Vol.2022-SLP-140 No.41
2022/3/2

ソース・フィルタ・チャネル分解に基づく自己教師ありニューラル音声復元

佐伯 高明[1,a)]　高道 慎之介[1,b)]　中村 友彦[1]　丹治 尚子[1]　猿渡 洋[1]

概要：現代の音声工学研究に用いられる音声データは，一般に高品質な録音機器や整備された環境で収録されるが，過去の年代の録音音声などの低品質な音声データの活用が求められる場面も多い．しかし，当該録音の話者の高品質データや録音機器の情報が得られないため，劣化音声と高品質音声との対データを用いて音声復元モデルを教師あり学習することは困難である．そこで，本研究では，劣化音声データのみを用いた自己教師ありニューラル音声復元手法を提案する．提案手法は，劣化音声から復元音声の音声特徴量を推定する分析モジュール，音声特徴量から復元音声を生成する合成モジュール，復元音声に録音機器由来の乗算歪みを付与するチャネルモジュールからなり，入力波形と出力波形の再構成誤差を最小化することで，劣化音声のみから音声復元モデルを学習できる．さらに，提案手法は，劣化音声の生成過程をモデル化することでソース・フィルタ・チャネル成分を別々に抽出できる．実験的評価では，提案手法による音声復元性能および劣化音声特徴の操作性を評価し，提案手法の有効性を示す．

キーワード：音声復元，音声表現学習，自己教師あり学習，ニューラルボコーダ

Takaaki Saeki[1,a)]　Shinnosuke Takamichi[1,b)]　Tomohiko Nakamura[1]　Naoko Tanji[1]
Hiroshi Saruwatari[1]

1. はじめに

現代の音声工学研究に用いられる音声データは，一般に高性能なディジタル録音機器や整備された環境で収録されることが多い．一方で，ディジタル録音技術の普及以前の音声データも，本研究分野に有用な資源である．1900 年前後の蓄音機の販売から 1990 年代のコンパクトディスクの普及まで，音声はアナログ機器に保存されてきた．その音声データは，当時の人物の音声言語を保存しており，社会文化と音声言語文化を通時的に扱うための貴重な資料である．しかしながら，機器自体の品質とその経時劣化により，その音声データは，現代の高音質な音声データと比較して音響的な歪み（チャネル歪み）が大きい．そこで，音響的な歪みのある劣化音声から，歪みの無い音声特徴量を分析したり，高品質音声を復元することは，重要な課題である．また，その音響的歪みにも当時の社会文化が潜在するため，劣化音声から音響的歪みを抽出することは有用である．しかし，当該劣化音声の高品質データや音響的歪みは一般に得られないため，劣化音声と高品質音声との対データを用いて音声復元モデルを教師あり学習することは困難である．

そこで，本研究では，劣化音声のみから音声復元モデルを自己教師あり学習する手法を提案する．劣化音声の生成過程では，まず時変なソース・フィルタ特徴により高品質な音声が生成され，次に録音機器由来の時不変なチャネル特徴が付与される．本研究の提案手法は，この生成過程を模擬し，劣化音声から復元音声の音声特徴量および録音機器由来のチャネル特徴を抽出する分析モジュール，音声特徴量から復元音声を生成する合成モジュール，復元音声にチャネル特徴を付与するチャネルモジュールからなる．入力の劣化音声波形と，モデルが推定する劣化音声波形の再構成誤差を最小化することで，劣化音声のみから音声復元モデルを学習する．ここで，チャネル成分が復元音声に含まれないように自己教師あり学習するため，分析モジュールでは，時変なソース成分とフィルタ成分，時不変なチャネル成分を別々に抽出する．さらに，本研究では，分析部

[1] 東京大学 大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo, Bunkyo, Tokyo 113–8656, Japan.
[a)] takaaki_saeki@ipc.i.u-tokyo.ac.jp
[b)] shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

が復元音声の音響特徴量を出力することを保証するための学習手法を導入する．1 つは双方向学習法であり，この手法では，任意の高品質データをチャネルモジュールと分析モジュールに通し，高品質音声の音響特徴量が推定されるように正則化を行う．もう一方の擬似データ生成による事前学習法では，高品質データから擬似的な劣化音声データを生成し，それらをペアデータとして用いた教師あり学習により音声復元モデルの事前学習を行うことで，自己教師あり学習のためのモデルパラメータの初期値を得る．実験的評価では，提案する自己教師あり学習法により，1) 擬似データで教師あり学習したモデルを適用した場合よりも有意に音質・話者類似性が改善すること，2) 提案手法による復元音声は，劣化音声の分析再合成音声よりも有意に高音質となること，3) チャネルマッチングにより高品質音声に劣化音声のチャネル成分を付与できることを示す．

