卒業研究論文

論 文 題 目

ドラム演奏のための DNN に基づく リアルタイム叩打音量可視化システム

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

香川高等専門学校

DNN-based real-time gain visualization system for drum performance

Kagemoto Okita

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

Abstract

This thesis addresses the difficulty of assessing drum-part loudness balance during practice from a single microphone recording, because performers and listeners perceive different balances. I propose a DNN-based method that estimates frame-wise hit gains of kick drum , snare drum, and hi-hat directly from the mixed waveform and visualizes the balance for feedback. The model is a causal deep recurrent regression network that does not rely on explicit source separation and is designed for sequential processing. Training uses mixed drum signals generated from isolated sourcewise stems, and evaluation is conducted on actual recording signals. Compared with a conventional method based on supervised nonnegative matrix factorization under matched conditions, the proposed method yields lower estimation error, indicating more reliable feedback for balancing drum performance.

Keywords: *Drum performance, Drum hit gain estimation, Deep neural network, Gated recurrent unit*

（和訳）

演奏者と聴取者では音量バランスの感じ方が異なるため，単一マイクロホンの録音から練習時のドラム各パートの音量バランスを評価することは難しい．本研究では，混合波形からキックドラム，スネアドラム，ハイハットの時間フレームごとの叩打ゲインを直接推定し，音量バランスを可視化してフィードバックする深層ニューラルネットネットワークに基づいた手法を提案する．本手法は明示的な音源分離に依存しない因果的な深層再帰回帰モデルであり，逐次処理を想定して設計した．学習には各音源のステムから生成した混合ドラム信号を用い，評価は実録音信号に対して行った．教師あり非負値行列因子分解に基づく従来法と同条件で比較した結果，提案法の推定誤差が小さく，ドラム演奏の音量バランス把握により信頼性の高いフィードバックが得られることを示した．

目次

第 1 章

緒言

1.1 背景

一般的なドラムセットは Fig. 1.1 に示すように，複数の音源を組み合わせた楽器である．具体的には，キックドラム（kick drum: KD），スネアドラム（snare drum: SD），及びハイハットシンバル（hi-hat cymbal: HH），ハイタム，ロータム，フロアタム，及びクラッシュシンバル等の音源で構成されている．これらの複数の音源をリズムパターンに合わせて演奏することが一般的であり，楽曲の進行を担う中心的な役割を持った楽器である．

ドラム演奏において，KD，SD，及び HH の音量バランスは演奏全体の印象を大きく左右する．一般的には KD が最も強いエネルギーで演奏されるべきであり，次いで SD 及び HH の順で相対的に小さくなる音量バランスが標準とされており，このような音量バランスを基本演奏技術として習得することが重要である．そのうえで，曲調や個性に応じた音量の調整を行うことが望ましい [1].

ドラム演奏者が音量バランスを認識することの難しさとして，Fig. 1.2 に示すように，演奏者の位置で聞いた際の各音源の音量バランスと，例えばドラムセットの正面 1 m で聞いた際の各音源の音量バランスは明確に異なるという問題が挙げられる．原因は，空間的な聴取位置の違いに加えて，一般的に音響信号の伝搬特性が周波数毎に大きく異なることが原因の一つと考えられる．具体的には，低周波数帯域の音程全方位に拡散し，高周波数帯域の音は直線的に伝搬する．他にも，各音源で指向特性（音の放射方向の空間的な特性）が異なることや，室内の反射音といった演奏環境の影響にも起因して，ドラムセットの各音源の音量バランスの聞こえ方は変化しうる [2, 3]．このような聞こえ方の違いによってドラム演奏における各音源のバランスは演奏者と聴衆で異なるため，演奏者が聴衆に提示したい音量バランスを自分で把握することは困難である．しかしながら，聴衆にとって理想的な音量バランスで演奏する技術の習得はドラム演奏において重要であるため，演奏の録音を自分で聴取することや，ドラム演奏のトレーナーなどの技術的なアドバイスを受けながら身に着けていく過程が求められる．

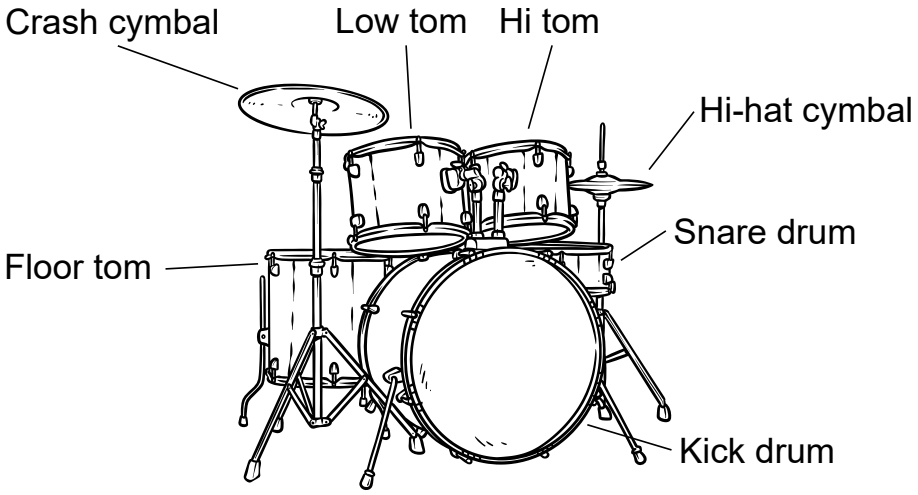


Fig. 1.1 Drum set layout consisting of multiple sound sources.

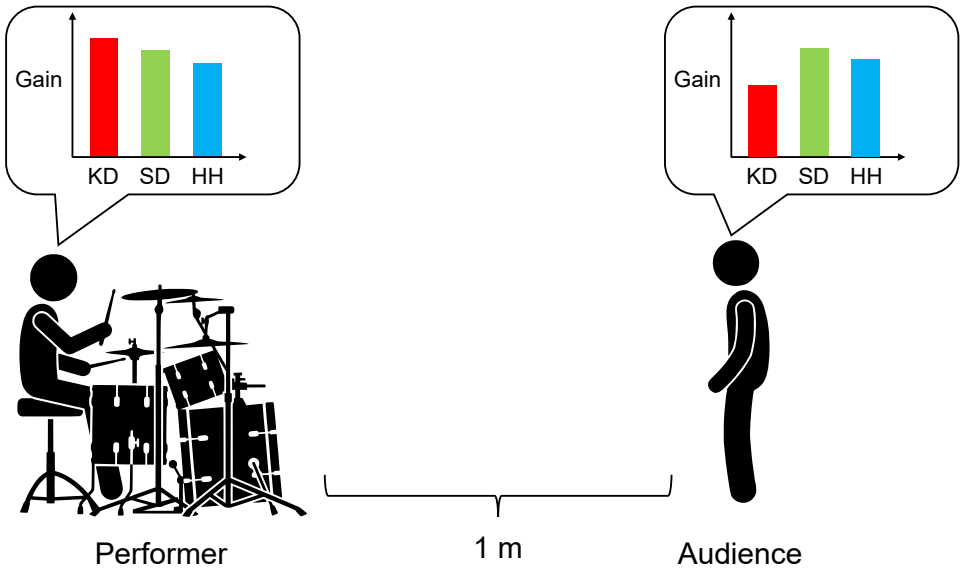


Fig. 1.2 Perceived gain balance at the performer position and at 1 m in front of the drum set.

1.2　目的

前節で述べたドラム演奏における音量バランスの問題の解決方法として，特定の位置での聞こえ方（各音源の音量バランス）を演奏者にリアルタイムに可視化・フィードバックするシステムが提案されている [1]．このシステムの概要を Fig. 1.3 に示す．このシステムでは，把握したい位置に 1 個のマイクロホンを配置し，その場所の音響信号をシステムに取り込むことで，KD，SD，及び HH 等の各音源の音量バランスをリアルタイムにグラフで表示し，演奏者に提示することを実現している．演奏者はドラムの演奏と同時に各音源の（マイクロホン位置

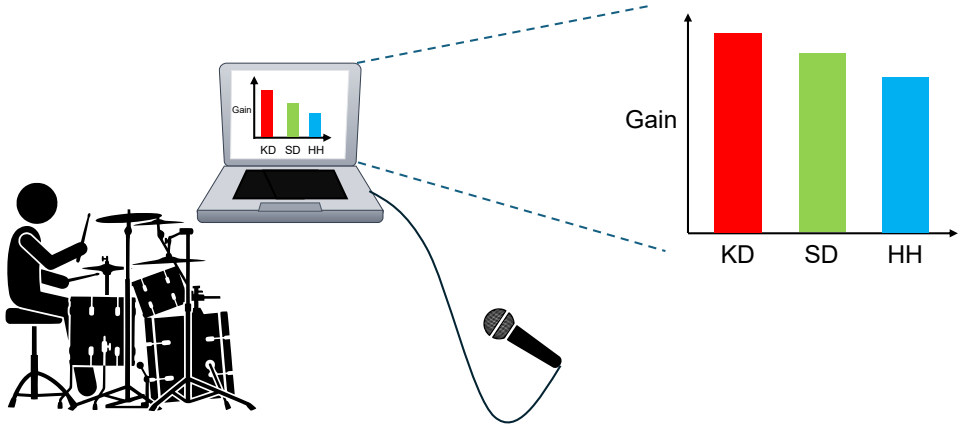


Fig. 1.3 Overview of the real-time drum-hit gain visualization system.

での）音量バランスを視覚的に把握できるため，演奏者と聴衆の聞こえ方の差の理解や適切な音量バランスでの演奏技術修得に役立つ．これによって，録音の手間やトレーナーの助けを借りる必要がなくなる利便性がある．本論文では以後，この Fig. 1.3 のシステムを「リアルタイム叩打音量可視化システム」や単に「提案システム」と呼ぶ．ただし，提案システムは音量バランスの良し悪しを自動判定するものではない．観客位置に設置したマイクロホンで収音された音響信号に基づく各音源の音量関係を可視化し，演奏者がその情報を参照して主体的に演奏を調整するための支援を目的としている．

一般にドラム演奏におけるリズムパターンは，KD と SD や SD と HH 等，複数の音源を同時に叩打するタイミングが多く存在する．そのため，前述のリアルタイム叩打音量可視化システムでは，Fig. 1.4 に示すように，同時に演奏された複数の音源の信号を分離したうえで，各音源の音量を推定・表示する必要がある．文献 [1] では，非負値行列因子分解（nonnegative matrix factorization: NMF）[4, 5] に基づく音源分離手法が採用されている．具体的には，サンプルとして KD，SD，及び HH などの各音源のサンプル音をシステムに事前に入力し，そのサンプル音を活用した教師あり NMF（supervised NMF: SNMF）[6, 7, 8] で音源分離している．

本研究では，既存のシステムよりも高精度な音量バランスを可視化と演奏者へのフィードバックを目的とし，深層ニューラルネットワーク（deep neural network: DNN）に基づく音源分離手法を組み合わせたリアルタイム叩打音量可視化システムを提案及び開発する．既存のシステムは事前に録音したサンプル音に基づく音源分離を適用しているが，実際に使用する際のドラム演奏音と各音源のサンプル音の音色の違いには敏感であり，叩打方法や叩打強度が異なる場合に音源分離精度が劣化してしまう問題が残る．本研究ではより柔軟な音源分離として，ドラムのデータセットで学習した DNN モデルを用いることで，サンプル音の事前録音を不要としながら音色の違いに対して頑健な叩打音量の可視化を実現することを目指す．

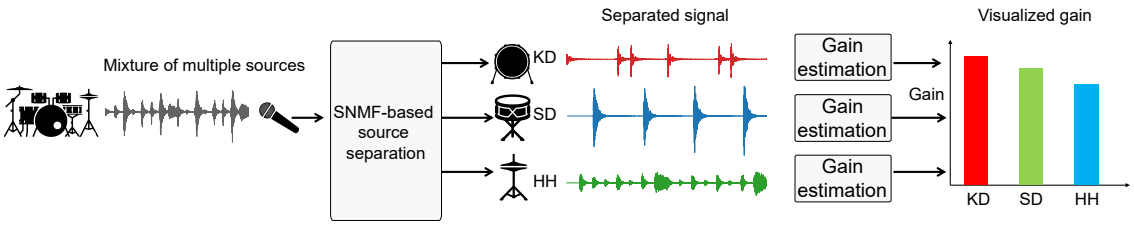


Fig. 1.4 Overview of the conventional visualization system.

