観測信号のみを用いた教師なし DNN 音声強調に関する研究

代表研究者　　宮　崎　亮　一　　徳山工業高等専門学校　情報電子工学科　准教授

1 はじめに

音声を利用したシステムにおいて，音声信号に重畳する雑音は音声の明瞭度や品質を低下させるため望ましくない．そこで，音声強調［1--13］が音声品質の低下を防ぐために用いられている．音声強調の中でも，スペクトル減算法［1］やウィーナフィルタ［2]，最小平均二乗誤差規範短時間振幅スペクトル推定器［3］などの，雑音の推定情報と与えられた音声のパワースペクトルを用いて処理を行う，古典的な非線形音声強調手法は広く使用されている．他方，深層ニューラルネットワーク（Deep Neural Network: DNN）を利用した音声強調[4--6] は高い性能を発揮できることが知られている．DNN を利用した音声強調では，雑音（noise）信号が混入した（noisy）信号と noise 信号が重畳していない（clean）信号のペアを大量に用意し，そのペアを用いて noisy 信号から clean 信号へ変換する DNN を学習する，教師あり学習を行うことが一般的である．学習後の DNN は任意の noisy 信号を入力として受け取り，それに対応した clean 信号を出力すること音声強調を行う．一方で，clean 信号は反響のない特殊な環境で収録する必要があるため，大量の clean 信号を入手するのは難しいという問題も抱えている．

近年この問題を解決するために，clean 信号を必要としない DNN 音声強調手法が盛んに研究されている[7--13]．この研究分野における方向性は主に二つに大別できる．一つは，大量の noisy 信号を利用した自己教師あり学習により音声強調 DNN を学習する手法である．例えば，文献［9］では noisy 信号と clean 信号のペアの代わりに，noisy 信号に追加の noise 信号を加えたものと noisy 信号のペアを用いる．DNN は，追加の noise 信号を除去するように学習され，推論時には入力された noisy 信号から noise 信号を除去するように動作する．また，文献［10］では非侵入型の音声品質評価指標を損失関数に用いることで，clean 信号なしで noisy 信号の品質を改善する，即ち音声強調を施す DNN の学習を実現している．

もう一つは，事前学習なしで音声強調を実現する手法である［11-14]．これらの手法は，Deep Image Prior (DIP)［15］と呼ばれる画像処理分野で示された性質から着想を得たものであり，DNN 自体の性質を利用することで音声強調を実現している.DIP は,畳み込みニューラルネットワーク（Convolutional Neural Network: CNN)の学習において，エッジ等の規則性を持つ clean 画像の方が規則性を持たない noise 画像よりも生成されるまでの学習回数が少ないという性質である．そのため，入力の一様乱数が noisy 画像を表現するように学習を進める場合，学習過程において noise 画像成分は clean 画像成分よりも生成が難しく，clean 画像成分が生成された後に noise 画像成分が生成される．clean 画像のみが生成された時点で学習を停止することで，音声強調を実現できる．

DIP を画像分解問題に応用した Double-DIP［16］と呼ばれる手法も提案されている．画像分解問題は例えば，1 枚の画像を背景画像とオブジェクト画像に分解するオブジェクト抽出や，2 枚の画像から成る 1 枚の混合画像のみが与えられた場合に，元の 2 枚の画像へと分解する透過画像分解など，混合画像を所望の複数枚の画像に分解する問題の総称である．Double-DIP では複数の CNN の組み合わせにより，画像分解問題を解決する事ができる．文献［16］では，CNN は複数の画像が混ざり合った不規則な混合画像よりも規則的な個別の画像の方が生成しやすいことが示されている（画像分解問題における DIP)．DIP を踏まえて，2 つの CNN 出力を混合して 1 つの混合画像を表現することを考えると，各 CNN は混合画像の一部を規則なく生成するのではなく，規則的な個別な画像に分解して生成するように学習が進む．そのため，混合前の CNN 出力を得ることで画像分解された所望の画像を得ることができる．

文献［15, 16]のような枠組みを音声に応用した手法［11-14］も提案されている．音声の複素スペクトログラムで DIP と同様の性質を本稿では Deep Audio Prior: DAP と呼び，DAP に着目した音声強調を DAP Speech Enhancement: DAP-SE と呼ぶ．文献［12］では，振幅スペクトログラムに DIP と同様の枠組みで学習を行うことで，トーン性のミュージカルノイズを除去できることを示している．また文献［11, 14]では，複素スペクトログラムに DIP の枠組みを適用することでガウス性雑音が除去できる，文献［13］では Demucs [17］モデルを用いることで，DIP の枠組みで時間領域でガウス性雑音を除去できることが示されている．

本研究では２つの観測信号のみを用いた教師なし DNN 音声強調手法を提案する．１つ目は振幅スペクト

ログラムの統計量に着目したDouble-DIPに基づく音声強調である．提案手法では，Double-DIP と同様に 2 つの CNN を利用する．観測信号の noisy 信号を尖度の高い clean 信号と尖度の低い noise 信号の 2 つに分解して生成するように学習を進め，最終的に尖度の高い clean 信号のみを得ることで音声強調を行う．また，一方のDNNは clean を生成しやすく，もう一方は noise を生成しやすくなるような Deep prior を設計する．２つ目はまた，DAP-SEの音声強調性能を向上させることを目的としてDAP-SEを複数回反復する手法 Iterative DAP-SE を提案する．この手法はDAP-SEの学習が十分に終了した時点における出力信号がnoiseを少量除去できていることを利用して反復的に音声強調を行う．さらに，Iterative DAP-SEに関して，反復処理によって clean の高域成分が削れてしまうことを防ぐために，複素スペクトログラムの位相修正(Instantaneous Phase Correction: iPC) [18]を導入する．

2 関連研究

2-1 Deep Image Prior (DIP)

