微分可能なコンプレッサーの深層展開に基づくパラメータ推定と信号復元*

☆王檬，赤石夏輝，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，音響信号の振幅が大きい部分を圧縮することで音量を均一化するエフェクタである。コンプレッサーを用いることで信号の全体的な音量を均一にできる一方で，信号に歪みを生じさせる。この歪みが後続の処理に悪影響を及ぼす場合があるので，コンプレッサー処理後の信号から元信号を復元する手法が望まれる。しかし，コンプレッサーは複雑な非線形システムであるため，元信号の復元は難しい。

それに対して，我々はこれまで，微分可能な信号処理を用いて，コンプレッサー処理された信号からコンプレッサーのパラメータを推定する手法を提案した [1]。さらに，非線形システムによって処理された信号を変分不等式に基づいて復元する手法をコンプレッサーに適用し，元信号を復元できることを確認した [2]。しかし，推定されたパラメータを用いて信号を復元しようとすると，パラメータの推定誤差が復元結果に影響を及ぼしてしまう。そこで本稿では，信号の復元とパラメータの推定を一貫させるために，これらを組み合わせた深層展開に基づく手法を提案する。

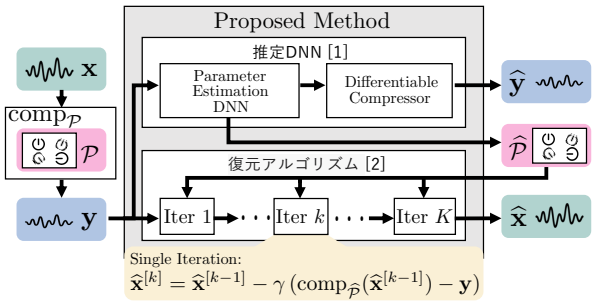
2　これまでのコンプレッサーの信号復元

我々は以前，未知のコンプレッサーで処理された信号からパラメータを推定する手法を提案した [1]。この手法は，微分可能な信号処理に基づくコンプレッサーと深層ニューラルネットワーク（DNN）をまとめて学習したモデルでスレッショルド T，レシオ R，アタックタイム τ，ニー W を推定する。本稿では，この手法を推定 DNN と呼ぶ。

また，コンプレッサー処理された信号から元信号を復元する手法を提案した [2]。真のパラメータの組 $\mathcal{P}$ に近い推定パラメータの組を $\widehat{\mathcal{P}}$，コンプレッサー処理を行う関数を $\mathrm{comp}_{\mathcal{P}}$，元信号を $\mathbf{x}$ とする。コンプレッサー処理された信号は $\mathbf{y} = \mathrm{comp}_{\mathcal{P}}(\mathbf{x})$ で表せる。このとき，変分不等式に基づく反復更新

$$\widehat{\mathbf{x}}^{[k]} = \widehat{\mathbf{x}}^{[k-1]} - \gamma\left(\mathrm{comp}_{\widehat{\mathcal{P}}}(\widehat{\mathbf{x}}^{[k-1]}) - \mathbf{y}\right) \qquad (1)$$

によって，元信号の推定信号 $\widehat{\mathbf{x}}$ が得られる。ただし，k は 1 から最大反復回数 K までの値であり，$\gamma \in (0, 2)$ はステップサイズである。この復元アルゴリズムは，$\mathbf{y}$ に極端な歪みがなく，真のパラメータの組 $\mathcal{P}$ が与えられた場合に元信号を精度良く復元できる [2]。



図–1　提案する深層展開に基づく信号復元モデルの構造。同じ色で囲まれた入出力同士で損失を計算する。

推定 DNN で得られたパラメータを復元アルゴリズムで用いることで，コンプレッサー処理された信号から元信号の復元ができると考えられる。しかし，推定 DNN で得られたパラメータは真の値との誤差が大きく，そのままでは復元アルゴリズムによって精度良く復元ができない。そこで，パラメータ推定と信号復元を一貫させ，復元アルゴリズムを考慮した推定 DNN を学習できれば，復元性能の向上が期待できる。

3　提案手法

本稿では，パラメータ推定と信号復元を一貫させるために，推定 DNN と復元アルゴリズムを組み合わせた深層展開に基づく手法を提案する。深層展開は，反復アルゴリズムを DNN とみなして学習する方法である [4]。提案手法では，パラメータの推定と信号復元の学習を end-to-end で行う。これによって，復元アルゴリズムの性質を反映させたパラメータの推定を行うことができ，そのパラメータを用いることで元信号の復元ができると考えられる。