1.3　本論文の構成

本論文の構成は次の通りである．第 2 章では深層学習の基礎と比較対象である SNMF を整理する．第 3 章では提案手法として，DNN による叩打音量推定の枠組み，教師データセットの作成，ネットワーク構造を述べる．第 4 章では実験条件・評価指標を示し，提案法と従来法の比較結果を示す．第 5 章では結言と今後の課題を述べる．

第 2 章

基礎知識

2.1 まえがき

本章では，本研究で用いる深層学習に基づく推定手法と，比較対象とする従来手法である教師あり NMF の基礎を整理する．第 3 章では，混合ドラム音から KD，SD，及び HH の叩打音量をフレームごとに推定するモデルを提案するため，ここでは回帰としての学習の考え方，時系列波形に対する畳み込みによる特徴抽出，時間方向の文脈を扱うための再帰型モデルの基本を述べる．さらに，従来法である教師あり NMF がどのような処理の流れで推定を行うかを示し，リアルタイム性や頑健性の観点での性質を整理する．

2.2 DNN

2.2.1 DNN の基礎と本研究における入出力

深層ニューラルネットワーク（deep neural network: DNN）は，入力から出力への対応関係を多数のパラメータを含む非線形写像で表現し，データから目的を他生するための最適なパラメータを学習することで，複雑な入出力関係を近似できる．本研究では，KD, SD, 及び HH 複数の音源が混合したモノラルのドラムセットの観測音響信号を入力として，KD，SD，及び HH の叩打音量を推定するために DNN を用いる [9, 10].

今，DNN の入力を $\boldsymbol{x} \in \mathbb{R}^{D_{\text{in}}}$，ラベルを $\boldsymbol{y} \in \mathbb{R}^{D_{\text{out}}}$，予測を $\hat{\boldsymbol{y}} \in \mathbb{R}^{D_{\text{out}}}$ と定義する．ここで，D_{in} 及び D_{out} はそれぞれ DNN の入力の次元と出力の次元である．このとき，一般に DNN の入出力は次式のように表される．

$$\hat{\boldsymbol{y}} = \text{DNN}(\boldsymbol{x}; \theta) \tag{2.1}$$

ここで，θ は DNN に含まれる学習可能なパラメータの集合を表す．したがって，DNN$(\cdot)$ は $\mathbb{R}^{D_{\text{in}}} \to \mathbb{R}^{D_{\text{out}}}$ なる非線形写像である．

もっとも単純な全結合型 DNN は，次式で示される 1 層の非線形写像を複数層重ねた構造を

持つ．

$$\boldsymbol{p}^{(u)} = \phi^{(u)}\left(\boldsymbol{A}^{(u)}\boldsymbol{x}^{(u)} + \boldsymbol{b}^{(u)}\right) \tag{2.2}$$

ここで，$u = 1, 2, \cdots, U$ は第何層目かを表すインデクスであり，$\boldsymbol{x}^{(1)} = \boldsymbol{x}$ かつ $\boldsymbol{p}^{(U)} = \hat{\boldsymbol{y}}$，第 u 層の入出力は $\boldsymbol{x}^{(u)} \in \mathbb{R}^{D_{\text{in}}^{(u)}}$ 及び $\boldsymbol{p}^{(u)} \in \mathbb{R}^{D_{\text{out}}^{(u)}}$ と定義する．また，$\phi^{(u)}(\cdot)$ は活性化関数（非線形関数）を表す．この層に含まれる学習可能なパラメータは $\boldsymbol{A}^{(u)} \in \mathbb{R}^{D_{\text{out}}^{(u)} \times D_{\text{in}}^{(u)}}$ 及び $\boldsymbol{b}^{(u)} \in \mathbb{R}^{D_{\text{out}}^{(u)}}$ である．

式 (2.1) 及び (2.2) より，全 U 層で構成される全結合型 DNN に含まれる学習可能なパラメータは $\theta = \{\boldsymbol{A}^{(1)}, \cdots, \boldsymbol{A}^{(U)}, \boldsymbol{b}^{(1)}, \cdots, \boldsymbol{b}^{(U)}\}$ となる．これらのパラメータの学習は次の最適化として実現される．

$$\underset{\theta}{\text{Minimize}}\ \mathcal{D}(\boldsymbol{y}|\hat{\boldsymbol{y}})\ \ \forall \boldsymbol{x}\ \ \text{s.t. } \hat{\boldsymbol{y}} = \text{DNN}(\boldsymbol{x}; \theta) \tag{2.3}$$

ここで，$\mathcal{D}(\cdot|\cdot)$ は 2 つの入力間の誤差を返す関数である．すなわち最適化問題 (2.3) は，すべての（学習データ中の）入力 $\boldsymbol{x}$ を与えた際の予測 $\hat{\boldsymbol{y}}$ とラベル $\boldsymbol{y}$ 間の誤差が最小化されるパラメータ θ を求める問題である．損失関数関数 $\mathcal{D}(\cdot|\cdot)$ は用途に応じて様々な形で設計される．回帰の DNN においては，平均二乗誤差（mean squared error: MSE）が広く用いられる代表的な損失関数関数であり，一般的に損失関数と呼ばれる．

$$\mathcal{D}(\boldsymbol{y}|\hat{\boldsymbol{y}}) = \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2 \tag{2.4}$$

ここで，$\|\cdot\|_2$ は ℓ_2 ノルムを表す．

全結合型 DNN は膨大なパラメータを持つため高い表現力を有する．一方で過学習を起こしやすいため，現在ではより効率的な学習のために目的に応じて明示的な学習を促すネットワーク構造を用いることが一般的である．次項と次々項では，音響信号処理等の時系列信号を対象とした DNN でよく用いられるネットワーク構造として，畳み込みニューラルネットワーク（convolutional neural network: CNN）及び再帰型ニューラルネットワーク（reccurent neural network: RNN）についてそれぞれ述べる．

2.2.2　畳み込みニューラルネットワーク

畳み込みニューラルネットワーク（convolutional neural network: CNN）は [11]，入力に含まれる局所的な構造の特徴を抽出するのに適したニューラルネットワークである．例えば音響信号は時間方向に連続する時系列信号であり，短い時間範囲に局所パターンとして特徴的な構造が現れる．また，二次元画像においても局所的な領域に特徴的な構造をもつ場合が多い．CNN で用いられる畳み込み層はそのような局所パターンを捉えるフィルタを学習し，パラメータの学習がより効果的になるような特徴量列へと変換する．

本稿では複数チャネルからなる 1 次元の時系列信号の入力を例とし，時間方向に畳み込みを適用する 1 次元畳み込み層の CNN について述べる．今，信号長が N でチャネル数が C の入力を $\boldsymbol{X} \in \mathbb{R}^{C \times N}$ と定義し，その要素を $x_c[n]$ と表す．ここで，$c = 1, 2, \cdots, C$ 及び

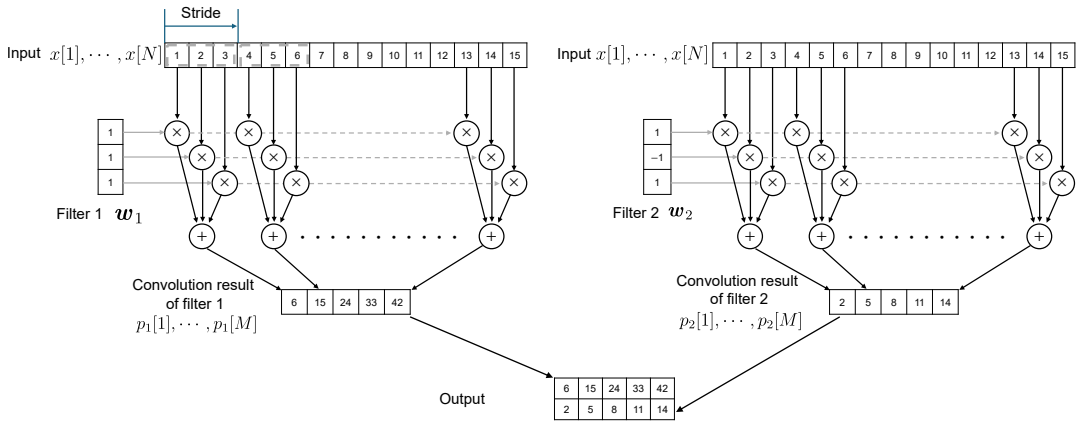


Fig. 2.1 Conceptual diagram of 1D convolution.

$n = 1, 2, \cdots, N$ はそれぞれチャネルのインデクス及び離散時間インデクスを表す．このとき，1 次元畳み込み層の入出力関係は次式で表せる．

$$\begin{aligned} p_j[m] &= b_j + \sum_c \boldsymbol{w}_{j,c} * \boldsymbol{x}_c[m] \\ &= b_j + \sum_c \sum_{r=1}^{R} w_{j,c}[r] x_c[\eta(m-1) + (r-1) + 1] \end{aligned} \tag{2.5}$$

ここで，$\boldsymbol{w}_{j,c}$ 及び $\boldsymbol{x}_c[m]$ はそれぞれ学習可能フィルタ及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．

$$\boldsymbol{w}_{j,c} = [w_{j,c}[1], w_{j,c}[2], \cdots, w_{j,c}[r], \cdots, w_{j,c}[R]]^{\mathrm{T}} \in \mathbb{R}^R \tag{2.6}$$

$$\boldsymbol{x}_c[m] = [x_c[\eta(m-1)+1], x_c[\eta(m-1)+2], \cdots, x_c[\eta(m-1)+R]]^{\mathrm{T}} \in \mathbb{R}^R \tag{2.7}$$

また，$j = 1, 2, \cdots, J$ は出力のチャネルインデクス，$m = 1, 2, \cdots, M$ は出力の時間フレームインデクス，η はストライド長，R はフィルタ長（カーネル長），$p_j[m]$ は j 番目のチャネルの時間フレーム m に対応する出力，b_j は j 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. 2.1 は式 (2.5) 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 $J = 2$，信号長 $N = 10$，フィルタ長 $R = 3$，ストライド長 $\eta = 2$ の例を図示しており，さらに式 (2.5) 中のバイアス b_j 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. 2.1 のように計算した結果を出力する．このような計算過程で用いられるフィルタ $\boldsymbol{w}_{j,c}$ を損失関数の最小化として学習することで，入力の局所時間の効果的・効率的な非線形写像が獲得できる利点がある．また，通常の全結合層と比較してパラメータ数を大幅に少なくできるため過学習のリスクを低減できるほか，畳み込み層を複数回通すネットワークに拡張することで，最初の入力 $\boldsymbol{X}$ に対する時間方向のフィルタの受容野を効率的に広げることも可能となる．このような畳み込み層を DNN の入力の直後に何層か用いて，出力側では全結合層を 1 層以上接続するネットワーク構造を CNN と呼び，その有用性から広く用いられている．

2.2.3　再帰型ニューラルネットワークとゲーツ付き再帰ユニット

RNN は，時系列データを扱うために設計されたニューラルネットワークである．RNN は，現在の入力と過去の状態から現在の状態を更新し，その状態に基づいて出力を計算する．これにより，ある時刻の推定に過去の情報も用いられる構造となるため時間的に連続的な推定結果を得る用途などで広く用いられる．

基本的な RNN では，時刻 t における隠れ状態 $\boldsymbol{h}_t \in \mathbb{R}^{D_{\text{hidden}}}$ を次式で更新する．

$$\boldsymbol{h}_t = \phi\left(\boldsymbol{A}_h \boldsymbol{x}_t + \boldsymbol{U}_h \boldsymbol{h}_{t-1} + \boldsymbol{b}\right) \tag{2.8}$$

ここで $\boldsymbol{x}_t \in \mathbb{R}^{D_{\text{in}}}$ は時刻 t の入力，$\boldsymbol{A}_h \in \mathbb{R}^{D_{\text{h}} \times D_{\text{in}}}$ 及び $\boldsymbol{U}_h \in \mathbb{R}^{D_{\text{hidden}} \times D_{\text{hidden}}}$ はそれぞれ $\boldsymbol{x}_t$ 及び $\boldsymbol{h}_{t-1}$ のパラメータ行列，$\boldsymbol{b} \in \mathbb{R}^{D_{\text{hidden}}}$ はバイアスベクトルである．