雑音を持つ画像 $\boldsymbol{X}_{\text{img}}$ は $\boldsymbol{X}_{\text{img}} = \boldsymbol{S}_{\text{img}} + \boldsymbol{N}_{\text{img}}$ で表される．ここで，$\boldsymbol{S}_{\text{img}}, \boldsymbol{N}_{\text{img}}$ はclean画像，雑音である．$\boldsymbol{X}_{\text{img}}$ のみが与えられた際に，$\boldsymbol{X}_{\text{img}}$ から $\boldsymbol{N}_{\text{img}}$ を取り除き，$\boldsymbol{S}_{\text{img}}$ を復元するタスクが画像処理分野における雑音除去である．画像処理分野における雑音除去でも，DNN を用いた雑音除去手法が高い性能を発揮できることが示されている [19, 20]．一般的な雑音除去手法では，I 個のclean-noisyペア $\{\boldsymbol{S}_{\text{img}}^{(i)}, \boldsymbol{X}_{\text{img}}^{(i)}\}$ を用意し，以下の式に基づきDNNパラメータ θ の最適化を行うことで雑音除去を施すDNNを学習する．

$$\min_{\theta} \mathcal{L}\left(g_{\theta}(\boldsymbol{X}_{\text{img}}^{(i)}), \boldsymbol{S}_{\text{img}}^{(i)}\right) \tag{1}$$

ここで，$g_{\theta}(\cdot)$ はDNN，$\mathcal{L}(\cdot)$ は損失関数である．学習には通常，大量のデータが用いられることが多いため，ペアデータ数 I は大きな値となる．

一方で文献 [15] では，大量のデータと clean 画像が不要な雑音除去手法している．この手法では，$g_{\theta}(\cdot)$ としてCNNベースのDNNを用いて，図１ (a)に示すように学習を行う．

$$\min_{\theta} \mathcal{L}\left(g_{\theta}(\boldsymbol{Z}_{\text{img}}), \boldsymbol{X}_{\text{img}}\right) \tag{2}$$

$$\hat{\boldsymbol{X}}_{\text{img}(t)} = g_{\theta(t)}(\boldsymbol{Z}_{\text{img}}) \tag{3}$$

ここで，t は学習のステップ数，$\hat{\boldsymbol{X}}_{\text{img}(t)}$ はステップ t におけるDNN出力，$\boldsymbol{Z}_{\text{img}}$ は $[0, 0.1]$ の値を取る一様分布 $U(0, 0.1)$ からサンプリングされた入力乱数である．DIPとは，式 (3) における学習で$\boldsymbol{X}_{\text{img}}$ が十分生成されるステップ数 t_x と $\boldsymbol{S}_{\text{img}}$ が十分生成されるステップ数 t_s が，$t_s < t_x$ になる性質のことである．これは，$\boldsymbol{S}_{\text{img}}$ が有する滑らかかつ明確な構造が，不規則性を有する $\boldsymbol{N}_{\text{img}}$ よりも生成しやすいことに起因している．この性質を利用し，図１(b)に示すように t_s で学習を停止することにより，推定 clean 画像 $\hat{\boldsymbol{S}}_{\text{img}} = \hat{\boldsymbol{X}}_{\text{img}(t_s)} = g_{\theta(t_s)}(\boldsymbol{Z}_{\text{img}})$ が得られる．

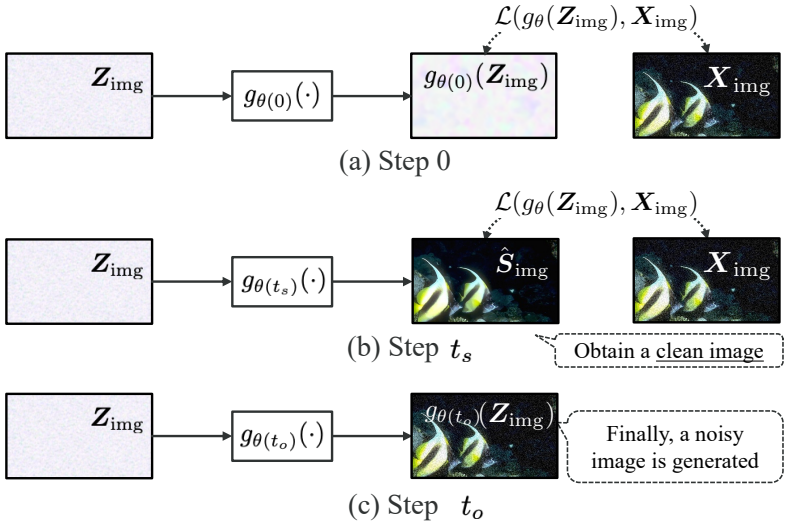

図 1　DIP を用いた画像雑音除去の概要図

2-2 Double-DIP

Double-DIP［16］は DIP を画像分解問題に適用した手法である．ここでは，いくつかある画像分解問題の中から透過画像分解について説明する．透過画像分解は式 (4) で定式化される混合画像を分解し，原画像 $\boldsymbol{S}_{1\text{img}}, \boldsymbol{S}_{2\text{img}}$ を得ることを目的としている．

$$\boldsymbol{X}_{\text{img}} = \boldsymbol{M} \odot \boldsymbol{S}_{1\text{img}} + (1 - \boldsymbol{M}) \odot \boldsymbol{S}_{2\text{img}} \tag{4}$$

ここで，$\boldsymbol{M}$ は 0.0 から 1.0 の値を持つ透過度マスクであり，$\odot$ はアダマール積を表す．透過画像分解における DIP は，式(3)において $\boldsymbol{S}_{1\text{img}}, \boldsymbol{S}_{2\text{img}}, \boldsymbol{X}_{\text{img}}$ のそれぞれを対象に学習を行なった場合，生成までのステップ数 t_{s1}, t_{s2}, t_x において $t_{s1}, t_{s2} < t_x$ の関係があることを指す．言い換えると，DNN は透過画像よりも clean 画像をより少ないステップ数で生成できるということである．図 2 に示す 3 つの DNN を用いた学習を導入し，DIP を活用することで画像分解を実現している．

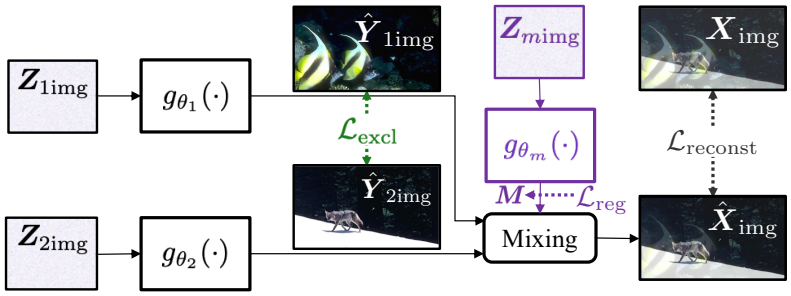

図 2　DIP を用いた画像雑音除去の概要図

$$\hat{\boldsymbol{Y}}_{1\text{img}} = g_{\theta_1}(\boldsymbol{Z}_{1\text{img}}) \tag{6}$$
$$\hat{\boldsymbol{Y}}_{2\text{img}} = g_{\theta_2}(\boldsymbol{Z}_{2\text{img}}) \tag{7}$$
$$\hat{\boldsymbol{M}} = g_{\theta_m}(\boldsymbol{Z}_{m\text{img}}) \tag{8}$$
$$\hat{\boldsymbol{X}}_{\text{img}} = \hat{\boldsymbol{M}} \odot \hat{\boldsymbol{Y}}_{1\text{img}} + (1 - \hat{\boldsymbol{M}}) \odot \hat{\boldsymbol{Y}}_{2\text{img}} \tag{9}$$