2. 関連研究

かねてより低品質な音声を高品質な音声に復元することを目的とした研究が行われている．特定の音声復元タスクに着目した研究として，残響除去 [1] やデクリッピング [2] などの特定の音響的歪みを除去する研究や，帯域拡張の研究 [3] が存在する．本研究の提案手法では，特定の音響的歪みに着目するのではなく，チャネル歪みを表す特徴を自己教師あり学習により自動的に獲得することで音声復元を行う．本研究と同様に複数の音声復元タスクを行う研究 [4] も存在する．とりわけ本研究の提案手法に最も近い研究として，2 段階の分析・合成モデルによって音声復元を行う研究が行われている [5]．この研究では，劣化音声の音響特徴量から復元音声の音響特徴量を出力する分析モジュールと，復元音声の音響特徴量から復元音声波形を生成する合成モジュールを別々に教師あり学習することによって音声復元を行う．本研究とこの研究との最も大きな相違点は，本研究では，劣化音声のみから自己教師あり学習によって学習を行うため，ペアデータを必要としないことである．これによって，本研究の提案手法は，高品質な音声データの存在しない過去の年代の録音などを活用できる．4 節でも議論を行うように，教師あり学習した高品質な音声復元モデルをドメイン外の実データに対して適用すると大きく性能が下がることがあり，本研究の提案手法はその問題の解決策となりうる．また，本研究の提案手法では，音声復元を行うと同時にチャネル成分を抽出できるため，高品質音声に対して所望録音のチャネル歪みを付与するという音声加工 (チャネルマッチング) も可能となる．

また，近年，音声の自己教師あり表現学習法についての研究が行われている．wav2vec 2.0 [6] や HuBERT [7] をラベルなし音声データで自己教師あり学習し，様々な下流タスクに対して fine-tuning [8] することで，高い性能を達成することが確認されている．このようなモデルは，ラベルのない大規模データに対してデータの潜在的な特徴を学習できるため，より多くのデータを活用できるという利点がある．本研究の提案手法はこのメリットを享受しつつ，ソース・フィルタ・チャネル特徴が付与される過程に基づくモデル化により，各特徴を学習する．DDSP Autoencoder [9] は，本研究の提案手法と同様に，音響信号データから，disentangle された特徴を自己教師あり学習し，各特徴を加工できる手法である．本研究の提案手法も，チャネル特徴やソース特徴などを別々に推定するため，これらを独立に制御することにより 3 節に示すチャネルマッチングや劣化音声を用いた音声変換などの用途に適用可能である．DDSP Autoencoder では，基本的に正弦波ボコーダで合成可能な音声や，チャネル成分として残響のみを想定しているが，本研究の提案手法はより表現力の高い波形合成モジュールおよびチャネルモジュールを用いている．