Fig. 2.2 は式 (2.5) の再帰層及び隠れ状態を含む DNN の時間展開を示した図である．ただし，図中の $\hat{\boldsymbol{y}}$ は再帰層を含む DNN 全体の出力である．時刻 t ごとに式 (2.5) の演算が繰り返され，隠れ状態が未来の時刻へ受け渡される形となっていることがわかる．

単純な再帰型モデルは長い系列で学習が不安定になりやすいことが知られている．この問題への対処として，情報の保持と更新を制御するゲート機構を持つゲート付き再帰ユニット (gated recurrent unit: GRU) が提案されている [12]．GRU は更新ゲートとリセットゲートにより，過去の状態をどの程度保持するか，現在の入力をどの程度反映するかを調整する構造を持つ．GRU の更新式を次に示す．

$$\boldsymbol{z}_t = \sigma\left(\boldsymbol{A}_z \boldsymbol{x}_t + \boldsymbol{U}_z \boldsymbol{h}_{t-1} + \boldsymbol{b}_z\right) \tag{2.9}$$

$$\boldsymbol{r}_t = \sigma\left(\boldsymbol{A}_r \boldsymbol{x}_t + \boldsymbol{U}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r\right) \tag{2.10}$$

$$\tilde{\boldsymbol{h}}_t = \tanh\left(\boldsymbol{A}_h \boldsymbol{x}_t + \boldsymbol{U}_h \left(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}\right) + \boldsymbol{b}_h\right) \tag{2.11}$$

$$\boldsymbol{h}_t = \left(1 - \boldsymbol{z}_t\right) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t \tag{2.12}$$

ここで $\boldsymbol{z}_t$ は更新ゲート，$\boldsymbol{r}_t$ はリセットゲート，$\tilde{\boldsymbol{h}}_t$ は候補状態である．σ はシグモイド関数，$\odot$ は要素ごとの積である．$\boldsymbol{A}_z$，$\boldsymbol{A}_r$，$\boldsymbol{A}_h$，$\boldsymbol{U}_z$，$\boldsymbol{U}_r$ 及び $\boldsymbol{U}_h$ はパラメータ行列，$\boldsymbol{b}_z$，$\boldsymbol{b}_r$ 及び $\boldsymbol{b}_h$ はバイアスベクトルである．

GRU セルの内部構造を Fig. 2.3 に示す．更新ゲートは過去状態と候補状態の混合比を決め，リセットゲートは候補状態の算出時に過去情報をどの程度利用するかを制御する．このような機構を用いることで長期的な時系列の記憶と短期的な時系列の活用の両方を同時に実現できる RNN を構築できることが知られており，GRU を用いた RNN も広く用いられる．

時系列信号を対象とする RNN は，リアルタイム信号処理への応用が容易である利点がある．これは，一定時間毎の入力を常に入れ続け，その都度出力を得ることができるという利点に起因している．本論文においても，一定の時間区間の音響観測信号を入力し，その時間区間に対する推定を出力することを反復的に継続できる DNN モデルとして，この RNN 構造を利用する．

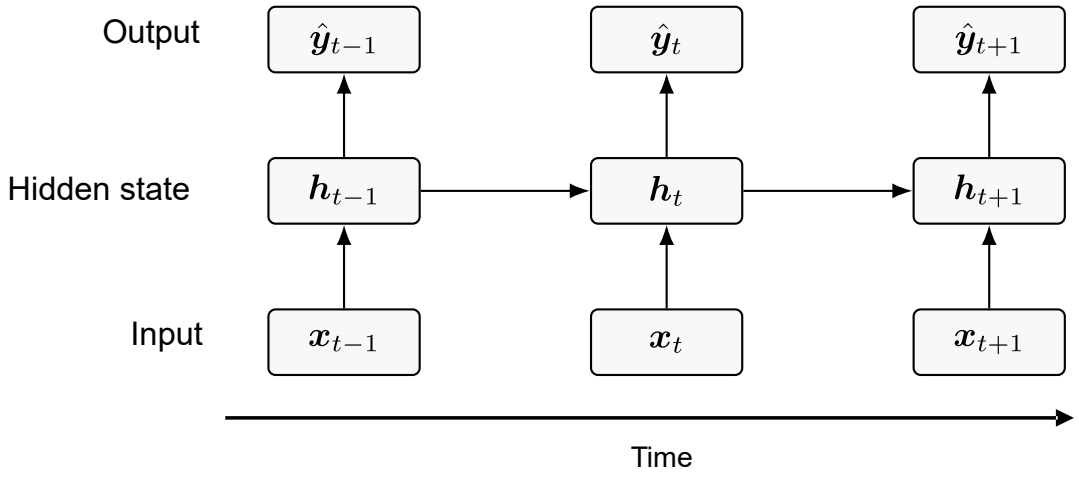


Fig. 2.2 Temporal unrolling of a recurrent model.

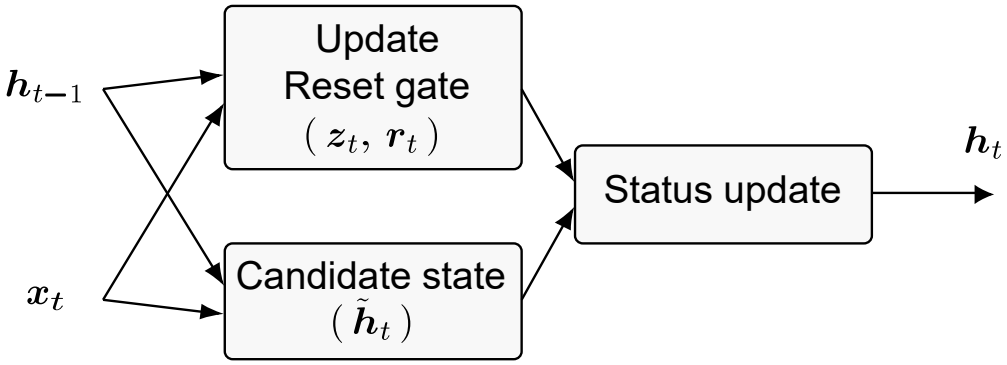


Fig. 2.3 Block diagram of a GRU cell.

2.3 教師あり NMF を用いたドラムセットの叩打音量推定

本節では，本論文で目的とするドラムセットの各音源の叩打音量のリアルタイム推定において，既に提案されている技術である SNMF を用いた叩打音量推定 [1] について説明する．この手法は，SNMF を用いた技術であるため，まず NMF 及びその教師あり拡張である SNMF について述べる．

NMF は，非負の観測行列を非負の基底行列と非負の係数行列の行列席で低ランク近似する手法である．音響分野では，Fig. 2.4 に示すように，観測信号のスペクトログラム（時間周波数行列）の振幅値を NMF を適用することで，複数の基底ベクトル（振幅スペクトルのテンプレート）の線形結合として近似することができる．これによって得られた基底スペクトルやその係数を用いて音源分離などの用途で利用されてきた．[13]．また，NMF を教師ありに拡張した SNMF では，あらかじめ各音源に対応する基底をサンプルとなる音響信号から事前に学

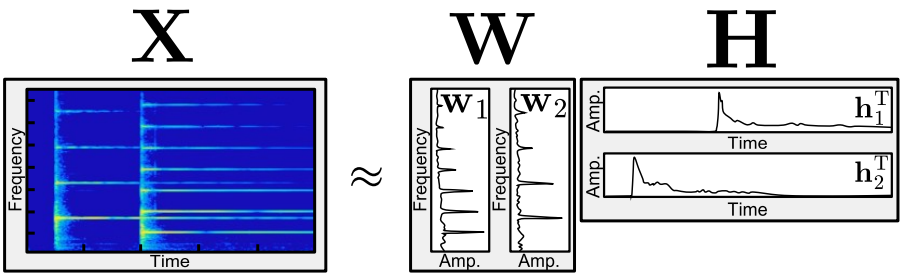


Fig. 2.4 Conceptual diagram of NMF factorization.

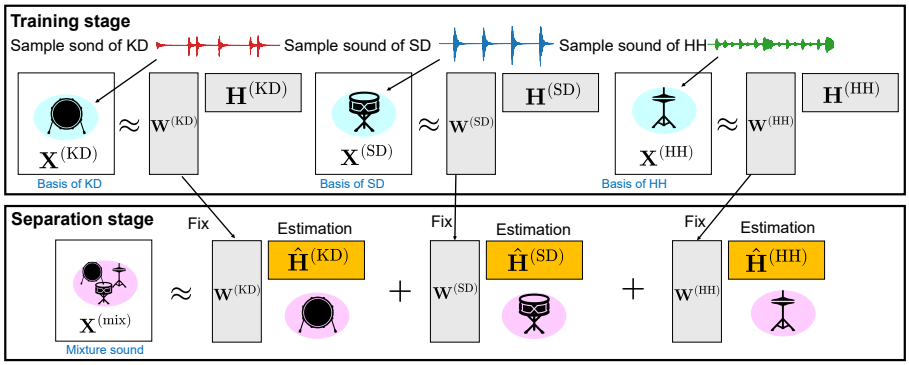


Fig. 2.5 Overview of SNMF-based drum source separation using pre-recorded sourcewise sample souds.

習し，複数の音源が混合した観測信号に対しては学習済みの基底を固定化したうえで係数のみを推定する [1, 6].

従来の SNMF を用いたドラムセットの叩打音量推定法 [1] について，処理の概要を Fig. 2.5 に示し，その詳細を以下に述べる．まず，KD，SD，及び HH のサンプルとなる音響信号を得る必要がある．この時，複数の音源が混合しないように，KD，SD，及び HH を個別に録音したサンプル信号を得る必要がある．実用的には，この叩打音量システムを用いる直前に，初期化として各音源のサンプル信号を録音する工程を想定している．こうして得られた KD，SD，及び HH の個別の音響信号の振幅スペクトログラム $\mathbf{X}^{\mathrm{KD}}$，$\mathbf{X}^{\mathrm{SD}}$，及び $\mathbf{X}^{\mathrm{HH}}$ に対して，Fig. 2.5 の学習ステージに示すようにそれぞれ NMF を適用し，各音源に対応する基底行列 $\mathbf{W}^{\mathrm{KD}}$，$\mathbf{W}^{\mathrm{SD}}$，及び $\mathbf{W}^{\mathrm{HH}}$ と，それぞれの係数行列 $\mathbf{H}^{\mathrm{KD}}$，$\mathbf{H}^{\mathrm{SD}}$，及び $\mathbf{H}^{\mathrm{HH}}$ を次式のように得る．

$$\begin{cases} \mathbf{X}^{\mathrm{KD}} \approx \mathbf{W}^{\mathrm{KD}}\mathbf{H}^{\mathrm{KD}} \\ \mathbf{X}^{\mathrm{SD}} \approx \mathbf{W}^{\mathrm{SD}}\mathbf{H}^{\mathrm{SD}} \\ \mathbf{X}^{\mathrm{HH}} \approx \mathbf{W}^{\mathrm{HH}}\mathbf{H}^{\mathrm{HH}} \end{cases} \tag{2.13}$$

この学習ステージで得られる基底行列 $\mathbf{W}^{\mathrm{KD}}$，$\mathbf{W}^{\mathrm{SD}}$，及び $\mathbf{W}^{\mathrm{HH}}$ は，それぞれ KD，SD，及び HH の音源の振幅スペクトルをテンプレートとして列に持つ行列と解釈できる．次に，分離ステージでは学習済みの $\mathbf{W}^{\mathrm{KD}}$，$\mathbf{W}^{\mathrm{SD}}$，及び $\mathbf{W}^{\mathrm{HH}}$ を固定して複数の音源が混合した信号の

振幅スペクトログラム $\mathbf{X}^{\mathrm{mix}}$ を次式のように近似分解する．

$$\mathbf{X}^{\mathrm{mix}} \approx \mathbf{W}^{\mathrm{KD}}\mathbf{H}^{\mathrm{KD}} \tag{2.14}$$

混合音のスペクトログラムを $\boldsymbol{X}$，基底行列を $\boldsymbol{W}$，活性化行列を $\boldsymbol{H}$ とすると，$\boldsymbol{X} \in \mathbb{R}_{+}^{I\times J}$，$\boldsymbol{W} \in \mathbb{R}_{+}^{I\times K}$，$\boldsymbol{H} \in \mathbb{R}_{+}^{K\times J}$ であり，I は周波数ビン数，J は時間フレーム数，K は基底数を表す．NMF の基本モデルは次式で表せる．Fig. 2.4 に NMF の分解概念を示す．

$$\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H} \tag{2.15}$$

教師ありの場合，$\boldsymbol{W}$ は固定し，$\boldsymbol{H}$ を推定する．典型的には，$\boldsymbol{X}$ と $\boldsymbol{W}\boldsymbol{H}$ の誤差が小さくなるように $\boldsymbol{H}$ を反復更新する．推定された $\boldsymbol{H}$ から各パートの時間変化を得て，叩打音量へ変換し可視化することでフィードバックに用いる．処理の流れは，時間周波数変換，活性化推定，音量算出，可視化という段階的構成となる．