ここで，$\boldsymbol{Z}_{1\text{img}}, \boldsymbol{Z}_{2\text{img}}, \boldsymbol{Z}_{m\text{img}}$ は一様分布 $U(0, 0.1)$ からサンプリングされた入力乱数であり，$g_{\theta_1}, g_{\theta_2}, g_{\theta_m}$ は

CNN ベースの DNN，式 (9) は図２の Mixing に対応する．DNN の学習は式 (10) に従う．

$$\min_{\theta_1,\theta_2,\theta_m} \mathcal{L}\left(\hat{\boldsymbol{Y}}_{1\mathrm{img}}, \hat{\boldsymbol{Y}}_{2\mathrm{img}}, \hat{\boldsymbol{M}}, \boldsymbol{X}_{\mathrm{img}}\right) \tag{10}$$

$$\mathcal{L}(\hat{\boldsymbol{Y}}_{1\mathrm{img}}, \hat{\boldsymbol{Y}}_{2\mathrm{img}}, \hat{\boldsymbol{M}}, \boldsymbol{X}_{\mathrm{img}}) = \mathcal{L}_{\mathrm{rec}}(\hat{\boldsymbol{X}}_{\mathrm{img}}, \boldsymbol{X}_{\mathrm{img}}) + \alpha_e \mathcal{L}_{\mathrm{excl}}(\hat{\boldsymbol{Y}}_{1\mathrm{img}}, \hat{\boldsymbol{Y}}_{2\mathrm{img}}) + \alpha_r \mathcal{L}_{\mathrm{reg}}(\hat{\boldsymbol{M}}) \tag{11}$$

ここで，α_e, α_r は$\mathcal{L}_{\mathrm{excl}}, \mathcal{L}_{\mathrm{reg}}$ の重みを表すパラメータである．$\mathcal{L}_{\mathrm{rec}}$ は予測された混合画像と与えられた混合画像の再構成誤差，$\mathcal{L}_{\mathrm{excl}}$ は $\boldsymbol{X}_{\mathrm{img}}$ に含まれるエッジ成分は $\boldsymbol{S}_{1\mathrm{img}}$ か $\boldsymbol{S}_{2\mathrm{img}}$ のどちらか一方のみにより発生するものであるという仮定に基づき，$\hat{\boldsymbol{X}}_{1\mathrm{img}}$ と$\hat{\boldsymbol{X}}_{2\mathrm{img}}$ のエッジの相関を最小化することで各出力を $\boldsymbol{S}_{1\mathrm{img}}, \boldsymbol{S}_{2\mathrm{img}}$ へ誘導する．$\mathcal{L}_{\mathrm{reg}}$ は $\hat{\boldsymbol{M}}$ の正則化項である．

$\mathcal{L}_{\mathrm{excl}}$ により, $\boldsymbol{X}_{\mathrm{img}}$ が有するエッジは $\hat{\boldsymbol{Y}}_{1\mathrm{img}}, \hat{\boldsymbol{Y}}_{2\mathrm{img}}$ のどちらか一方にのみ生成される. また, DIP により，各 DNN 出力はそれぞれ $\boldsymbol{S}_{1\mathrm{img}}, \boldsymbol{S}_{2\mathrm{img}}$ へ誘導される．これらの働きにより，$\boldsymbol{S}_{1\mathrm{img}}, \boldsymbol{S}_{2\mathrm{img}}$ がそれぞれどちらか一方の DNN で生成されるように学習が進む．最終的に，十分なステップ数 t_o での各 DNN の出力が $\hat{\boldsymbol{S}}_{1\mathrm{img}} = g_{\theta_1(t_o)}(\boldsymbol{Z}_{1\mathrm{img}}), \hat{\boldsymbol{S}}_{2\mathrm{img}} = g_{\theta_2(t_o)}(\boldsymbol{Z}_{2\mathrm{img}})$ となり，画像分解問題を解くことができる．

2-3 Deep Audio Prior (DAP)

離散時間領域において， noisy $x(l)$ は以下のように表せる．

$$x(l) = s(l) + n(l) \tag{12}$$

ここで, l は離散時間のインデックス，$s(l)$ は clean，$n(l)$ は noise を表す．また，フレーム長 N の短時間フーリエ変換 (Short-Time Fourier Transform: STFT)により，$x(l)$ の複素スペクトル $X(m,k)$ を次式より得る．

$$X(m,k) = \sum_{n=0}^{N-1} x_m(l - ma) e^{-j2\pi kn/N} \tag{13}$$

ここで，a はシフト長，m は時間フレームのインデックス，k は周波数ビンのインデックス，$x_m(l-ma) = w(l-ma)x(l)$ であり，$w(l)$ は窓長 N の窓関数である．そして，式 (13) で得られた $X(m,k)$ を集約することで複素スペクトログラムを $\boldsymbol{X}$ が得られる．

DAP-SE は DIP と同様の性質である DAP を用いて音声強調を行う手法である．具体的には，次式に基づき，$\boldsymbol{X}$ と同形状の一様乱数 $\boldsymbol{Z}$ を DNN の入力とし，目的信号である$\boldsymbol{X}$ に近づけるように学習することで音声強調を行う．

$$\min_{\theta} \mathcal{L}(\mathcal{F}_{\theta^t}(\boldsymbol{Z}), \boldsymbol{X}) \tag{14}$$

図３に DAP-SE の概要図を示す． DAP-SE の t_s 時点では $\hat{\boldsymbol{S}}^{t_s} = \mathcal{F}_{\theta^{t_s}}(\boldsymbol{Z})$ で表される clean に近い音声（音声強調された信号），t_n 時点では $\hat{\boldsymbol{S}}^{t_n} = \mathcal{F}_{\theta^{t_n}}(\boldsymbol{Z})$ で表される noisy に近い音声がそれぞれ得られ，DIP と同様に t_s で学習を停止することで音声強調を実現できる．