提案するモデルの構造を図–1 に示す。まず，$\mathbf{y} = \mathrm{comp}_{\mathcal{P}}(\mathbf{x})$ で生成された信号 $\mathbf{y}$ を推定 DNN に入力すると，推定パラメータの組 $\widehat{\mathcal{P}}$ が得られる。また，$\widehat{\mathbf{y}} = \mathrm{comp}_{\widehat{\mathcal{P}}}(\mathbf{x})$ によってコンプレッサー処理された信号の推定信号 $\widehat{\mathbf{y}}$ が得られる。それと同時に，$\mathbf{y}$ を入力とした復元アルゴリズムの反復に $\widehat{\mathcal{P}}$ を用いると，推定信号 $\widehat{\mathbf{x}}$ が得られる。このとき，復元アルゴリズムは微分可能なコンプレッサーを用いるため，上記の処理に対する勾配を計算することができ，復元アルゴリズムの性質を反映させた推定 DNN の学習ができる。

提案するモデルを学習するために，元信号 $\mathbf{x}$，コンプレッサー処理された信号 $\mathbf{y}$，パラメータの組 $\mathcal{P}$ に

*Parameter estimation and signal recovery based on deep unrolling with differentiable compressor. By Meng WANG, Natsuki AKAISHI, Koki YAMADA and Kohei YATABE (Tokyo University of Agriculture and Technology).

表–1　実験条件及び結果。評価指標は全データの中央値を示している。

Model		Learning Components	L_{overall}	Loss for $L_{\mathbf{x}}, L_{\mathbf{y}}$	L1 ↓	MSE ↓	MRSTFT ↓	V-MOS ↑	V-NSIM ↑
T-UNet	L1	T-UNet [3]	$L_{\mathbf{x}}$	L_{L1}	0.000128	0.0746	0.10570	4.7076	0.9964
	MRSTFT	T-UNet [3]	$L_{\mathbf{x}}$	L_{MRSTFT}	0.042550	1263	0.06177	4.7303	0.9994
	L1 + MRSTFT	T-UNet [3]	$L_{\mathbf{x}}$	$L_{\text{MRSTFT}} + 100L_{\text{L1}}$	**0.000123**	**0.0675**	**0.02163**	**4.7310**	**0.9995**
	MSE	T-UNet [3]	$L_{\mathbf{x}}$	L_{MSE}	0.000379	0.2679	0.22550	4.6016	0.9875
Prop.1	L1	推定 DNN [1]	$0.1L_{\mathbf{y}} + L_{\mathcal{P}}$	L_{L1}	0.001046	0.8760	0.05785	4.7299	0.9996
	MRSTFT	推定 DNN [1]	$0.1L_{\mathbf{y}} + L_{\mathcal{P}}$	L_{MRSTFT}	**0.001020**	0.8606	0.05761	4.7296	0.9996
	L1 + MRSTFT	推定 DNN [1]	$0.1L_{\mathbf{y}} + L_{\mathcal{P}}$	$L_{\text{MRSTFT}} + 100L_{\text{L1}}$	0.001023	**0.8421**	**0.05747**	4.7299	**0.9997**
	MSE	推定 DNN [1]	$0.1L_{\mathbf{y}} + L_{\mathcal{P}}$	L_{MSE}	0.001106	0.8822	0.05891	**4.7314**	0.9997
Prop.2	L1	推定 DNN [1], 復元アルゴリズム [2]	$0.1L_{\mathbf{x}} + 0.5L_{\mathbf{y}} + 15L_{\mathcal{P}}$	L_{L1}	0.001020	0.8749	0.05734	4.7298	0.9996
	MRSTFT	推定 DNN [1], 復元アルゴリズム [2]	$0.1L_{\mathbf{x}} + 0.5L_{\mathbf{y}} + 15L_{\mathcal{P}}$	L_{MRSTFT}	0.001010	0.7887	0.05698	**4.7300**	**0.9997**
	L1 + MRSTFT	推定 DNN [1], 復元アルゴリズム [2]	$0.1L_{\mathbf{x}} + 0.5L_{\mathbf{y}} + 15L_{\mathcal{P}}$	$L_{\text{MRSTFT}} + 100L_{\text{L1}}$	0.001004	0.8578	0.05700	4.7297	0.9997
	MSE	推定 DNN [1], 復元アルゴリズム [2]	$0.1L_{\mathbf{x}} + 0.5L_{\mathbf{y}} + 15L_{\mathcal{P}}$	L_{MSE}	**0.000922**	**0.7174**	**0.05456**	4.7293	0.9996

関する損失 $L_{\mathbf{x}}, L_{\mathbf{y}}, L_{\mathcal{P}}$ を用いた以下の損失関数

$$L_{\text{overall}} = \alpha L_{\mathbf{x}} + \beta L_{\mathbf{y}} + \lambda L_{\mathcal{P}} \tag{2}$$

