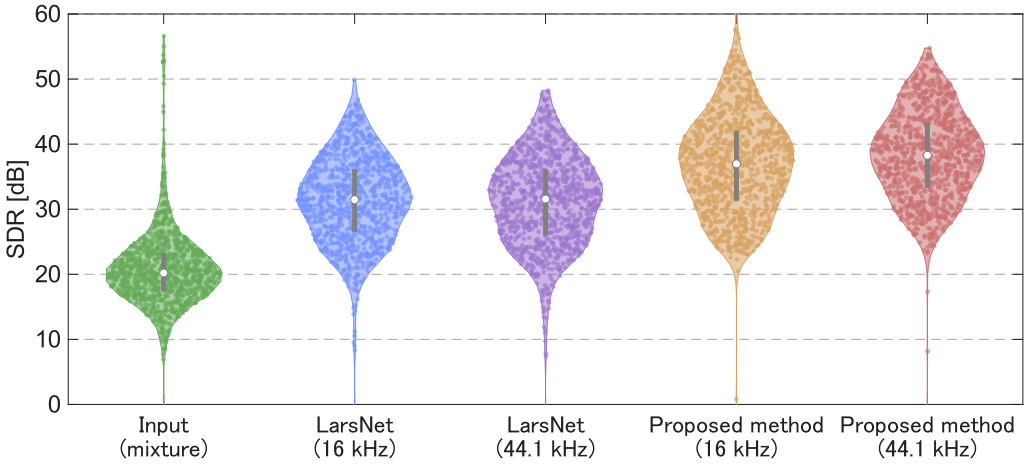
A.2 結果と比較

サンプリング周波数が 44.1 kHz のデータセットで学習および評価をした結果を Table A.1 および Fig. A.1 に示す．評価は，4 章と同一条件で行った．Table A.1 に示す公開モデル LarsNet の平均 SDR は，16 kHz の KD，SD，および HH で 30.57 dB，16.71 dB，および −5.29 dB となり，44.1 kHz で 29.95 dB，17.99 dB，および −5.41 dB となった．KD および HH は SDR が悪化したが，SD は改善した．また，提案手法の平均 SDR は，16 kHz で 36.37 dB，22.87 dB，および 2.25 dB，44.1 kHz で 37.07 dB，25.52 dB，および 3.50 dB となった．サンプリング周波数が 44.1 kHz のデータセットを用いて，公開モデル LarsNet で SDR を算出した場合でも提案手法の SDR を超えることはできないことが分かる．LarsNet

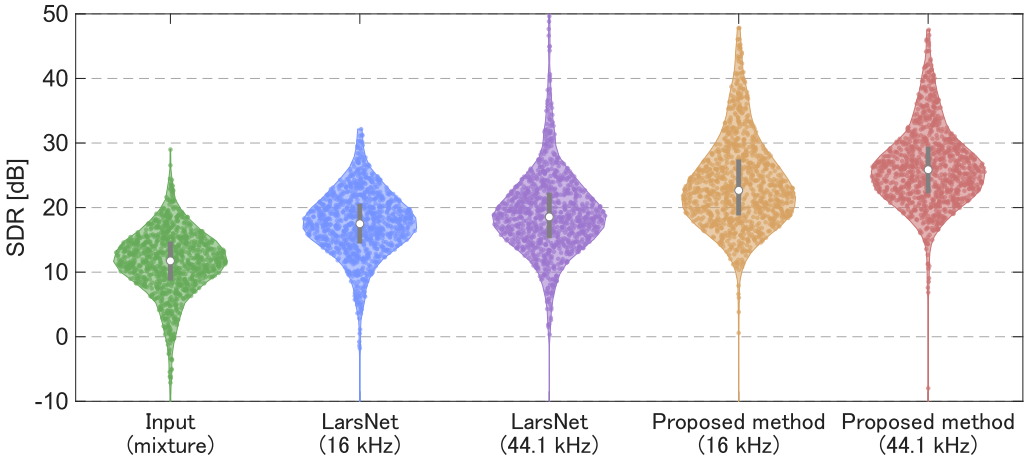
Table A.1 Average SDR [dB] on test data under 16 kHz and 44.1 kHz conditions

Target source	**Input (mixture)**	**LarsNet (16 kHz)**	**LarsNet (44.1 kHz)**	**Proposed method (16 kHz)**	**Proposed method (44.1 kHz)**
KD	19.75	30.57	29.95	36.37	37.07
SD	10.63	16.71	17.99	22.87	25.52
HH	−22.32	−5.29	−5.41	2.25	3.50

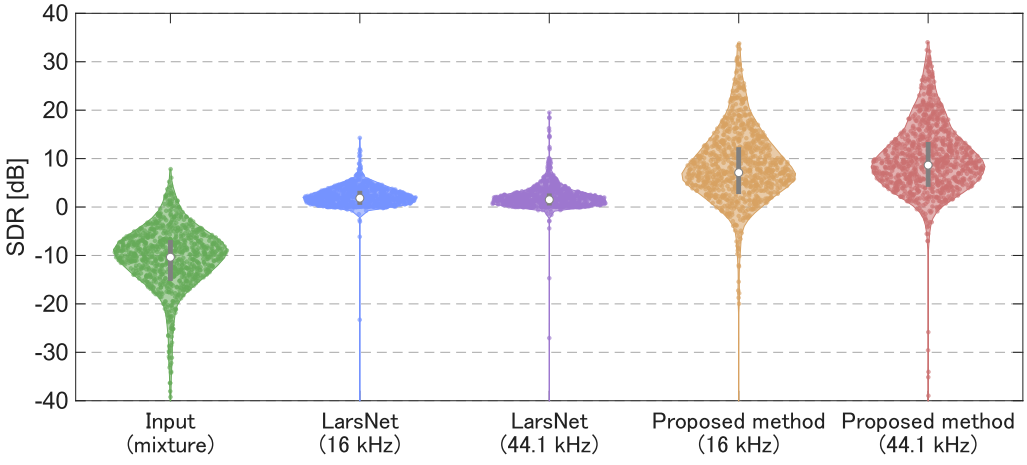
では音源によって SDR が改善する場合と改善しない場合があり，公開モデル LarsNet においてはサンプリング周波数を 44.1 kHz にすることによる優位性は不確かである．ただし，データセット形式が学習データの形式と大きく異なるためこの結果は参考値である．提案手法の 16 kHz と 44.1 kHz を比較すると 44.1 kHz の SDR が全ての音源で高いことが分かる．提案手法においては，高周波の情報がある場合の被り音抑圧性能が高くなる結果となった．



(a)



(b)



(c)

Fig. A.1 Violin plots of SDR on test data for (a) KD, (b) SD, and (c) HH, comparing input mixture, LarsNet (16 kHz), LarsNet (44.1 kHz), proposed method (16 kHz), and proposed method (44.1 kHz).