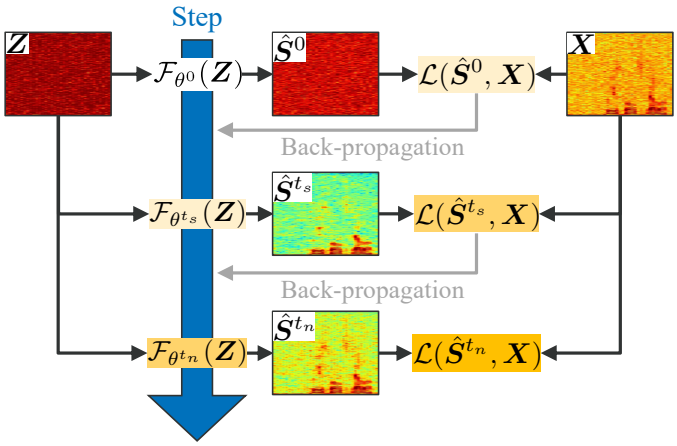

図３　DAP-SE の概要図

3 提案手法１：Double-DIP ベースの音声強調

3-1 提案手法１の概要

振幅スペクトログラム $|\boldsymbol{S}|, |\boldsymbol{N}|$ について次式の加法性が成り立つと仮定する．

$$|\boldsymbol{X}| \simeq |\boldsymbol{S}| + |\boldsymbol{N}| \tag{15}$$

提案手法１では図４に示す構成で，２つの推定信号 $|\hat{\boldsymbol{S}}|_M = g_{\theta_1}(\boldsymbol{Z}_{1,M}), |\hat{\boldsymbol{N}}| = g_{\theta_2}(\boldsymbol{Z}_2)$ の和 $|\hat{\boldsymbol{X}}|_M = |\hat{\boldsymbol{S}}|_M + |\hat{\boldsymbol{N}}|$ が与えられた $|\boldsymbol{X}|$ に近づくようにDNNを学習する．ここで Mはバッチサイズを表す．

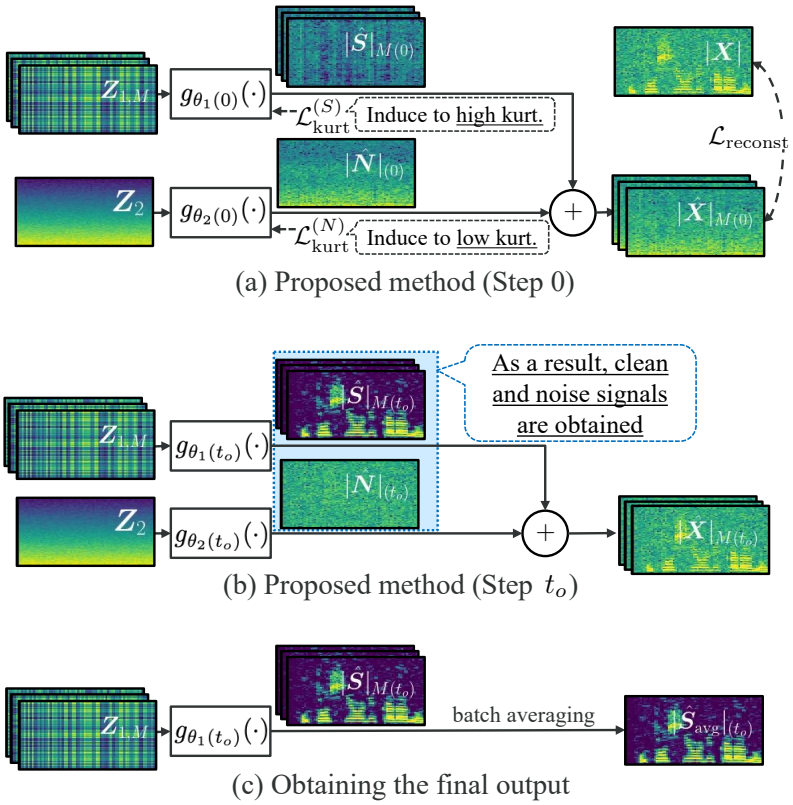

図４　提案手法１の概要図

提案法では次式により学習を行う．

$$\min_{\theta_1,\theta_2} \mathcal{L}\left(|\hat{\boldsymbol{S}}|_M, |\hat{\boldsymbol{N}}|, |\boldsymbol{X}|\right) \tag{15}$$

$$\mathcal{L}\left(|\hat{\boldsymbol{S}}|_M, |\hat{\boldsymbol{N}}|, |\boldsymbol{X}|\right) = \mathcal{L}_{\text{reconst}}(|\hat{\boldsymbol{X}}|_M, |\boldsymbol{X}|) + \mathcal{L}^{(S)}_{\text{kurt}}(|\hat{\boldsymbol{S}}|_M, |\boldsymbol{X}|) + \mathcal{L}^{(N)}_{\text{kurt}}(|\hat{\boldsymbol{N}}|, |\boldsymbol{X}|) \tag{16}$$

$\mathcal{L}_{\text{reconst}}$ は $|\hat{\boldsymbol{X}}|_M$ と $|\boldsymbol{X}|$ の再構成誤差であり，ここでは平均絶対誤差を採用した．$\mathcal{L}^{(S)}_{\text{kurt}}, \mathcal{L}^{(N)}_{\text{kurt}}$ は Double-DIP の式（11）における $\mathcal{L}_{\text{excl}}$ に対応する損失項である．これらの項は clean 信号と noise 信号の分解を促進する働きを持つ．$\mathcal{L}^{(S)}_{\text{kurt}}, \mathcal{L}^{(N)}_{\text{kurt}}$ と，Deep prior の働きにより，図４（b)に示す十分なステップ数 t_o の反復後に，推定 clean 信号 $|\hat{\boldsymbol{S}}|_{m(t_o)}$，推定 noise 信号 $|\hat{\boldsymbol{N}}|_{(t_o)}$ が得られる．最終的な出力は，推定 clean 信号 $|\hat{\boldsymbol{S}}|_{m(t_o)}$ のバッチ平均 $|\hat{\boldsymbol{S}}_{\text{avg}}|$ である（図４（c))．