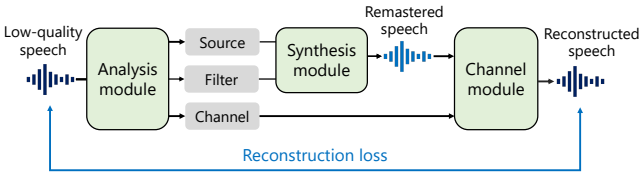


図 1 提案する自己教師ありニューラル音声復元

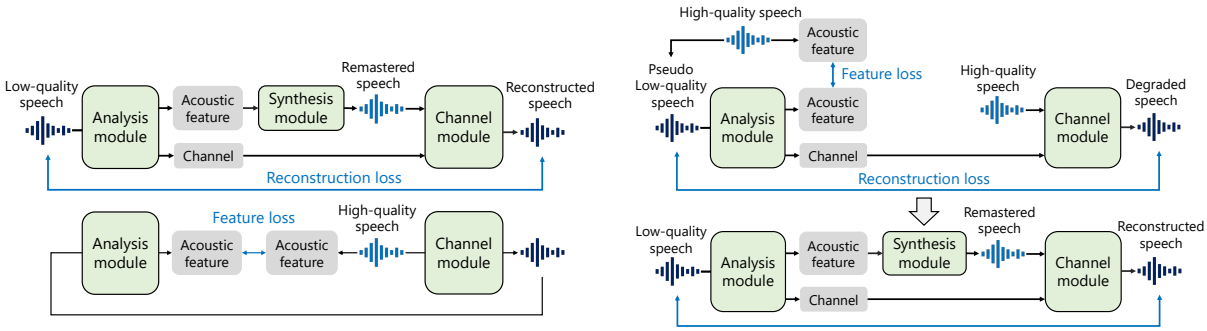
3. 提案手法

本節では，提案する自己教師あり学習によるニューラル音声復元手法についての基本的な枠組みとその学習手法を述べる．

3.1 自己教師あり学習によるニューラル音声復元

音声復元モデルを提案する前に，まず劣化音声の録音過程を整理する．まず，ソース・フィルタ過程を経て高品質音声が口外に放射される．このソース・フィルタ過程は時変であり，その特徴はメルケプストラム・F0 といったボコーダ特徴やメルスペクトログラムで表現可能である．その後，この音声に録音機器由来の歪みが付与され，最終的な録音音声となる．この歪み (チャネル特徴) は，録音機器が変わらなければ時不変であり，音声波形に対する線形変換とは限らない．

提案する音声復元モデルは，この録音過程を模擬するデコーダと，このデコーダを駆動するための特徴を抽出するエンコーダから構成される．図 1 に示すように，劣化音声から復元音声の音声特徴量および録音機器由来のチャネル特徴を抽出する分析モジュール，音声特徴量から復元音声を生成する合成モジュール，復元音声にチャネル特徴を付与するチャネルモジュールからなる．ここで，劣化音声の音声波形を $\boldsymbol{x}_{\mathrm{low}}$，$\boldsymbol{x}_{\mathrm{low}}$ を音声分析して得られる音響特徴量系列を $\boldsymbol{y}_{\mathrm{low}}$ とする．分析モジュールは 2 次元畳み込みに基づく U-Net 型のモデルであり，$\boldsymbol{y}_{\mathrm{low}}$ から復元音声の音響特徴量系列 $\hat{\boldsymbol{z}}_{\mathrm{rem}}$ を推定する．さらに，この UNet の



(a) Dual-learning for analysis and channel consistency

(b) Pretraining with pseudo low-quality speech

図 2 (a) 双方向学習法と (b) 擬似データ生成による事前学習法

中間特徴量を別の 2 次元畳み込みに基づくモデルに入力し，最終的に時不変なチャネル特徴 $\boldsymbol{c}$ を推定する．合成モジュールには，非自己回帰型ニューラルボコーダである HiFi-GAN [10] を用いており，$\hat{\boldsymbol{z}}_{\mathrm{rem}}$ を入力として復元音声の音声波形 $\hat{\boldsymbol{x}}_{\mathrm{rem}}$ を生成する．ここで，学習の安定化のため，別の高品質音声コーパスでニューラルボコーダを学習しておき，パラメータを固定する．チャネルモジュールには，復元音声 $\hat{\boldsymbol{x}}_{\mathrm{rem}}$ と，分析部で抽出されたチャネル特徴 $\boldsymbol{c}$ が入力され，劣化音声の音声波形 $\hat{\boldsymbol{x}}_{\mathrm{low}}$ を推定する．この時，モデルの入力となる劣化音声 $\boldsymbol{x}_{\mathrm{low}}$ と，チャネルモジュールの出力として推定された劣化音声 $\hat{\boldsymbol{x}}_{\mathrm{low}}$ との再構成誤差を最小化するようにモデルを学習する．再構成誤差関数として，以下のように線形項と log 項を持つ multi-resolution spectrogram loss [9] を定義する．

$$\mathcal{L}_{\mathrm{recons}} = \sum_{i} \{ ||\boldsymbol{s}_i - \hat{\boldsymbol{s}}_i||_1 + \alpha || \log \boldsymbol{s}_i - \log \hat{\boldsymbol{s}}_i ||_1 \} \quad (1)$$

ここで，$\boldsymbol{s}_i$ と $\hat{\boldsymbol{s}}_i$ は，それぞれ入力の劣化音声と再構成された劣化音声の振幅スペクトログラムであり，添字 i はフーリエ変換の窓長であり，α は log 項の重みである．