教師あり NMF は解釈性が高い一方で，基底を得るために事前サンプルが必要になり，音色や録音条件が変化すると適合しにくい．また，活性化推定が反復更新を前提とするため，更新回数やチャンク長に応じて計算遅延が発生しやすい．リアルタイム動作を目指す場合，遅延の制約や計算資源の制約の下で性能を維持する設計が難しくなる．このような背景から，第 3 章では分離と推定を一括で扱い，逐次更新による低遅延化を意図した深層学習に基づく回帰モデルを提案する．

2.4 本章のまとめ

本章では，深層学習による回帰推定の基本，音声波形に対する 1 次元畳み込みによる特徴抽出，時系列の文脈を扱う再帰型モデルと GRU の基礎を整理した．さらに比較対象として教師あり NMF の枠組みと処理の流れを示し，事前サンプル依存や反復更新に伴う遅延といった性質を整理した．次章では，これらの基礎を踏まえ，混合ドラム音から叩打音量の低遅延推定を目指すためのデータセット構築とネットワーク構造を提案する．

第 3 章

提案手法

3.1　まえがき

本章では，本論文で提案するリアルタイム叩打音量可視化システムの詳細を示す．3.2 節では，入力信号と出力叩打音量の対応関係を整理し，時間フレーム単位で逐次推定する枠組みを説明する．3.3 節では，時間フレーム内畳み込み，因果畳み込み，GRU，FC 層からなるネットワーク構造を示し，因果性を保った推定の設計意図を述べる．3.4 節では，教師データセットの作成手順を示し，ダウンサンプルとラベル付与の方法を明確にする．3.5 節では，学習条件と評価指標を整理し，提案法の学習がどの前提で行われるかを示す．3.6 節では，本章をまとめる．

3.2　DNN を用いた叩打音量推定

Fig. 3.1 に，DNN に基づくリアルタイム叩打音量可視化システムの概要図を示す．提案システムは従来手法と同様に，1 本のマイクロホンで観測されたモノラルの音響信号を入力とする．この観測信号はドラムセットの演奏の録音を想定しており，KD，SD，及び HH などの複数の音源が混合した信号である．本研究は基礎的な検討であるため，文献 [1] と同様に KD，SD，及び HH の 3 音源の混合のみを想定している．この観測信号は提案手法の DNN へと入力され，各音源の音量が DNN の予測値として出力される．この出力の音量値を演奏者に可視化することで，演奏しながら各音源の音量バランスを確認することが可能となる．ただし，この用途ではシステムの動作のリアルタイム性が重要となるため，提案手法の DNN は観測信号全体を 1 度にまとめて入力するのではなく，Fig. 3.2 のように短時間区間毎に観測信号を入力し，その短時間区間での各音源の音量を推定・出力することを反復的に実行できるモデルとなっている．

提案手法では，観測信号のサンプリング周波数を 16 kHz と定義し，Fig. 3.2 のように短時間区間への分割を行う．分割された 1 つの短時間区間を以後フレームと呼び，提案手法では 1 フレーム 512 サンプル（32 ms）と定める．さらに，その 1 フレームの半分である 256 サン

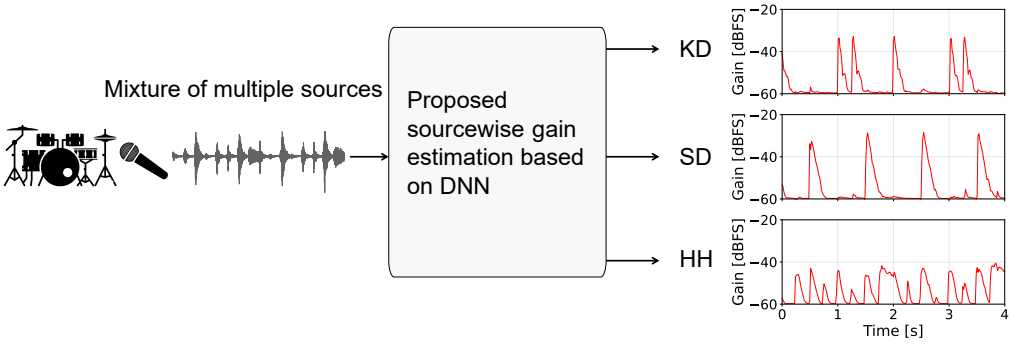


Fig. 3.1 Overview of proposed sourcewise gain estimation method based on DNN.

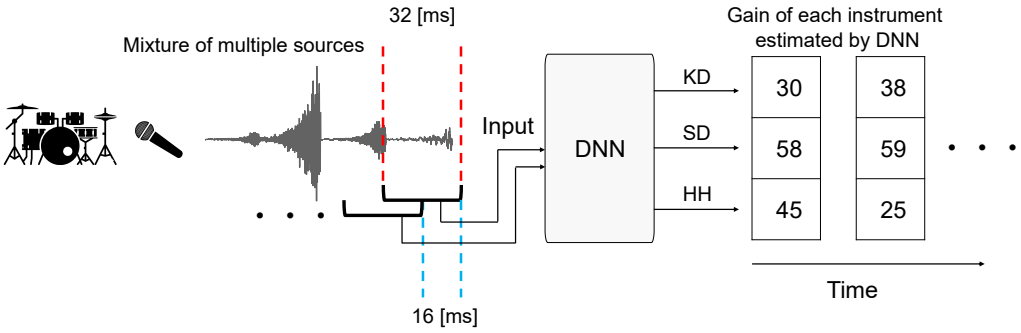


Fig. 3.2 DNN-based time-frame-wise gain estimation from a mixed drum signal.

プル（16 ms）毎に DNN に入力し，各音源の音量を推定する設定を採用している．この設定は，ドラムの打撃音のアタック成分や時間的な変化を十分な時間解像度で捉えつつ，DNN の予測に必要な計算時間を考慮しても過度な遅延をとならない値として実験的に得られた最適値である．

提案手法の DNN は，音響信号の A/D 変換時のダイナミックレンジのフルスケールを基準（0 dB）として，各音源の音量（振幅）の絶対的な値を decibels relative to full scale（dBFS）という単位で予測するモデルとなっている．dBFS はデジタル信号の最大振幅を基準とした相対レベルであり，人間の聴覚が周波数ごとに持つ感度差を直接表す指標ではない．一方で，同一のマイクロホン及び同一の収音条件であれば，各音源間の相対的な音量関係を再現性よく比較できるため，本研究では可視化のための客観指標として dBFS を用いる．

3.3 ネットワーク構造

提案手法で用いるネットワークは，1 次元畳み込み層と GRU を組み合わせた RNN である．このネットワークの設計の指針は次に示すとおりである．

(a) 時間フレーム内の時系列構造を効率的に学習すること
(b) 直近の複数の時間フレームの短期的な依存性を統合的に扱うこと

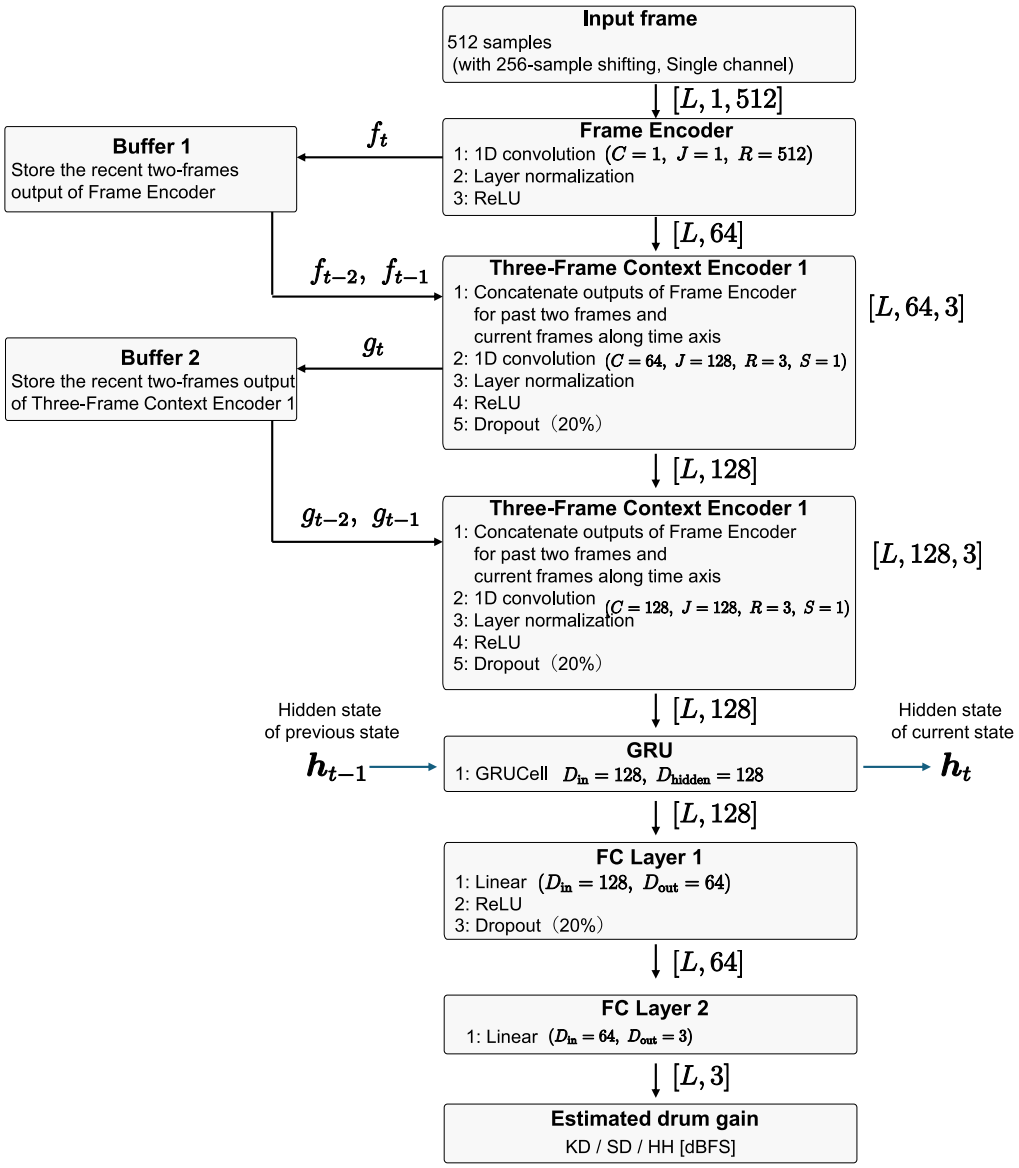


Fig. 3.3 Proposed DNN architecture for drums gain estimation.