3-2 提案手法１における Deep Prior

提案手法１において，clean 信号の生成では $g_{\theta_1}(\cdot)$ で生成されるステップ数 t_{g1}，$g_{\theta_2}(\cdot)$ で生成されるステップ数 t_{g2} の間に $t_{g1} < t_{g2}$ の関係が，noise 信号の生成では $t_{g2} < t_{g1}$ の関係があることが望ましい．なぜなら，これらの関係が存在する場合，clean 信号の生成は $g_{\theta_1}(\cdot)$ 側へ，noise 信号の生成は $g_{\theta_2}(\cdot)$ 側へと

誘導されると考えられるためである．

そのために，入力特徴量 $\boldsymbol{Z}$ と出力層に工夫を施す．Double-DIP における動画処理や Double-DIP を応用した音源分離［21］では，入力乱数の時間方向の値の一貫性を持たせることで出力の時間方向の一貫性が向上することが示されている．これらの研究から着想を得て，時間周波数方向で一貫した，高品質な clean 信号を得るために次式に従い $\boldsymbol{Z}_1$ を定義する．

$$Z_1[k,l] = \frac{1}{2}(u_k + u_l) \tag{17}$$

ここで，u_k, u_l はそれぞれ $U(0,0.1)$ からサンプリングされた入力乱数であり，u_k は周波数ビンごとに異なる値を持つが，時間フレームでは同一の値を持つ．u_l は逆に時間フレームごとに異なる値，周波数ビンで同一の値を持つ．また，$\boldsymbol{Z}_1$ の取りうる範囲は $[0,0.1]$ の範囲に限定される．$\boldsymbol{Z}_1$ は時間方向 l，周波数方向 k で同一の値を持つ乱数の和により構成されており，時間周波数で一貫した出力の取得が期待できる．

$\boldsymbol{Z}_2$ については低周波側から高周波側にかけて滑らかに値が減少していく meshgrid を用いる．meshgrid は滑らかな出力を誘導する事前分布として働くことが文献［15］で述べられており，noise 信号の滑らかな特徴を表現する働きが期待できる．提案手法 1 では，$\boldsymbol{Z}_2$ を meshgrid に小さな乱数 Δu を加えた次式により定義する．

$$Z_2[k,l] = 0.09\frac{K-k}{K} + \Delta u \tag{18}$$

ここで，Δu は $U(0,0.01)$ からサンプリングされた乱数であり，$\boldsymbol{Z}_1$ と同様に，取りうる範囲は $[0,0.1]$ となるように定式化した．

さらに，信号のスパース性を考慮した Softplus 出力層を設計する．一般的に，$|\boldsymbol{S}|$ はスパースな信号であることが広く知られており，一方で，$|\boldsymbol{N}|$ はスパースでない信号である．これらの性質を考慮するために，提案手法では出力層にパラメータの異なる Softplus 関数 $S(x) = \frac{1}{\beta}\log(1+e^{\beta x})$ を用いる．Softplus 関数のパラメータ β は大きければ大きいほど ReLU に近いグラフ形状を持つ．零付近で急峻な形状を持つ関数であるほどスパースな出力を誘導するため，$g_{\theta_1}(\cdot)$ では大きな β を持つ Softplus 関数を，$g_{\theta_2}(\cdot)$ では小さな β を持つ Softplus 関数を出力層へ用いることで，$|\boldsymbol{S}|$ は $g_{\theta_1}(\cdot)$ へ，$|\boldsymbol{N}|$ は $g_{\theta_2}(\cdot)$ へ誘導されることが期待できる．

3-3 分割尖度を用いた損失項の設計

提案手法 1 で用いられる分割尖度に基づく損失項の設計について述べる．まず初めに，時間周波数領域を分割した各分割領域における尖度を考える．時間周波数各方向について信号を分割した，分割領域における期待値 $\mathcal{E}$ を定義する．

$$\left(\mathcal{E}(|\boldsymbol{Y}|^2)[\tilde{k},\tilde{\tau}]\right)_{\tilde{k}=0,\tilde{\tau}=0}^{\tilde{K}-1,\tilde{T}-1} = \frac{1}{r_k r_\tau}\sum_{k_r=0}^{r_k-1}\sum_{\tau_r=0}^{r_\tau-1}|\boldsymbol{Y}|^2[r_k\tilde{k}+k_r, r_\tau\tilde{\tau}+\tau_r] \tag{19}$$

ここで，r_k, r_τ はそれぞれ周波数，時間方向の 1 分割に含まれる要素数であり，$\tilde{K}, \tilde{T}$ それぞれ周波数，時間

方向の分割数である．信号を分割せずに計算された尖度では，文献［22］で示された結果と同様に，雑音の種類に応じて異なる尖度を取ることがわかる．これは，周波数や時間方向の動的なパワー変化が理由である．対して，分割領域における尖度では動的なパワー変化が緩和されるため，雑音の種類に関わらず，noise 信号は低い値を，clean 信号は高い値を取る．よって，分割領域における尖度は noisy 信号を clean 信号と noise 信号に分けて生成するために有効な手掛かりとなる．実際の学習には，次式に示す損失関数を用いる．

$$\mathcal{L} = \mathcal{L}_{\text{reconst}} + \mathcal{L}_{\text{kurt}}^{(S)} + \mathcal{L}_{\text{kurt}}^{(N)} \tag{20}$$

ここで，$\mathcal{L}_{\text{kurt}}^{(S)}, \mathcal{L}_{\text{kurt}}^{(N)}$ は尖度に基づく正則化項である．提案手法では，一方の出力信号が有する尖度を高く，他方の出力信号が有する尖度を低くなるように誘導する．

3-4 客観評価実験

clean 信号 $\boldsymbol{s}$ として，JNAS 音声コーパス［23］より 5 サンプルを用いた．また，noise 信号として DEMAND［24］より 8 種類のノイズを用いた．noisy 信号 x の SNR は 5, 10, 15［dB］の 3 種類の，全組み合わせとなる 120 サンプルを音声強調の対象信号とする．$\boldsymbol{X}$ に変換する際の FFT 長，窓長，シフト長は 512, 512, 128 とした．DNN アーキテクチャとして Une［25］ベースの構造の DNN を用いた．最適化アルゴリズムには学習率 $\lambda = 0.001$ の Adam［26］を用いた．客観評価指標として，音声の品質を表す PESQ［27］と音声強調後の音声の歪み度合いを表す Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)［28］を用いた．提案手法におけるハイパーパラメータについて述べる．各損失関数の重みは，$(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (0.00001, 0.001, 0.00001, 2.0)$ とした．