音声復元を行う際は，劣化音声の音響特徴量系列 $\boldsymbol{y}_{\mathrm{low}}$ をモデルに入力し，合成モジュールで復元音声の音声波形 $\hat{\boldsymbol{x}}_{\mathrm{rem}}$ を合成することによって復元音声を得る．また，学習した分析モジュールとチャネルモジュールから，チャネルマッチングを行うことも可能である．$\boldsymbol{x}^{\mathrm{src}}_{\mathrm{low}}$ から学習済み分析モジュールで劣化音声のチャネル特徴 $\boldsymbol{c}^{\mathrm{src}}$ を抽出する．これを用いて，別の高品質音声 $\boldsymbol{x}^{\mathrm{tar}}_{\mathrm{high}}$ をチャネルモジュールに入力し，$\hat{\boldsymbol{x}}^{\mathrm{tar}}_{\mathrm{src-low}}$ を出力する．これにより，劣化音声データのチャネル特徴を別の高品質音声に付与する．

先述の枠組みにより，高品質音声を教師データとして用いることなく，劣化音声のみから音声復元モデルを自己教師あり学習し，各特徴を操作することが可能である．しかし，一般にこの学習は不安定である．これは，モデルが劣化音声の再構成誤差のみで学習され，更に，各モジュールの表現能力が高いことに起因する．具体的には，あるモジュールが他モジュールの効果を表現してしまい，生成過程についての仮定が成立せず，分析部の出力が高品質音声の音響特徴量となる保証がない．本研究では，3.2 節に示す双方向学習法及び 3.3 節に示す事前学習法を導入する．

3.2 双方向学習法

分析部で推定される音声特徴量が高品質音声の音声特徴量となることを保証するため，図 2(a) に示す双方向学習法を導入する．ここでは，劣化音声と対応しない任意の高品質音声データを用いる．この高品質データの音声波形 $\boldsymbol{X}_{\mathrm{high}}$ をまずチャネルモジュールに通すことで，チャネル特徴が付与された劣化音声波形 $\hat{\boldsymbol{X}}_{\mathrm{low}}$ を生成する．次に，この劣化音声の音声特徴量 $\hat{\boldsymbol{Y}}_{\mathrm{low}}$ を分析部に入力し，復元音声の音響特徴量 $\hat{\boldsymbol{Z}}_{\mathrm{rem}}$ を推定する．この時，このサブタスクでの高品質データが存在するため，分析部の出力 $\hat{\boldsymbol{Z}}_{\mathrm{rem}}$ と，高品質音声から得られる音響特徴量 $\boldsymbol{Z}_{\mathrm{high}}$ の間で以下のように L1 loss を定義する．

$$\mathcal{L}_{\mathrm{feature}} = ||\boldsymbol{Z}_{\mathrm{high}} - \hat{\boldsymbol{Z}}_{\mathrm{rem}}||_1 \quad (2)$$

最終的な損失関数を，メインタスクの再構成誤差と音響特徴量間の損失関数の重み付け和として以下のように与える．

$$\mathcal{L} = (1-\beta) * \mathcal{L}_{\mathrm{recons}} + \beta * \mathcal{L}_{\mathrm{feature}} \quad (3)$$

これにより，分析部の出力が高品質音声の音声特徴量となることが保証されるが，一方で $\mathcal{L}_{\mathrm{feature}}$ の重みを大きくしすぎると，劣化音声の再構成ではなく，高品質パラメータが推定しやすい方向に学習されるため，チャネルモジュールでチャネル特徴が付与されないように学習が進む可能性がある．したがって，双方向学習の重みを適切に設定する必要がある．

3.3 擬似データ生成による事前学習

3.2 節とは別のアプローチとして，高品質データから擬似的な劣化音声データを生成し，それらをペアデータとして用いた教師あり学習により音声復元モデルの事前学習を行う．図 2(b) に本手法の概念図を示す．擬似データの生成の際は，元の高品質データに対してランダムにチャネル歪みを付与し，様々な周波数帯域・量子化ビットを持つ擬似データを生成する．高品質音声波形から得られる音声特

徴量を $\boldsymbol{Z}_{\mathrm{high}}$ とし，擬似劣化音声波形 $\boldsymbol{X}_{\mathrm{pseudo-low}}$ から分析される音声特徴量を $\boldsymbol{Z}_{\mathrm{rem}}$ とするとき，音響特徴量同士で以下のような feature loss を定義する．

$$\mathcal{L}_{\mathrm{feature}} = ||\boldsymbol{Z}_{\mathrm{high}} - \hat{\boldsymbol{Z}}_{\mathrm{rem}}||_1 \tag{4}$$