(c) 長期的な時間フレーム間の依存性を状態として保持すること
(d) 叩打音量を dBFS という単位で直接的に回帰すること

上記の指針の (a) は 1 次元畳み込み層を複数回適応することで実現している．また，(b) は直近の時間フレームをバッファとして保持し現在の時間フレームの処理時に結合する手法を採用している．さらに (c) は GRU を用いた RNN により実装され，(d) はネットワークの出力側で FC 層を接続することで実現している．

Fig. 3.3 に提案手法のネットワークの全体構造を示す．Frame Encoder は時間フレーム内畳み込み層による特徴抽出を指す．Three-Frame Context Encoder は過去 2 フレームと現在

時間フレームを連結した短期的な時間フレーム間の文脈を扱う畳み込み層であり，第 2 畳み込み層及び第 3 畳み込み層に対応する．Buffer 1 は Frame Encoder の出力の過去 2 フレームを保持するバッファであり，Buffer 2 は Three-Frame Context Encoder 1 の出力の過去 2 フレームを保持するバッファである．その後に接続されている GRU では，より長期的な時間フレームにまたがる特徴量の依存性を保持し，隠れ状態ベクトル $\boldsymbol{h}_t$ は未来の時間フレームの予測へと受け継がれる．最終的に接続されている FC Layer 1 及び FC Layer 2 は 2 層の FC 層であり，適切な次元への非線形圧縮を実現している．DNN の出力は単位を dBFS とする音量の数値が KD，SD，及び HH の 3 音源についてそれぞれ出力されるため，3 次元ベクトルとなる．このネットワークは 1 フレームずつの逐次処理であり，未来の時間フレームの情報を用いない因果性を保証している．従って，時間フレーム単位でのリアルタイムの入出力が可能であり，提案システムの要件を満たす構成となっている．

以後，ネットワーク中の特徴量のより詳細な説明を入力層側から順番に説明する．まず，入力時間フレーム（512 サンプル）は Frame Encoder で特徴ベクトルへ変換される．このとき入力は $[L, 512]$ であり，Frame Encoder に入力する前に $[L, 1, 512]$ に整形する．ここで L は学習データのバッチサイズである．Frame Encoder では 1 次元畳み込み層のフィルタサイズを 512，ストライドを $\eta = 1$ とし，1 フレーム全体を同サイズのフィルタで畳み込んでおり，短時間の波形形状や帯域的なエネルギー分布を 1 層で統合できる．これはフーリエ変換のような固定の時間周波数変換ではなく，タスクに最適なフィルタバンクを学習する役割を持つ．出力は $[L, 64, 1]$ となるため，次元を圧縮して $[L, 64]$ の特徴ベクトルとする．その後，学習安定化のために，レイヤー正規化を通し，非線形関数には rectified linear unit（ReLU）を適用している．

次に，直近の時間フレーム間を統合した短期文脈を考慮するため，Three-Frame Context Encoder 1 では過去 2 フレームの特徴量（Frame Encoder の出力）を保持するバッファから特徴量を呼び出し，現在時間フレームの特徴量と連結する．この操作によって特徴量の次元は $[L, 64, 3]$ となる．Three-Frame Context Encoder 1 中の畳み込み層の入力チャネル数は $C = 64$，出力チャネル数は $J = 128$，フィルタサイズは $R = 3$ である．この畳み込み層の出力 $[L, 128, 1]$ を $[L, 128]$ に整形し，レイヤー正規化，ReLU，20% の確率のドロップアウトを順に適用する．ここで，ドロップアウトは学習時にユニット（ニューロン）を一定確率で無効化することで過学習を抑制する効果がある．その後に続く Three-Frame Context Encoder 2 もほとんど同様の構成となっているが，畳み込み層の入力チャネル数が $C = 128$ となっている点のみ異なる．このように畳み込み層を含む処理ブロックを連結することで，フィルタの受容野が拡大し，打撃の立ち上がりや減衰などの短期的な変化を捉えられる．

その後，GRU セルを用いて長期的な時間依存をモデル化する．GRU はゲート機構により必要な情報を保持・更新でき，単純 RNN より学習が安定しやすい．隠れ状態次元は $D_{\text{hidden}} = 128$ とし，時刻 t の入力と時刻 $t-1$ の隠れ状態から $\boldsymbol{h}_t$ を更新する．推論と同様の逐次処理を行うため，初期状態はゼロベクトルとして与え，過去から未来の時間フレーム方向に隠れ状態を継承する．これにより，連続打撃や余韻といった長い時間文脈を反映できる．

最後に，FC 層を 2 層適用して叩打音量を回帰する．1 層目の FC 層は入力次元が $D_{\text{in}}^{(1)} = 128$，出力次元が $D_{\text{out}}^{(1)} = 64$ であり，ReLU と 20% の確率のドロップアウトを適用する．2 層目の FC 層は入力次元が $D_{\text{in}}^{(2)} = 64$，出力次元が $D_{\text{out}}^{(2)} = 3$ である．最終出力は各時間フレームにおける KD，SD，及び HH の叩打音量に対応し，その数値の単位は dBFS である．

以上の処理は次のように表せる．入力時間フレームを $\boldsymbol{x}_t \in \mathbb{R}^{512}$ とすると，

$$\boldsymbol{f}_t = \phi(\boldsymbol{A}_1 \boldsymbol{x}_t + \boldsymbol{b}_1)\,, \tag{3.1}$$
$$\boldsymbol{g}_t = \phi(\mathcal{C}_2([\,\boldsymbol{f}_{t-2},\, \boldsymbol{f}_{t-1},\, \boldsymbol{f}_t]))\,, \tag{3.2}$$
$$\boldsymbol{u}_t = \phi(\mathcal{C}_3([\boldsymbol{g}_{t-2},\, \boldsymbol{g}_{t-1},\, \boldsymbol{g}_t]))\,, \tag{3.3}$$
$$\boldsymbol{h}_t = \mathcal{G}(\boldsymbol{u}_t, \boldsymbol{h}_{t-1}), \tag{3.4}$$
$$\hat{\boldsymbol{y}}_t = \boldsymbol{A}_o \phi(\boldsymbol{A}_f \boldsymbol{h}_t + \boldsymbol{b}_f) + \boldsymbol{b}_o. \tag{3.5}$$

ここで $\boldsymbol{A}_1$，‘$\boldsymbol{A}_f$，及び $\boldsymbol{A}_o$ は重み行列，$\boldsymbol{b}_1$，$\boldsymbol{b}_f$，及び $\boldsymbol{b}_o$ はバイアスベクトルである．$\mathcal{C}_2(\cdot)$ 及び $\mathcal{C}_3(\cdot)$ はそれぞれ Three-Frame Context Encoder 1 及び 2 中の畳み込み層の演算，$\mathcal{G}(\cdot)$ は GRU セルによる状態更新演算を表す．$\phi(\cdot)$ は ReLU を表し，$\boldsymbol{f}_t$ は Frame Encoder の出力特徴量，$\boldsymbol{g}_t$ は Three-Frame Context Encoder 1 の出力特徴量，$\boldsymbol{u}_t$ は Three-Frame Context Encoder 2 の出力特徴量，$\hat{\boldsymbol{y}}_t$ は最終的な叩打音量の推定ベクトルである．

3.4　教師データセットの作成

本研究では，教師あり学習のためのドラムセット音源データセットとして，StemGMD [14] を用いた．StemGMD は，Groove MIDI Dataset [15] を基に構築された大規模なドラム演奏のデータセットであり，ドラムセットを構成する各音源の音が分離された状態で収録されている点が特徴である．このようにドラムセットを構成する各音源のみの信号をステムと呼ぶ．Groove MIDI Dataset は，10 人のドラマーによる演奏データから構成されており，合計 13.6 時間分の musical instrument digital interface (MIDI) データと対応する混合信号が収録されている [15]．StemGMD では，これらの MIDI データを整理し，KD，SD，及び HH を含む 9 種類の基本的なドラム音源に再分類した上で，10 種類のドラム音源ソフトウェアを用いて再合成を行っている [14]．StemGMD では前述の通りドラムセットを構成する音源のステム信号が合計 1224 時間含まれている．

StemGMD は 44.1 kHz の音源であるが，本研究では 16 kHz にダウンサンプリングし，各ドラムパートのステムを用いて教師データを作成した．まず，各ステムから固定長 2.048 s（32768 samples）の短時間信号を非重複で切り出し，KD，SD，及び HH のそれぞれについて音源毎のデータを生成した．切り出す際信号長が 2.048 s 満たない場合は信号の末尾にゼロ埋めを行った．各短時間信号に対して窓長 512，シフトサイズ 256 で時間フレーム化を施し，二乗平均平方根（root mean square: RMS）に基づく音量のラベルを毎時間フレーム付与した．今，時間フレーム t における音源 k の波形ベクトルを $\boldsymbol{s}_{t,k} \in \mathbb{R}^N$（$N = 512$）とすると，RMS

に基づくラベルは次式で計算している．

$$\rho_{t,k} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} |s_{t,k}[n]^2|} \tag{3.6}$$

$$\rho_{t,k}^{(\mathrm{dB})} = 20\log_{10}(\rho_{t,k} + \varepsilon) \tag{3.7}$$

$$y_{t,k} = \begin{cases} \rho_{t,k}^{(\mathrm{dB})} & \left(-60 \leq \rho_{t,k}^{(\mathrm{dB})}\right) \\ -60 & \left(\rho_{t,k}^{(\mathrm{dB})} < -60\right) \end{cases} \tag{3.8}$$

ここで，ε は対数計算の安定化のための微小量であり，本データセット作成では $\varepsilon = 10^{-12}$ を用いた．また，時間信号 $s_{t,k}[n]$ は $\boldsymbol{s}_{t,k}$ の要素であり，すべてのステムにおいて常にダイナミックレンジが範囲 $[-1, 1]$ として定義されていることを想定しており，それゆえ式 (3.8) の最大値は 0 dB である．したがって式 (3.8) として与えられるラベル $y_{t,k}$ の単位は dBFS である．さらに，式 (3.8) ではラベルの下限値を -60 dBFS としており，これを時間区間の音量ラベルは一律 -60 dBFS に固定している．これは，-60 dBFS を下回る波形の音量が人間の聴覚にとってほとんど認知できないレベルであることを考慮して，-60 dBFS 未満の予測精度は不要と考えられるためである．

次に，短時間区間信号の KD，SD，及び HH の各ステムをランダムに抽出し，これらの時間波形を加算することで混合観測信号の短時間区間信号を作成した．このとき，無音に近いステムが使われることを除外するため，すべての t 及び k に関する $y_{t,k}$ の最大値が -40 を超えるもののみを採用した．さらに，各音源の音量比のバリエーションを持たせるために，各音源のステムには区間 $[-12, 0]$ の一様分布から生成したゲイン変化を各音源に付与してから，加算による混合観測信号を作成した．混合観測信号の波形の絶対値の最大値が 0.99 を超える場合は，混合観測信号を再度正規化してクリッピングが生じることを防いだ．この混合観測信号に対する各音源のラベルを改めて次式のように計算して用意した．

$$\tilde{y}_{t,k} = \begin{cases} y_{t,k} + \gamma_k & (-60 \leq y_{t,k} + \gamma_k) \\ -60 & (y_{t,k} + \gamma_k < -60) \end{cases} \tag{3.9}$$

ここで，γ_k は各音源に与える $[-12, 0]$ の一様分布から生成したゲイン値である．以上の混合観測信号及びラベルの作成を行うことで，学習用に 20000 個，検証用に 3000 個，テスト用に 3000 個の短時間区間の入力とラベルを作成した．

3.5　学習の詳細な内容と結果

本節では，提案手法の学習条件と学習結果について述べる．学習・検証・テストには，3.4 節で述べたデータセットを用いた．提案手法の DNN の損失関数及び評価指標には MSE を用いた．時間フレーム t における楽器 k の教師ラベルを $y_{t,k}$，推定値を $\hat{y}_{t,k}$ とすると，MSE は

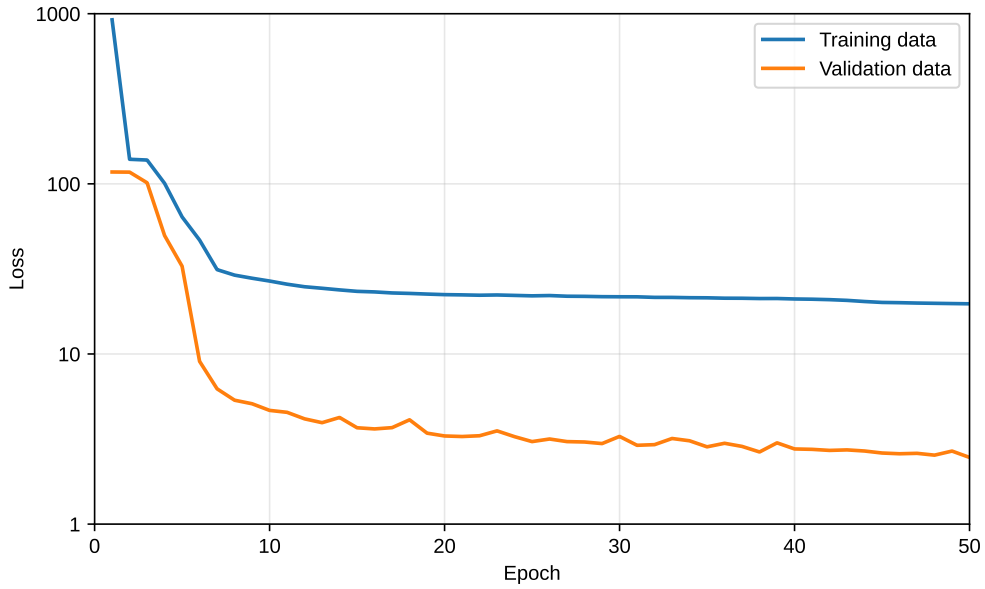


Fig. 3.4 Training and validation curves of MSE.