雑音の種類毎の性能評価を行う．ここでは，比較手法として文献［14］と同様の手法 (Single-DP) を用いた．評価結果を図 5 に示す．まず，提案手法では，全ての雑音に対して，比較手法よりも高いスコアを示していることがわかる．このことから，Double-DIP ベースの構造と振幅スペクトログラムに基づく提案法が，音声強調においてより有効であることがわかる．また，従来手法では音声強調が難しい環境雑音 (presto, tbus) においてもスコアを改善できることがわかる．これは，分割領域における尖度に基づく損失項が，雑音の種類に依らず，有効であるためだと考えられる．

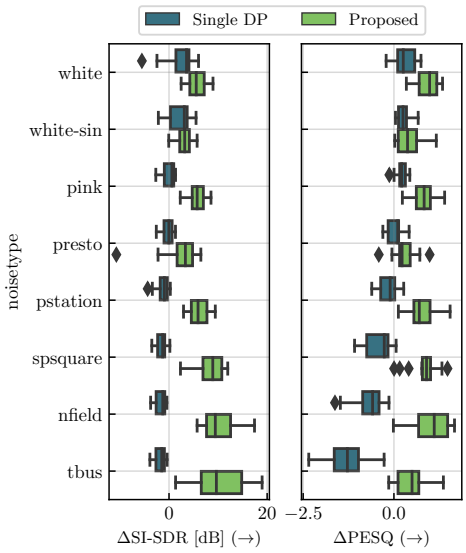

図 5　雑音の種類ごとの客観評価値改善量の比較

4 提案手法２：Iterative DAP-SE

4-1 提案手法２の概要

DAP-SE の予備実験より，t_n における各評価指標値が全ての入力 SNR においてわずかながら改善していることを確認した．よって t_n 時点においては少し noise が除去された音声が得られる．すなわち，t_n における出力信号を目的信号として用い，反復的に DAP-SE を行うと，少しずつ noise が除去されていく可能性があると考える．よって，DAP-SE の目的信号を変更しながら反復的に音声強調を行う Iterative DAP-SE を提案する．図６に提案手法２の概要図を示す．DAP-SE と同様に $\boldsymbol{Z}$ を DNN の入力として，目的信号である $\boldsymbol{X}$ に近づけるように DNN のパラメータ θ_1^t の最適化を t_n まで行う．そして，推論時の出力信号として noise が少量除去された $\hat{\boldsymbol{S}}_1$ を得る．これを一回の反復として，次の反復の目的信号を前の出力信号として DAP-SE を繰り返し，最適な反復回数 C で noise が十分に除去された音声 $\hat{\boldsymbol{S}}_C$ を得る．

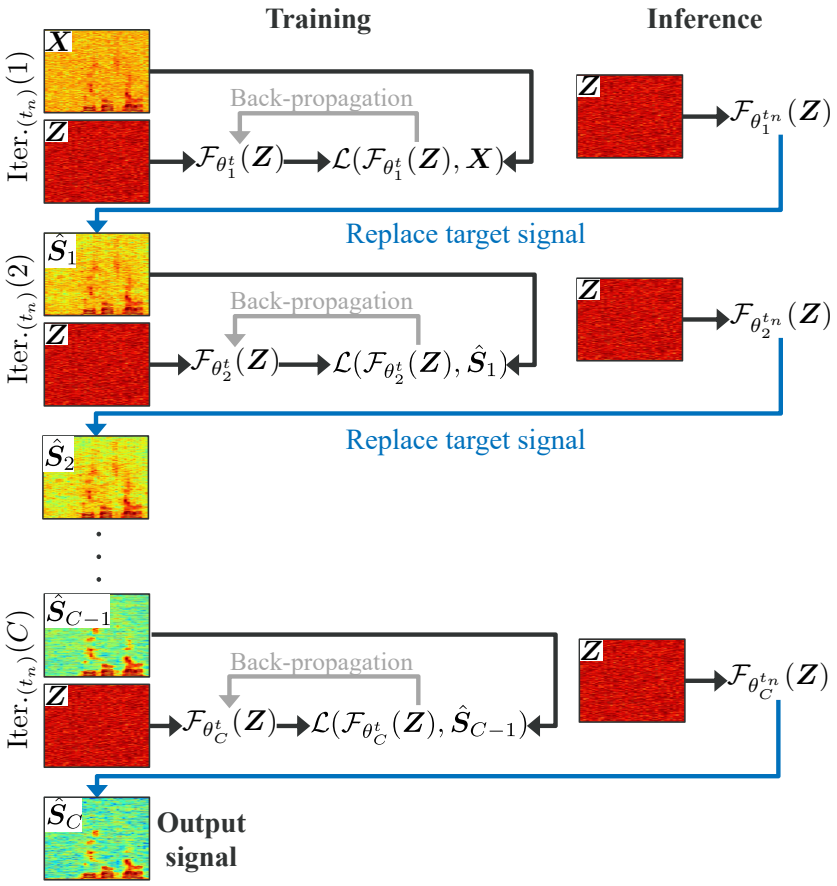

図６　DAP-SE の概要図

4-2 iPC の導入

予備実験より，提案手法２では反復処理によって noise が少しずつ除去されたが，同時に clean の高域成分が削れ，従来の DAP-SE と同程度の性能しか得ることができなかった．この問題を解決するために，目的信号を DAP で表現しやすい信号に変換することを考える．変換手法として，iPC を施すことを検討する．

位相の変化は各時間周波数の要素における瞬時周波数 $v(m,k)$ と密接に関連している．複数の正弦波からなる音声信号について，各正弦波が十分に分離され，他の正弦波からの干渉が無視できると仮定すると，$X(m+1,k)$ は次式で表せる．

$$X(m+1,k) = e^{2\pi j v(m,k) a/N} X(m,k) \tag{21}$$

ここで，$v(m,k)$ は各周波数ビンのインデックス k で異なるが，すべての時間フレーム m で同じ値を持つ．

つまり，複素スペクトログラムの位相は m が変化するに従って瞬間周波数 $v(m,k)$ の定数倍として変化する．この位相変化を打ち消すために，Instantaneous Phase Correction STFT（iPC-STFT）では次式で示される位相補正行列 $\boldsymbol{E}$ を適用して，より低ランクな複素スペクトログラム $\boldsymbol{X}_{\mathrm{iPC}}$ が得られる．

$$\boldsymbol{X}_{\mathrm{iPC}} = \boldsymbol{E} \odot \boldsymbol{X} \tag{22}$$

$$E(m,k) = \prod_{\eta=0}^{m-1} e^{-2\pi j v(\eta,k) a/N} \tag{23}$$

$\boldsymbol{E}$ の要素 $E(m,k)$ は式（23）で表され，全ての k について $E(0,k)=1$ が成り立つ．iPC を施す利点として，位相修正の効果を打ち消す際には $\boldsymbol{E}$ の複素共役を掛ければよいため，処理自体が可逆な変換なことが挙げられる．そのため，$\boldsymbol{X}$ を扱いやすい $\boldsymbol{X}_{\mathrm{iPC}}$ へ変換して信号処理を行なった後に，逆変換で元の領域に戻すことが可能であり，振幅スペクトルの低ランク性を用いた手法への応用がされている．

