# 微分可能なコンプレッサーのパラメータ推定に関する検討*

☆王檬（農工大），中村友彦（産総研），山田宏樹，矢田部浩平（農工大）

## 1　まえがき

コンプレッサーは，音響信号の振幅を圧縮する重要なサウンドエフェクトである。しかし，望みの音を得るためには，パラメータへの深い理解が必要であり，その調整は難しい。音響信号からコンプレッサーのパラメータが推定できれば，所望の音を得やすくなる。深層ニューラルネットワーク (DNN) を用いて音響信号からパラメータを回帰する方法が提案されているが，推定精度と解釈性の両面で低い性能を示していた [1]。近年，デジタル信号処理の機構を微分可能なモジュールとして実装し，DNN と組み合わせて用いる手法（DDSP）[2] に基づく微分可能なコンプレッサー [3] が提案され，有望な性能を示している。そこで本稿では，文献 [3] で提案された音響信号間のスタイル転移手法を教師ありのパラメータ推定問題に援用し，その推定性能について実験的に調査した。

## 2　微分可能なコンプレッサー

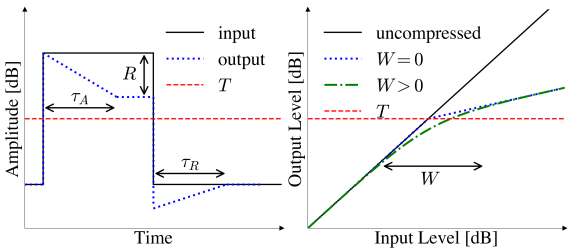
コンプレッサーの主な機能は，入力信号のダイナミックレンジを制御して，最大振幅と最小振幅の差を小さくすることである。典型的なデジタルコンプレッサーは，スレッショルド $T$ [dB] ，レシオ $R$，ニーの幅 $W$ [dB]，アタックタイム $\tau_A$ [ms]，リリースタイム $\tau_R$ [ms]，メイクアップゲイン $M$ [dB] の 6 つのパラメータで信号を制御する。$M$ 以外のパラメータの信号への寄与を図–1 に示す。コンプレッサーの動作はゲインの計算，レベルの検出，ゲインの調整の 3 つの段階に分かれている [4]。

ゲインの計算では，圧縮信号の振幅 $y_G$ [dB] を入力信号の振幅 $x_G$ [dB] から以下のように得る。

$$y_G = \begin{cases} x_G & (2(x_G - T) < -W) \\ x_G + \frac{(\frac{1}{R}-1)(x_G - T + \frac{W}{2})^2}{2W} & (2|x_G - T| \le W) \\ T + \frac{x_G - T}{R} & (2(x_G - T) > W) \end{cases} \tag{1}$$

$T$ を超える $x_G$ が $R$ に従って圧縮され，圧縮と非圧縮部分の変化カーブは $W$ によって滑らかになる。しかし，$y_G$ を直接出力すると，急激なゲインの変化により出力信号が不自然になるため，ゲイン低減量 $x_L = x_G - y_G$ の反応速度を緩やかにする処理が必要である。

レベルの検出では，より緩やかなゲイン低減量




図–1　コンプレッサーの処理過程。左図は，スレッショルド $T$，レシオ $R$，アタックタイム $\tau_A$，リリースタイム $\tau_R$ の処理を示す。右図は，ニーの幅 $W$ がどのようにゲインの変化を滑らかにするかを示す。

$y_L$ [dB] を $x_L$ [dB] から以下のように得る。

$$y_L[n] = \begin{cases} \alpha_A y_L[n-1] + (1-\alpha_A) x_L[n] & (x_L[n] > y_L[n-1]) \\ \alpha_R y_L[n-1] + (1-\alpha_R) x_L[n] & (x_L[n] \le y_L[n-1]) \end{cases} \tag{2}$$

ただし，係数 $\alpha_A$ と $\alpha_R$ はサンプリング周波数を $f_s$ として，$\tau_A$ と $\tau_R$ を用いて $\alpha = e^{-1/\tau f_s}$ で定義される。