次式で与えられる．ここで $k \in \{1, 2, 3\}$ であり，T は時間フレーム数である．

$$\mathrm{MSE} = \frac{1}{3T} \sum_{t=1}^{T} \sum_{k=1}^{3} (y_{t,k} - \hat{y}_{t,k})^2 \tag{3.10}$$

最適化手法には Adam を用い，学習率は 1×10^{-3}，バッチサイズは 128，エポック数は 50 とした．Adam は学習の安定性と収束性の観点から広く用いられており，本研究でも安定した学習が得られることを優先して採用した．学習率，バッチサイズ，エポック数は，学習の安定性と収束のバランスを考慮し，過度な発散や過学習を避けるための経験的な設定として決定した．

学習曲線を Fig. 3.4 に示す．縦軸は対数軸である．学習初期（1–6 エポック）では MSE が急減し，訓練時の MSE は 920 から 47，検証時の MSE は 117 から 9 まで低下した．10 エポック時点で検証 MSE は 4.66 となり，その後は 2.5–3.3 付近で小さく変動しながら緩やかに低下している．最終エポック（50 Epoch）では訓練 MSE が約 19.7，検証 MSE が約 2.47 となり，検証誤差の変動は限定的であるため，過度な過学習は確認されなかった．以上より，提案手法のモデルは適切に学習できていることが分かるため，次章で述べる実験ではこの学習済みモデルを用いた叩打音量の推定精度の客観精度について述べる．

3.6　本章のまとめ

本章では，混合ドラム音から叩打音量を直接推定する提案手法を示した．入力信号と出力叩打音量の対応関係を整理し，時間フレーム単位の逐次推定を前提とした枠組みを明確にした．

ネットワーク構造は時間フレーム内畳み込み，因果畳み込み，GRU，FC 層から構成し，短期的な変化と長期的な文脈の双方を扱える設計とした．また，StemGMD を 16 kHz にダウンサンプルして短時間信号を作成し，RMS に基づく dBFS ラベルを付与することで教師データセットを整備した．学習条件と評価指標を整理し，提案法がどの前提で学習・評価されるかを明確にした．以上により，次章では本章で述べた提案手法と SNMF に基づく従来手法を比較し，その精度について評価を行う．

第 4 章

叩打音量推定実験

4.1 まえがき

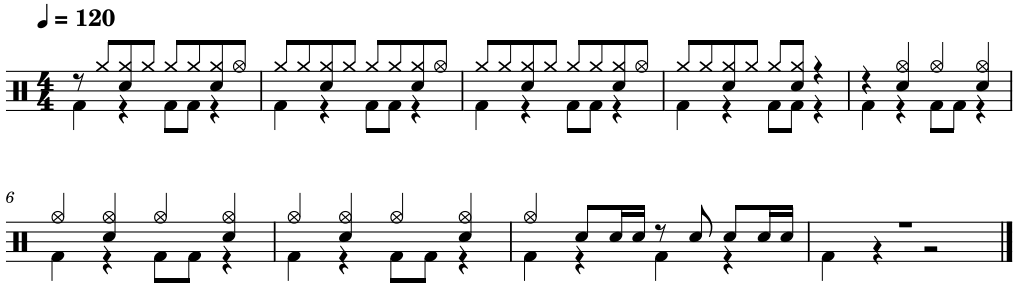
本章では，提案手法と SNMF に基づく叩打音量の推定精度を同一条件で比較し，検証する．4.2 節では，使用データと実験条件を示し，時間フレーム設定の違いに対する時間対応付けやスケール校正の手順を整理する．4.3 節では，MSE に基づく定量評価と時間推移の結果を示し，音源毎の差や手法間の傾向を考察する．4.4 節では，本章をまとめる．なお，リアルタイム性に関する実測遅延や実際のシステムとして運用した際の客観評価については本章の対象外とし，本章では推定精度の比較のみに焦点を当てる．

4.2 実験条件

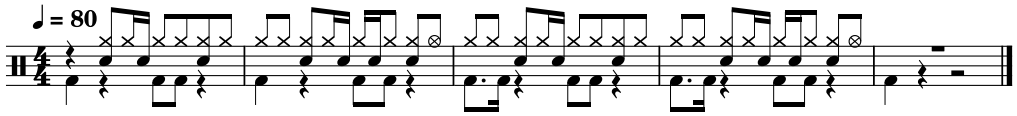
評価実験には，実際のドラムセットの音を収録したデータセット [16, 17] に含まれている音響信号を用いた．このデータセットに含まれる演奏リズムパターンとして，Fig. 4.1 に示す Pattern 1，Pattern 2，及び Pattern 3 を抽出し，本章での実験に用いた．ここで，一般的なドラム譜の凡例を Fig. 4.2 に示す．

本章では，DNN に基づく提案手法と SNMF に基づく従来手法で時間フレームの設定が異なるため，それぞれの処理条件を明記する．提案手法（DNN）は，混合信号をサンプリング周波数 $F_{\mathrm{sampling}} = 16$ kHz で扱い，窓長 $\Psi_{\mathrm{DNN}} = 512$，ホップ長 $\Gamma_{\mathrm{DNN}} = 256$ で時間フレーム化する．各時間フレームに対して KD，SD，及び HH の叩打音量を dBFS で推定し，出力系列の長さを T_{D} とする．

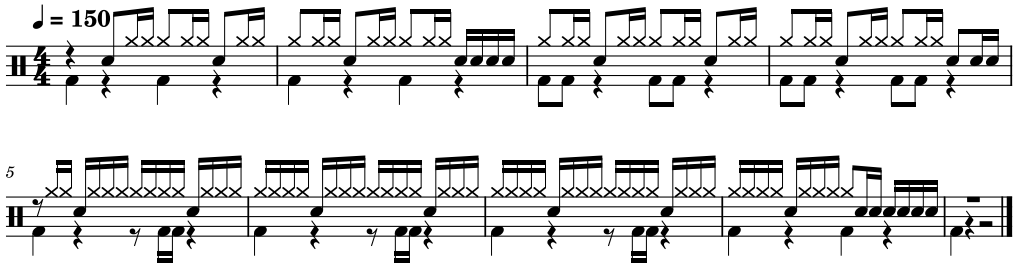
次に，SNMF による推定手順について述べる．SNMF は窓長 $\Psi_{\mathrm{SNMF}} = 1024$，ホップ長 $\Gamma_{\mathrm{SNMF}} = 256$ で信号を区切り，各時間フレーム毎に DFT を計算して得られる 1 本の周波数ベクトルを入力とする．なお，本実験では SNMF の学習ステージ（教師基底 $\mathbf{W}^{(\mathrm{KD})}$, $\mathbf{W}^{(\mathrm{SD})}$, 及び $\mathbf{W}^{(\mathrm{HH})}$ の学習）に用いるサンプル音のスペクトログラム $\mathbf{X}^{(\mathrm{KD})}$，$\mathbf{X}^{(\mathrm{SD})}$，及び $\mathbf{X}^{(\mathrm{HH})}$ として，混合する前の各音源の時間信号のスペクトログラムを用いた．これは実際には手に入れることができない混合前の各音源のデータを使用した事前学習であり，従来手法の SNMF にとっ



(a) Pattern 1



(b) Pattern 2



(c) Pattern 3

Fig. 4.1 Drum scores of Pattern 1, Pattern 2, and Pattern 3.

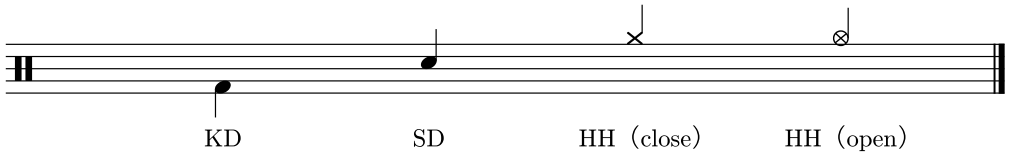


Fig. 4.2 Legend of standard drum notation.

て理想的な条件となっている．学習された教師基底 $\mathbf{W}^{(\mathrm{KD})}$, $\mathbf{W}^{(\mathrm{SD})}$, 及び $\mathbf{W}^{(\mathrm{HH})}$ は全て 1 個の列ベクトルから成る．従って，これをまとめた教師基底行列 $\mathbf{W} = \mathbf{W}^{(\mathrm{KD})}\ \mathbf{W}^{(\mathrm{SD})}\ \mathbf{W}^{(\mathrm{HH})}$ は 3 本の基底ベクトルをもち，その各々が KD，SD，及び HH の各音源のスペクトル教師に対応する．ベクトル $\mathbf{x}_{\bar{t}}$ に対して次式を近似する非負係数ベクトル $\mathbf{h}_{\bar{t}} = [h_{\bar{t},1}, h_{\bar{t},2}, h_{\bar{t},3}]^{\mathrm{T}}$ を求める．

$$\mathbf{x}_{\bar{t}} \approx \mathbf{W}\mathbf{h}_{\bar{t}} \tag{4.1}$$

推定された $\mathbf{h}_{\bar{t}}$ の各成分から各音源の RMS を計算し，dBFS へ変換して音量系列を得る．

両手法は窓長が異なるため，同一時刻の比較では時間フレーム番号を対応付ける必要があ

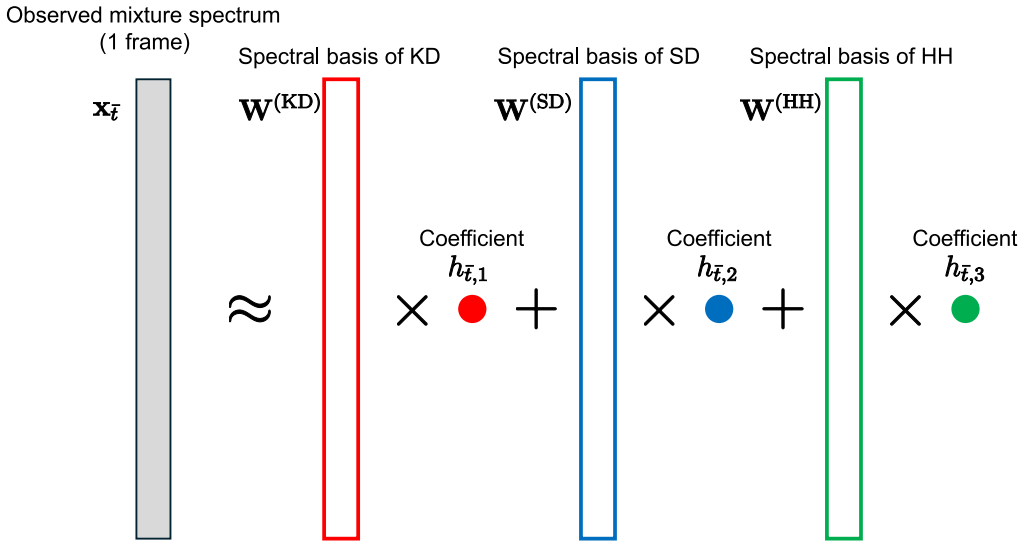


Fig. 4.3 SNMF-based gain estimation procedure used in the experiments.