STFT 領域と iPC-STFT 領域によって，DAP における clean の生成具合の違いを比較する．ここで，clean の生成具合の違いを確認しやすくするために，目的信号を clean の STFT 領域における複素スペクトログラム $\boldsymbol{S}$ として実験を行う．また，iPC-STFT 領域によって得られる clean の複素スペクトログラム $\boldsymbol{S}_{\mathrm{iPC}}$ を用いて学習する場合を次式のように定義する．

$$\tilde{\boldsymbol{S}}_{\mathrm{iPC}} = \mathcal{F}_{\theta^{tn}_{\mathrm{iPC}}}(\boldsymbol{Z}), \quad \min_{\theta_{\mathrm{iPC}}} \mathcal{L}(\mathcal{F}_{\theta^{t}_{\mathrm{iPC}}}(\boldsymbol{Z}), \boldsymbol{S}_{\mathrm{iPC}}) \tag{24}$$

図 7 に $\boldsymbol{S}, \tilde{\boldsymbol{S}}, \tilde{\boldsymbol{S}}_{\mathrm{iPC}}$ のスペクトログラムと SI-SDR を示す．図 7 （b）は $\boldsymbol{S}$ を目的信号とした場合でも clean の高域成分が生成されていない．一方，図 7 （c）では ，図 7 （b）と比べて clean の高域成分を生成できており， 客観評価地も改善していることが確認できた．これは，位相修正を施すことで複素スペクトログラムが低ランクで表現され，clean 成分が DAP で表現しやすい信号に変換されたためだと考える．

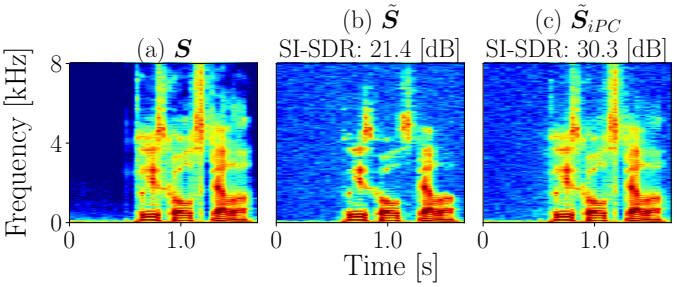

図 7　スペクトログラム：(a) clean, (b), (c) DAP-SE の目的信号を STFT, iPC-STFT 領域で処理された信

号としたときの出力信号

4-3 客観評価実験

$\mathrm{Conv.}_{(t_s)}, \mathrm{Iter.}_{(t_n)}$ における iPC を施す影響を確認することを目的として，客観評価実験を行う．本実験では DNN のアーキテクチャには文献 [14] と同様の拡張畳み込みを適用した U-Net を採用し，最適化アルゴリズムには学習率 0.001 の Adam [26] を用いた．また，損失関数は平均二乗誤差を用いた．学習回数 t_n は 7000 回とした．モデルの入力には実部と虚部で 2 チャンネルの，周波数ビン数，時間フレーム数が (512, 188) の一様乱数に従う複素スペクトログラムを用いた．窓関数にはハミング窓を用い，フレーム長，窓長は 1024，シフト長は 256 とした．clean には DEMAND [24] より５文，noise には白色ガウス雑音を用いた．目的信号を Signal-to-Noise Ratio~(SNR)が 2.5, 7.5, 12.5 [dB]の noisy とし，学習回数に対応する出力信号の Perceptual Evaluation of Speech Quality (PESQ) [27], Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [28],の改善量を客観評価指標として用いた．これらの音声のサンプリング周波数は 16 [kHz]とした.また，StepLRScheduler により学習回数 500 回ごとに学習率を 1/2 倍にした．

各手法におけるΔSI-SDRが最大となる学習回数・反復回数における各評価指標を 表１に，同条件における入力 SNR が 7.5 [dB] の出力信号のスペクトログラムを 図８にそれぞれ示す．なお，$\mathrm{Conv.}_{(t_s)}$ に関して STFT 領域，iPC-STFT 領域で処理された信号を用いる手法をそれぞれ $\mathrm{Conv.}_{(t_s)\mathrm{STFT}}, \mathrm{Conv.}_{(t_s)\mathrm{iPC}}$, のように定義し，$\mathrm{Iter.}_{(t_n)}$ も同様に記述する．表１より，$\mathrm{Iter.}_{(t_n)\mathrm{iPC}}$ は $\mathrm{Iter.}_{(t_n)\mathrm{STFT}}$ と比較して各評価指標値が改善していることが分かる．$\mathrm{Iter.}_{(t_n)\mathrm{STFT}}$ は反復処理によって clean の高域成分が削れてしまうが，iPC を施すことで，信号全体が DAP で表現しやすい信号に変換され，clean 成分と noise 成分がより表現されやすくなることが確認できた．よって $\mathrm{Iter.}_{(t_n)\mathrm{iPC}}$ は最適な反復回数が $\mathrm{Iter.}_{(t_n)\mathrm{STFT}}$ と比べて増加するが，clean の高域成分を表現しながら，少しずつ noise 成分を除去し，結果として性能が向上したと言える．また，図８(a)，(c)を比較すると，図８ (c) の方が clean の高域成分を表現しながら noise が除去されていることが確認できる．以上より，DAP-SE をベースとして iPC-STFT 領域で反復処理を行うことによって，従来手法の性能改善を実現できた．