最後のゲインの調整では，$x_G$ から $y_L$ を引くことによる信号レベル低下を $M$ によって補い，最終振幅 $y$ [dB] を以下のように出力する。

$$y = x_G - y_L + M \tag{3}$$

微分可能なコンプレッサーでは，パラメータと信号に関して出力からの勾配を計算できるように上記の過程が実装されている。ただし，レベル検出段階の $\tau_A$ と $\tau_R$ の再帰的な計算を IIR フィルタで近似する [3]。

$$y_L[n] = \alpha y_L[n-1] + (1-\alpha) x_L[n] \tag{4}$$

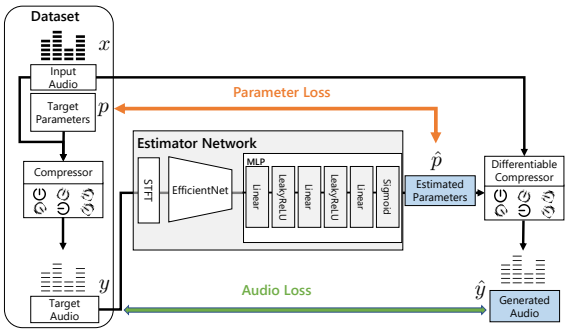
$\tau_A$ と $\tau_R$ は同一の値 $\tau$ に設定されているが，よい近似となることが実験的に確認されている [3]。

## 3　提案手法

文献 [3] の微分可能なコンプレッサーは，音響信号に所望のエフェクトをかけることを目的としており，自己教師あり学習モデルとして提案されている。本稿は，文献 [3] で提案された手法を教師ありの問題設定に拡張し，そのパラメータの推定精度を調査する。モデルの概略を図–2 に示す。

教師信号 $y$ は，入力信号 $x$ とランダムに生成された教師パラメータセット $p$ を用いてコンプレッサーによって生成される。$y$ を短時間フーリエ変換 (STFT)

*Study on parameter estimation of differentiable compressor. By Meng WANG (Tokyo University of Agriculture and Technology (TUAT)), Tomohiko NAKAMURA (National Institute of Advanced Industrial Science and Technology (AIST)), Koki YAMADA and Kohei YATABE (TUAT)




図–2　実験に用いたパラメータ推定モデルの構造

表–1　実験に用いたパラメータの生成範囲

| | $T$ [dB] | $R$ | $\tau$ [ms] | $W$ [dB] |
|---|---|---|---|---|
| 生成範囲 | [-50.0, -25.0] | [1.0, 20.0] | [0.1, 100.0] | [0.0, 12.0] |

でスペクトログラムに変換してネットワークに入力する。パラメータ推定ネットワークの構造は参考モデルと同一で，事前学習済みの EfficientNet B2 CNN [5] とその出力からコンプレッサーのパラメータセットを推定する多層パーセプトロン (MLP) からなる。最後に，推定されたパラメータセット $\hat{p}$ は $x$ と共に微分可能なコンプレッサーに渡され，推定信号 $\hat{y}$ が生成される。

提案手法のロス関数は，文献 [3] で用いられた推定信号のロス関数にパラメータのロス関数を追加する。

$$\mathcal{L}_{\text{overall}} = w_\alpha(\mathcal{L}_{\text{freq}} + w_\beta \mathcal{L}_{\text{time}}) + \mathcal{L}_{\text{param}} \qquad (5)$$

パラメータのロス関数 $\mathcal{L}_{\text{param}}$ は，$\hat{p}$ と $p$ の平均二乗誤差 (MSE) を適用した。信号のロス関数は，$\hat{y}$ と $y$ の時間領域の平均絶対誤差 (MAE) $\mathcal{L}_{\text{time}}$ と周波数領域の多重解像度 STFT (MR-STFT)[3] の誤差 $\mathcal{L}_{\text{freq}}$ の和を用いた。ここで，$w_\alpha = 0.1$，$w_\beta = 100$ は各ロスのバランスを取るための重み係数である。