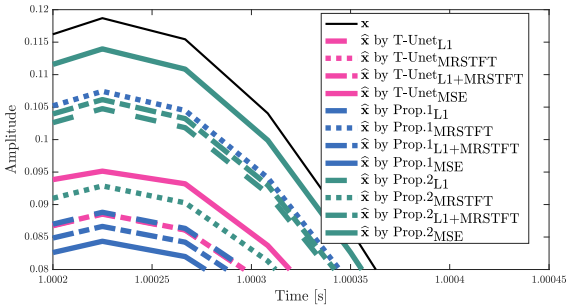
を用いる。ただし，α, β, λ はそれぞれの損失の重み係数である。$L_{\mathcal{P}}$ は $\widehat{\mathcal{P}}$ と $\mathcal{P}$ の平均二乗誤差 (MSE) を用いた。$L_{\mathbf{x}}$ と $L_{\mathbf{y}}$ には，時間領域の平均絶対誤差 L_{L1}，時間周波数領域の多重解像度誤差 (MRSTFT) L_{MRSTFT} または MSE を用いる。

4　実験

[1] と同じデータを用いて実験を行った。実験では，DNN に基づく信号復元手法のベースラインとして T-UNet [3] を用いた。また，深層展開が復元性能に与える影響を確認するため，復元アルゴリズムを含めずに学習した推定 DNN をそのまま用いる Prop.1 と，復元アルゴリズムまで含めて学習した推定 DNN を用いる Prop.2 の 2 つの手法を比較した。ただし，Prop.2 の推定 DNN を最初から学習するのは難しいため，初期値として Prop.1 の事前学習済みの推定 DNN を用いた。さらに，信号の損失関数の構成が提案手法に与える影響を確認するため，L_{L1} のみと L_{MRSTFT} のみを学習に用いた手法とも比較した。各手法の名前と構成は，表–1 の左から 1 列目と 2 列目に示す。

学習プロセスのエポック数とバッチサイズはそれぞれ Prop.1 で 500 と 6，Prop.2 で 300 と 4，T-UNet で 200 と 8 とした。真のパラメータの組 $\mathcal{P}$ の生成範囲はそれぞれ，T は $[-30, 1]$ dB，R は $[1, 20]$，τ は $[1, 100]$ ms，W は $[0, 12]$ dB とした。復元アルゴリズムは K を 30，γ を 1.9 とした。他の学習設定は文献 [1] と同じである。数値実験は，$\hat{\mathbf{x}}$ と $\mathbf{x}$ の各損失関数の結果と仮想音声品質評価である ViSQOL [5] の MOS と NSIM を用いた。

実験結果を表–1 に示す。また，各手法による復元結果 $\hat{\mathbf{x}}$ の例を図–2 に示す。表–1 の Prop.1 と Prop.2 の損失の結果を比べたとき，Prop.2 の信号の損失関数を MSE とした場合に Prop.1 のすべての条件よりも損失が小さくなったことが見て取れる。ここから，深層展開によってパラメータ推定および信号復元が



図–2　各手法による復元結果 $\hat{\mathbf{x}}$ の一例。T-UNet$_{\text{MRSTFT}}$ の復元結果は誤差が大きいため，表示されていない。

より高精度に行えるようになったと言える。しかし，V-MOS と V-NSIM はすべての手法で同程度だった。また，L_{MRSTFT} の場合を除く T-UNet の方が，提案手法のどの条件よりも損失が小さくなった。図–2 のように，提案手法の方が復元が上手くいくこともあったが，全体的に T-UNet の方が元信号を良く復元できていた。これは微分可能なコンプレッサーを用いた提案手法の学習が難しいことを示唆している。

5　むすび

本稿では，パラメータ推定と信号復元を一貫させる手法を提案した。これによって，それぞれを個別に用いるよりもパラメータと元信号をより良く推定できることが分かった。今後は，T-UNet に並ぶ性能を実現するためのモデル構造や学習方法を検討する。

参考文献

[1] 王櫟, 中村友彦, 山田宏樹, 矢田部浩平, “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,” 音講論集, pp. 259–260 (2023.9).

[2] 王櫟, 中村友彦, 山田宏樹, 矢田部浩平, “コンプレッサー処理された信号の復元に関する検討,” 音講論集, pp. 205–206 (2024.3).

[3] A. A. Nair and K. Koishida, “Cascaded time + time-frequency Unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps,” *IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 7153–7157 (2021).

[4] V. Monga, Y. Li and Y. C. Eldar, “Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing,” *IEEE Signal Process. Mag.*, **38**(2), 18–44 (2021).

[5] A. Hines, J. Skoglund, A. C. Kokaram and N. Harte, “ViSQOL: An objective speech quality model,” *EURASIP J. Audio Speech Music Process.*, **2015**(13), 1–18 (2015).