表１　それぞれの入力 SNR における ΔSI-SDR と ΔPESQ

SNR [dB]	Method	ΔSI-SDR [dB]	ΔPESQ
2.5	$\mathrm{Conv.}_{(t_s)\mathrm{STFT}}$	7.995	0.126
	$\mathrm{Conv.}_{(t_s)\mathrm{iPC}}$	3.300	0.014
	$\mathrm{Iter.}_{(t_n)\mathrm{STFT}}$	6.946	0.104
	$\mathrm{Iter.}_{(t_n)\mathrm{iPC}}$	**8.986**	**0.277**
7.5	$\mathrm{Conv.}_{(t_s)\mathrm{STFT}}$	6.173	0.214
	$\mathrm{Conv.}_{(t_s)\mathrm{iPC}}$	3.608	0.063
	$\mathrm{Iter.}_{(t_n)\mathrm{STFT}}$	5.827	0.212
	$\mathrm{Iter.}_{(t_n)\mathrm{iPC}}$	**7.360**	**0.535**
12.5	$\mathrm{Conv.}_{(t_s)\mathrm{STFT}}$	4.519	0.364
	$\mathrm{Conv.}_{(t_s)\mathrm{iPC}}$	3.284	0.131
	$\mathrm{Iter.}_{(t_n)\mathrm{STFT}}$	4.876	0.371
	$\mathrm{Iter.}_{(t_n)\mathrm{iPC}}$	**5.458**	**0.481**

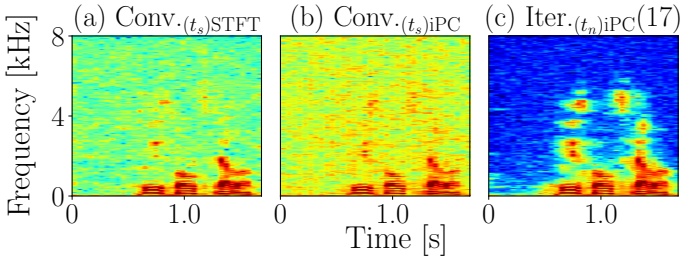

図８　スペクトログラム：(a) STFT 領域における従来手法，(b) iPC-STFT 領域における従来手法，(c) iPC-STFT 領域における提案手法２

まとめ

本研究では２つの観測信号のみを用いた教師なし DNN 音声強調手法を提案すした，提案手法１では，Double-DIP ベースの構造とスペクトログラム尖度の正則化に基づく雑音除去手法を提案した．尖度の正則化に基づき雑音除去が可能であること，特に従来手法では難しかった環境雑音に対しても雑音除去が可能であることを示した．提案手法２では，DAP-SE の「音声強調性能が不十分である点」を解決するために出力信号を次の DAP-SE の目的信号として反復的に音声強調を行う手法を提案し，手法の改良案として iPC を導入した．実験により，従来手法の DAP-SE と比べて音声の品質が改善することを実証した．現状では提案手法１と提案手法２を独立に研究を進めているが，今後は両者の利点を活かした音声強調の手法の実現に向けて研究を進めていく．

【参考文献】

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113–120, 1979.
[2] N. Wiener, “Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications,” The MIT press, 1949.
[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp.1109–1121, 1984.
[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 1, pp. 7–19, 2014.
[5] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder,” in Interspeech, pp. 436–440, 2013.
[6] S. Pascual, A. Bonafonte, and J. Serra, “Segan: Speech enhancement generative adversarial network,” in Interspeech, pp. 3642–3646, 2017.
[7] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, and J. Hershey, “Unsupervised sound separation using mixture invariant training,” Advances in Neural Information Processing Systems, vol. 33, pp. 3846–3857, 2020.
[8] N. Ito and M. Sugiyama, “Audio signal enhancement with learning from positive and unlabeled data,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5, 2023.
[9] T. Fujimura, Y. Koizumi, K. Yatabe, and R. Miyazaki, “Noisy-target training: A training strategy for dnn-based speech enhancement without clean speech,” in European Signal Processing Conference (EUSIPCO), pp. 436–440, 2021.

[10] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, andY. Tsao, “Metricgan-u: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7412–7416, 2022.
[11] Z. Zhang, Y. Wang, C. Gan, J. Wu, J. B. Tenenbaum, A. Torralba, and W. T. Freeman, “Deep audio priors emerge from harmonic convolutional networks,” in International Conference on Learning Representations, 2019.
[12] T. Fujimura and R. Miyazaki, “Removal of musical noise using deep speech prior,” Applied Acoustics, vol. 194, p. 108772, 2022.
[13] A. Turetzky, T. Michelson, Y. Adi, and S. Peleg, “Deep audio waveform prior,” in Interspeech, pp. 2938–2942, 2022.
[14] V. S. Narayanaswamy, J. J. Thiagarajan, and A. Spanias, “On the design of deep priors for unsupervised audio restoration,” in Interspeech, pp. 2167–2171, 2021.
[15] D. Ulyanov, V. Lempitsky, and A. Vedaldi, “Deep image prior,” International Journal of Computer Vision, vol. 128, no. 7, pp. 1867–1888, 2020.
[16] Y. Gandelsman, A. Shocher, and M. Irani, ““double-dip””: Unsupervised image decomposition via coupled deep-image-priors,” in IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11026–11035, 2019.
[17] A. Defossez, G. Synnaeve, and Y. Adi, “Real time speech enhancement in the waveform domain,” in Interspeech, pp. 3291–3295, 2020.
[18]K. Yatabe and Y. Oikawa, “Phase Corrected Total Variation for Audio Signals,” Proc. ICASSP, pp.855–859 2018.
[19] J. Xie, L. Xu, and E. Chen, “Image denoising and inpainting with deep neural networks,” Advances in Neural Information Processing Systems, vol. 25, 2012.
[20] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising,” IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3142–3155, 2017.
[21] Y. Tian, C. Xu, and D. Li, “Deep audio prior,” arXiv preprint arXiv:1912.10292, 2019.
[22] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, “Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics,” in International Workshop for Acoustic Echo and Noise Control, 2008.
[23] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, “Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research,” Journal of the Acoustical Society of Japan (E), vol. 20, no. 3, pp. 199–206, 1999.
[24] J. Thiemann, N. Ito, and E. Vincent, “Demand: A collection of multi-channel recordings of acoustic noise in diverse environments,” in Meetings Acoust, pp. 1–6, 2013.
[25] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer Assisted Intervention, pp. 234–241, 2015.
[26] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in International Conference on Learning Representations, 2015.
[27] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, “Perceptual evaluation of speech quality (pesq) the new itu standard for end-to-end speech quality assessment part ii: Psychoacoustic model,” Journal of the Audio Engineering Society, vol. 50, no. 10, pp. 765–778, 2002.
[28] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “Sdr–half-baked or well done?” in IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 626–630, 2019.