る．まず，SNMF の時間フレーム $\bar{t}$ の時刻を次式で定義する．

$$\tau_{\bar{t}} = \bar{t} \times \frac{\Gamma_{\mathrm{SNMF}}}{F_{\mathrm{sampling}}} \tag{4.2}$$

次に，DNN 側ではホップ長 Γ_{DNN} の時間フレーム系列のうち最も近い時間フレームを次式で対応させる．

$$t = \mathrm{round}\left(\frac{\tau_{\bar{t}} F_{\mathrm{sampling}}}{\Gamma_{\mathrm{DNN}}}\right) \tag{4.3}$$

ここで，round(·) は整数への丸め（四捨五入）を表す．また，SNMF の出力は RMS に基づく値であり，dBFS とはスケールが異なるため，各短時間音響信号及び各時間のそれぞれにおいて正解値 $y_{t,k}$ に最も近づく一次変換を施すことで校正した．SNMF の RMS 値を $\hat{y}_{t,k}^{(\mathrm{SNMF})}$ と表すと，この最適一次変換による校正は次式となる．

$$\hat{y}_{t,k} = \alpha_k \hat{y}_{t,k}^{(\mathrm{SNMF})} + \beta_k \tag{4.4}$$

係数 α_k 及び $beta_k$ は，正解値 $y_{t,k}$ との二乗誤差が最小となるように次の最小二乗法の解として与えた．

$$(\alpha_k, \beta_k) = \arg\min_{a,b} \sum_{t=1}^{T_{\mathrm{D}}} \left(a\hat{y}_{t,k}^{(\mathrm{SNMF})} + b - y_{t,k}\right)^2 \tag{4.5}$$

4.3　実験結果

推定精度の客観評価尺度には，次式で示す dBFS スケールでの MSE を用いた

$$\mathrm{MSE}_k = \frac{1}{T}\sum_{t=1}^{T} (y_{t,k} - \hat{y}_{t,k})^2 \tag{4.6}$$

Table 4.1 MSE comparison among all test data

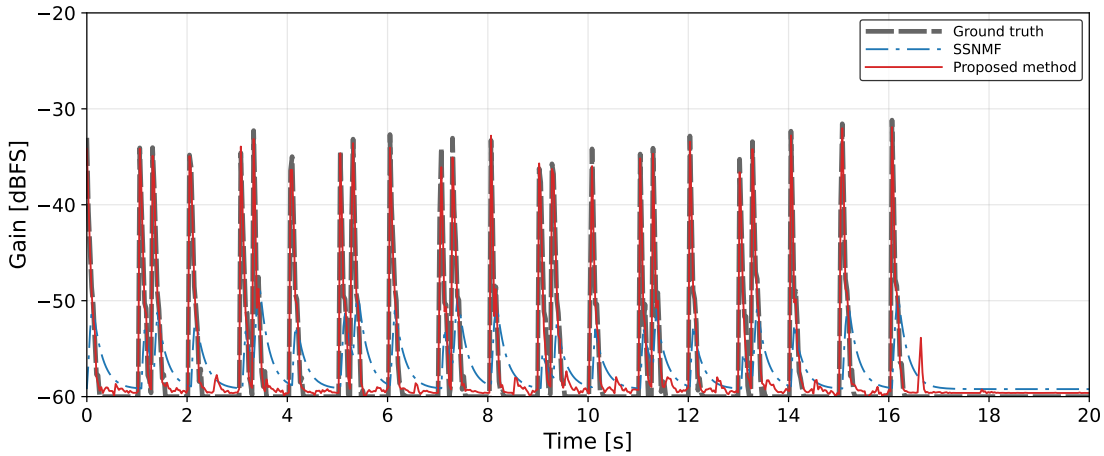
Source	SNMF	DNN
KD	23.618	**0.328**
SD	35.904	**0.740**
HH	14.234	**5.573**
Average	24.586	**2.214**

Pattern 1 から Pattern 3 の全区間をすべて含めた MSE の値を手法毎に Table 4.1 に示す．提案手法である DNN の推定結果は全音源で従来の SNMF に基づく推定結果よりも低い誤差を示し，平均 MSE は約 2.21 であった．一方，SNMF は平均約 24.59 であり，提案法が大幅に誤差を低減できていることが分かる．音源毎の差としては，DNN は HH の誤差が最も大きく，KD が最も小さい．一方，SNMF は SD の誤差が最も大きい傾向が確認できる．

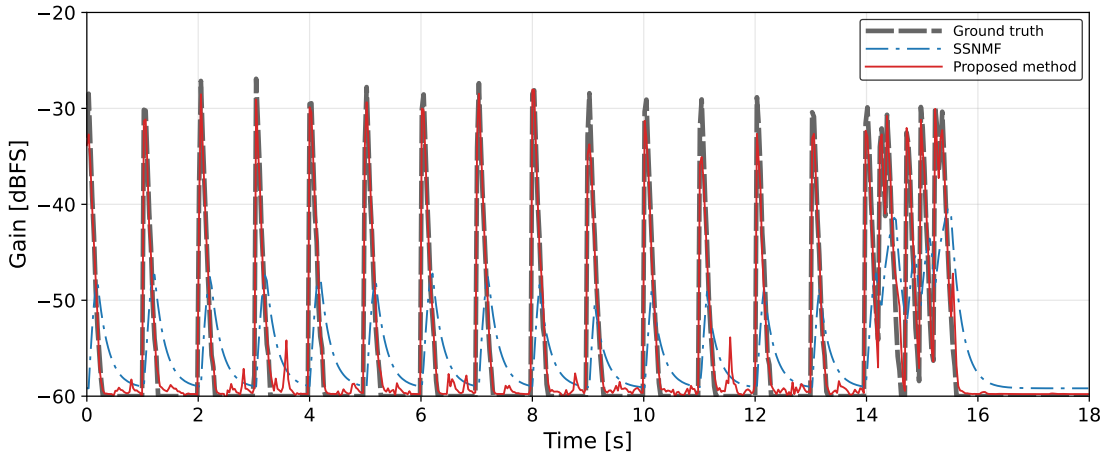
Fig. 4.4，4.5，及び 4.6 に，Pattern 1 から Pattern 3 の推移図を示す．時間フレーム長の違いとモデル構造を踏まえると，SNMF は窓長が長いため，叩打が密な時間区間では推定結果が平滑化されやすく，また同時に複数の音源を叩打した場合に音源間の干渉を起こして誤差が生じやすいことが結果から確認できる．しかしながら，単独音源の叩打では音色変化が小さい区間で事前学習した基底ベクトルとの一致が高く，SNMF の推定が安定しやすい傾向にある．一方で，提案手法においては，KD 及び SD ではの立ち上がり付近で DNN が正解に近い推定を達成する区間が多く，短い時間フレームと文脈利用の効果が現れているように見える．この傾向は，KD 及び SD での MSE 低減として現れている．HH については，DNN でも相対的に誤差が大きくなりがちであり，推定が難しいことが分かる．HH は広帯域のノイズ成分と長い減衰を持ち，SD と周波数帯が重なりやすいことから，推定精度の劣化につながったものと予想される．なお，HH は打撃間隔が短く変化が急な時間区間で両手法ともに推定のずれが見られる．SNMF は定常的なスペクトル形状に近い時間区間では比較的安定な推定が得られる一方で，変化が大きい時間区間では事前学習した基底ベクトルの不一致が顕在化しやすい．また，DNN は立ち上がりの変化には追従しやすいが，広帯域の残響成分が長く続く区間では推定が難しい傾向が残る．

4.4 本章のまとめ

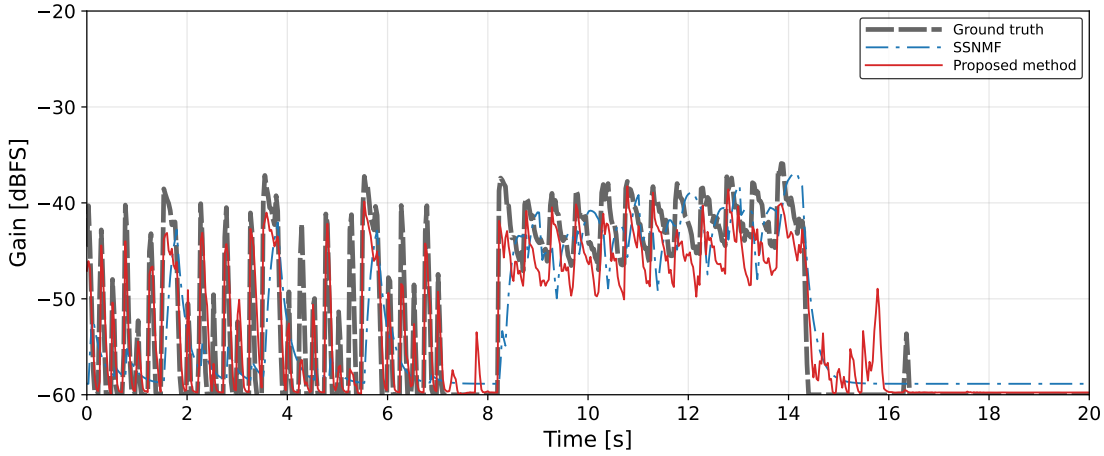
本章では，DNN に基づく提案手法と SNMF に基づく従来手法を同一条件で比較し，叩打音量推定の精度を検証した．ドラム演奏データに対して両手法の時間フレーム設定の違いを時間対応付けし，SNMF の出力については線形校正を行ったうえで dBFS スケールの MSE で評価した．その結果，提案法は全体として従来手法より小さい誤差を示し，特に KD と SD で改善が顕著であった．一方，HH は広帯域成分と残響の影響を受けやすく，両手法で誤差が残る傾向が見られた．以上より，提案手法は従来手法に比べて叩打音量推定の精度向上に有効であると結論づけられる．



(a) KD

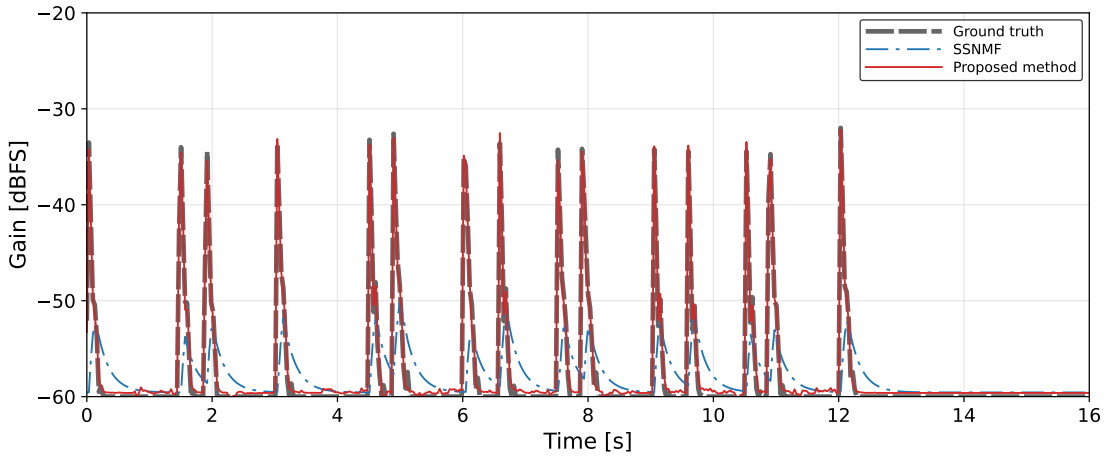


(b) SD

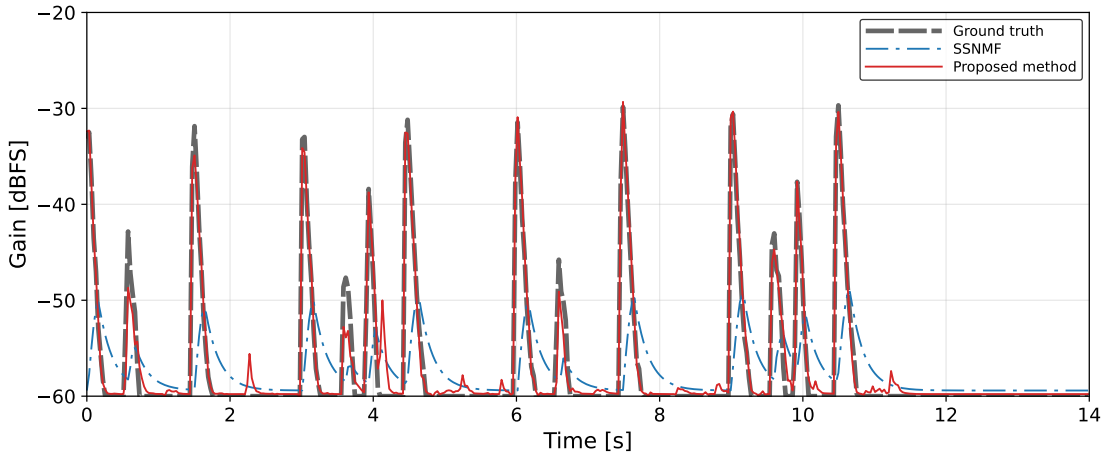


(c) HH

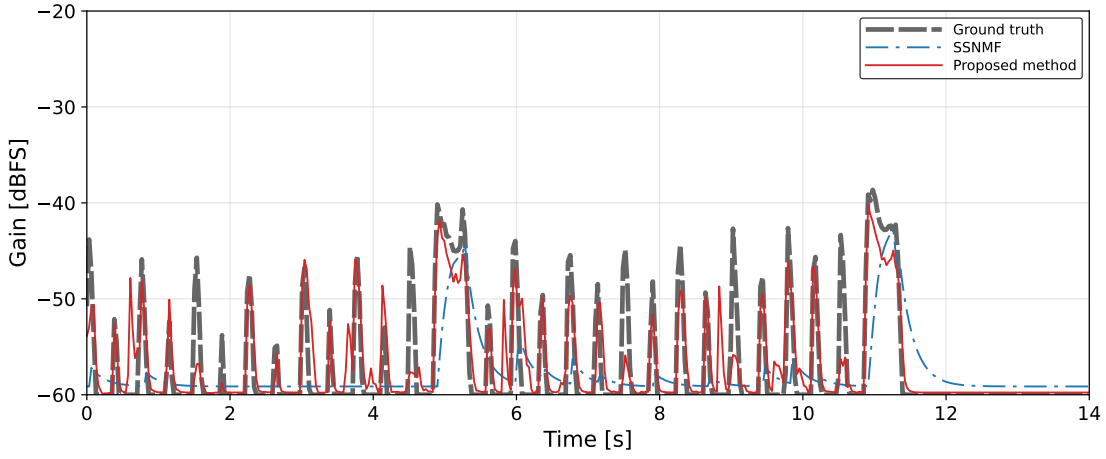
Fig. 4.4 Comparison of gain estimation for Pattern 1.