## 4　実験

学習データ，検証データ，テストデータの入力信号は，24 kHz の LibriTTS-R の音声データセット [6] から，事前に重複しないようにランダムに選び，約 10 秒に切り出した。推定問題をより簡単にするため，パラメータ $M$ の推定は行わず既知とした。パラメータの生成範囲は表–1 のように，$T$ は信号の最大振幅よりも常に小さくなるように制限した。STFT には，窓長が 4096 サンプルの Hann 窓を用い，シフト幅は 2048 サンプルとした。

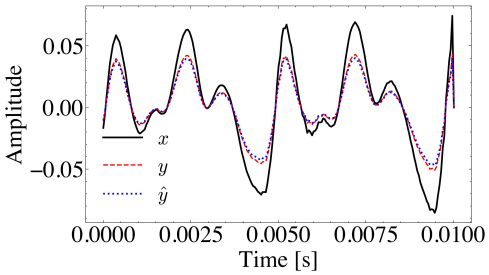
エポック数は 500，バッチサイズは 6 とした。各エポックで，学習データ数 10000 個と検証データ数 100 個を逐次生成した。初期学習率を $10^{-3}$ として Adam を用いて最適化した。学習率は，最大反復回数の 64%，76%，80%と 88%のときに 1/10 に減少するようにスケジュールした。

表–2　ランダムに生成された 100 回分のパラメータを推定したときの $p$ と $\hat{p}$ の平均絶対誤差

| | $T$ [dB] | $R$ | $\tau$ [ms] | $W$ [dB] |
|---|---|---|---|---|
| 100 回平均絶対誤差 | 1.54 | 4.05 | 9.4 | 2.94 |

表–3　パラメータの推定結果例

| | $T$ [dB] | $R$ | $\tau$ [ms] | $W$ [dB] |
|---|---|---|---|---|
| 教師パラメータ $p$ | -40.0 | 8.3 | 47.9 | 11.8 |
| 推定パラメータ $\hat{p}$ | -39.4 | 11.2 | 41.2 | 5.8 |




図–3　表–3 のパラメータでの入力，教師，推定信号

テストデータの推定結果を表–2 に示す。$T$ と $\tau$ のパラメータ推定精度は高いが，$R$ と $W$ の誤差が大きいことが見て取れる。表–3 に示す通り，$R$ と $W$ に大きな誤差がある。それにも関わらず，図–3 は推定信号 $\hat{y}$ が教師信号 $y$ に十分に近いことを示す。

コンプレッサーは出力信号の振幅を変化させるため，振幅への影響が小さいパラメータは推定が難しいと考えられる。$W$ は全ての範囲で出力信号への影響が小さく，推定が難しい。$R$ は概ね 8 以下は違いが現れやすいが，8 を超えると出力信号に違いが現れにくいため，推定が難しくなると考えられる。

## 5　むすび

本稿では，微分可能なコンプレッサーのパラメータ推定性能を調査した。今後は異なる範囲のパラメータ推定性能の比較と推定精度の向上に取り組む。

### 参考文献


[1] S. Hawley, B. Colburn, and S. I. Mimilakis, "Profiling audio compressors with deep neural networks," in *147th Audio Eng. Soc. Conv.*, (2019).

[2] J. Engel, L. H. Hantrakul, C. Gu, A. Roberts, "DDSP: Differentiable digital signal processing," in *Int. Conf. Learn. Represent. (ICLR)*, (2020).

[3] C. J. Steinmetz, N. J. Bryan, and J. D. Reiss, "Style Transfer of Audio Effects with Differentiable Signal Processing," *J. Audio Eng. Soc.*, **70**, 708–721 (2022).

[4] D. Giannoulis, M. Massberg, and J. D. Reiss, "Digital dynamic range compressor design—A tutorial and analysis," *J. Audio Eng. Soc.*, **60**, 399–408 (2012).

[5] M. Tan, Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Mach. Learn. (ICML)*, pp. 6105–6114 (2019).

[6] K. Yuma, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna. "LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus," in *Interspeech*, (2023).