(a) KD

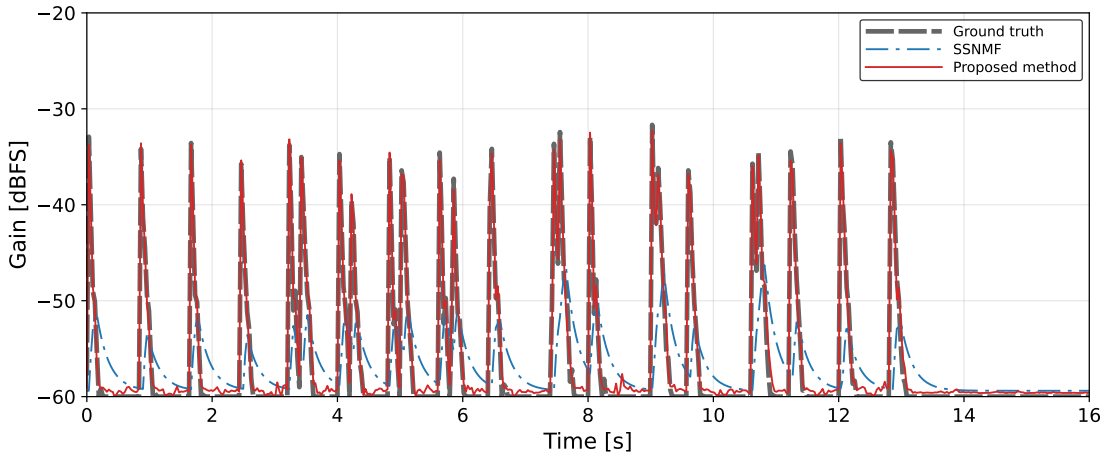


(b) SD

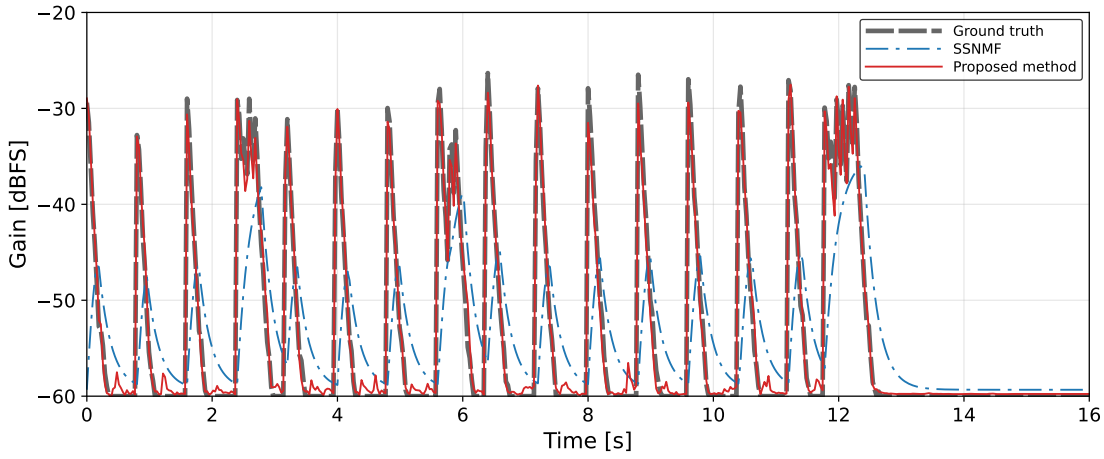


(c) HH

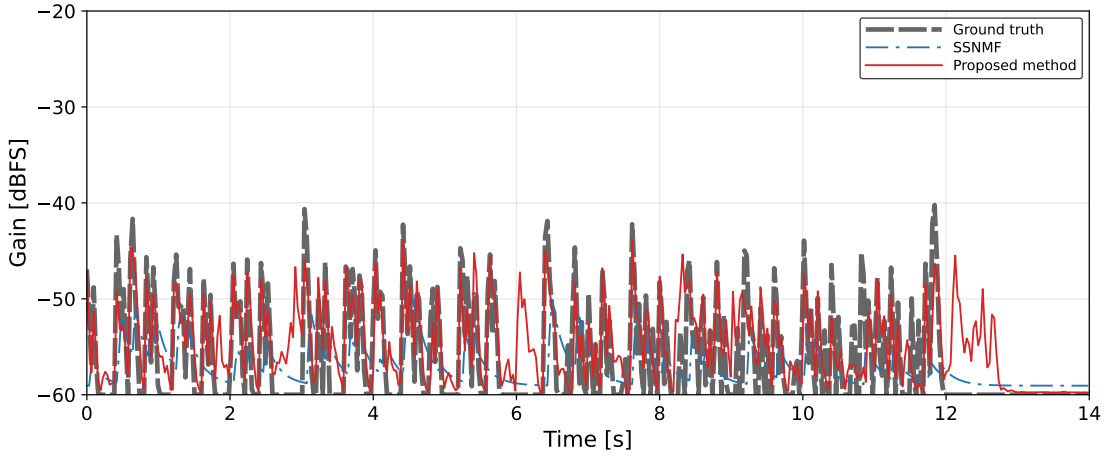
Fig. 4.5 Comparison of gain estimation for Pattern 2.



(a) KD



(b) SD



(c) HH

Fig. 4.6 Comparison of gain estimation for Pattern 3.

第 5 章

結言

本研究では，1 本のマイクロホンで収録した混合ドラム音から KD，SD，HH の叩打音量を時間フレーム毎に推定し，演奏者に可視化フィードバックする枠組みを検討した．従来の SNMF に基づく段階処理に対し，提案法では DNN 回帰によって分離と推定を一括化し，事前サンプルに依存しない推定の枠組みを示した．

提案法は時間フレーム単位の逐次処理を前提とし，時間フレーム内畳み込み，因果畳み込み，GRU，FC 層からなるネットワークで短期・長期の文脈を統合して叩打音量を推定する．学習には StemGMD を 16 kHz にダウンサンプルして作成した教師データセットを用い，RMS に基づく dBFS ラベルで学習を行った．

実験では，提案法と従来手法を同一条件で比較し，dBFS スケールの MSE で評価した．Pattern 1, Pattern 2, 及び Pattern 3 の全区間を統合した結果，提案法の平均 MSE は 2.214, SNMF は 24.586 であり，提案法が大幅に誤差を低減できることを確認した．楽器別では提案法は KD と SD で改善が顕著である一方，HH は広帯域成分や残響の影響を受けやすく，誤差が相対的に大きい傾向が残った．

以上より，提案法は従来法に比べて叩打音量推定の精度向上に有効であると結論づけられる．一方で，本システムは音量バランスの良し悪しを自動判定するものではなく，可視化された音量関係を演奏者が参照して調整するための支援を目的とする．また，dBFS は周波数ごとの知覚差を直接反映する指標ではないため，実環境での頑健性や実測遅延の評価と合わせて，知覚特性を考慮した提示方法の検討が今後の課題である．今後は，環境変動への耐性評価，対象楽器の拡張，学習データの多様化，及び可視化提示の改善を進めることで，実運用に近い条件での有効性を検証していく．

謝辞

本研究を進めるにあたり，終始懇切丁寧なご指導と多くのご助言を賜りました北村大地准教授に深く感謝いたします．研究の方向性に関する助言に加え，実験設計や論文執筆，発表資料の構成に至るまで細部にわたりご指導いただき，本研究をまとめ上げることができました．

また，副査として貴重なご指摘とご助言を賜りました重田和弘教授に心より感謝いたします．

さらに，北村研究室の先輩である加藤大輝氏，鈴木慶氏，和気佑弥氏，小川遼氏，谷野宮蒼士氏には，研究を進める上で有益な助言や議論をいただき，大変お世話になりました．特に，メンターとしてご指導いただいた和気佑弥氏には，論文および発表資料に関して具体的な改善案を数多くいただき，本研究の完成度向上に大きく寄与していただきました．

北村研究室に留学できていたエリョン ベカヴァ氏には先行研究を進めていただきました．また，研究活動で海外の方とかかわる貴重な経験をさせていただき感謝しています．

研究室同期の片山碧人氏，森末結氏には，日々の議論や情報共有を通じて多くの刺激を受けました．互いに支え合いながら研究を進められたことに感謝いたします．また，皆様がいてくれたからこそ，研究活動を乗り越えることができました．

また，クラスメイトの有岡優平氏，齊藤壮志氏，土井大地氏には，実験やテスト勉強，日々の学校生活において多くの助けをいただきました．皆様がいてくれたからこそ，この 5 年間を乗り越えることができました．ここに感謝の意を表します．

最後に，20 年間にわたり私を育て，金銭的にも精神的にも支えてくれ，いつでも温かい食事を用意してくれた両親に，心から感謝いたします．常に見守り続けてくれたおかげで，安心して学業と研究に取り組むことができました．感謝の念に堪えません．

参考文献

[1] 細谷 美月, 中村 聡史, 森勢 将雅, 吉井 和佳, “ドラム演奏の音量バランス習得に向けた音源分離を用いたリアルタイム叩打音量可視化システムの提案,” **情報処理学会研究報告**, vol. 2021-MUS-130, no. 27, pp. 1–8, 2021.

[2] ユルゲン・メイヤー, **ホールの響きと音楽演奏**, 市ケ谷出版社, 2015, 日高孝之 訳.

[3] N. H. Fletcher, T. D. Rossing, **楽器の物理学**, Springer-Verlag Tokyo, 2002.

[4] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

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

[6] P. Smaragdis, B. Raj, and M. Shashanka, “Supervised and semi-supervised separation of sounds from single-channel mixtures,” *in Proc. Int. Conf. Latent Variable Anal. Signal Sep. (LVA/ICA)*, pp. 414–421, 2007.

[7] D. Kitamura, H. Saruwatari, K. Yagi, K. Shikano, Y. Takahashi, and K. Kondo, “Music signal separation based on supervised nonnegative matrix factorization with orthogonality and maximum-divergence penalties,” *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E97-A, no. 5, pp. 1113–1118, 2014.

[8] Y. Iwase and D. Kitamura, “Supervised audio source separation based on nonnegative matrix factorization with cosine similarity penalty,” *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E105-A, no. 6, pp. 906–913, 2022.

[9] E. M. Grais, M. U. Sen, and H. Erdogan, “Deep neural networks for single-channel source separation,” *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3734–3738, 2014.

[10] S. Uhlich, F. Giron, and Y. Mitsufuji, “Deep neural network based instrument extraction from music,” *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 2135–2139, 2015.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder–decoder for

statistical machine translation," *in Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, pp. 1724–1734, 2014.

[13] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

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

[15] J. Gillick, A. Roberts, J. Engel, D. Eck, and D. Bamman, "Learning to groove with inverse sequence transformations," *in Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 2269–2279, 2019.

[16] 森末 結, 北村 大地, "マイクロホンアレイを用いたドラムセット音源分離のデータセット収録・公開," *in Proc.* **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, 京都, 2025 年 12 月.

[17] D. Kitamura, "Dataset for drums source separation using microphone arrays," Zenodo, 2025. https://doi.org/10.5281/zenodo.17706651.