この時，式 (3) により定義される損失関数を最小化するように事前学習する．その後，自己教師あり学習の際は，擬似データを用いた事前学習によって得られた学習済みモデルの重みを初期値とし，式 (1) により定義される損失関数を最小化するように学習する．ただし，この自己教師あり学習の際も，分析モジュールの出力が高品質音声のパラメータとなることが保証されない．そこで，チャネルモジュールのみを学習させてから，次に分析モジュールのみを学習するという段階的な重み更新を行う．これにより，高品質音声から劣化音声を推定するチャネルモジュール十分に学習した上で分析モジュールを学習することで，分析モジュールが劣化音声特徴量を推定する問題を緩和する．

4. 実験的評価

4.1 実験条件

劣化音声のデータセットには，シミュレーションデータと実データの両方を用いた．シミュレーションデータには，JSUT コーパス [11] に含まれる音声データを 16 kbps の高圧縮率 MP3 に変換したものを用いた．実データには，遠野物語の読み上げ音源 [12] を用いた．これは，9 人の話し手がそれぞれ別の昔話について語ったものを収録した音源であり，収録時期は 1960–1970 年である．カセットテープに収録されたアナログ音源を，ラジオカセットレコーダー東芝 TY-CDX91-S でディジタル音源に変換した．このデータには加算性雑音が多く含まれるため，前処理として spectral gating に基づく雑音除去 [13] を事前に適用した．シミュレーションデータ・実データ共に 22.05 kHz サンプリングのデータセットである．3.3 節での教師あり事前学習用のデータセットには JVS コーパス [11] を用いた．事前学習の際は，学習時にサンプリングした各音声波形に量子化及びリサンプリングを適用して擬似データを作成した．6 ビットから 10 ビットまでのランダムな量子化ビットで mu-law 量子化を行なった後，(8k, 11.25k, 12k, 16k) Hz のランダムなサンプリング周波数でリサンプリングした．

本研究では，$\hat{\boldsymbol{z}}_{\mathrm{rem}}$ として 2 種類の特徴量を検討した．1 つは，別々のソース・フィルタ特徴であり，それぞれ F0 とメルケプストラムを用いた．メルケプストラムの次元は 41 次元である．2 つ目は，メルスペクトログラムであり，これは先行研究 [5] でも用いられている復元音声の特徴量である．メルフィルタバンクの次元数を 80 とした．

分析モジュールには，2 次元畳み込みに基づく U-Net を用いた．UNet の各 down sampling では，residual convolution block を 4 層適用した後に average pooling を適用することで，時間解像度を 1/2 ずつにした．各 residual convolution block は，カーネルサイズ 3 の 2 次元畳み込みと batch normalization 層から成っており，skip connection を持つ．また，up sampling 時は，逆畳み込みを適用することで，時間解像度を 2 倍ずつにしていき，入力特徴量と同じ時間解像度をもつ時変な高品質音声の特徴量を出力する．特徴量 $\hat{\boldsymbol{z}}_{\mathrm{rem}}$ にソース・フィルタ特徴を用いる場合は，分析モジュールで高品質音声のメルケプストラムを推定し，Harvest [14] で推定した F0 と結合した．$\hat{\boldsymbol{z}}_{\mathrm{rem}}$ にメルスペクトログラムを用いる場合は，分析モジュールの出力は高品質音声のメルスペクトログラムである．チャネルモジュールは 1 次元畳み込みに基づく U-Net 構造である．$\hat{\boldsymbol{z}}_{\mathrm{rem}}$ にソース・フィルタ成分を用いる場合は，JVS コーパスを用いて，メルケプストラムと F0 を結合した $\boldsymbol{z}_{\mathrm{rem}}$ から音声波形を出力するように学習した．ただし，neural vocoder の学習データに，事前学習での test 話者が含まれないようにした．$\hat{\boldsymbol{z}}_{\mathrm{rem}}$ にメルスペクトログラムを用いる場合，公開されている多話者での学習済みモデル[*1]を用いた．

バッチサイズは 4，エポック数は 50 である．3.3 節の学習を行う際は，チャネルモジュールのみを 25 epoch 学習させてから，分析モジュールのみを 25 epoch 学習させた．最適化関数には Adam を用い，初期状態での学習率は 0.001 で，validation loss が 2epoch に渡って下がらなければ学習率を 0.5 倍するスケジューリングを適用した．

4.2 比較手法

提案手法の有効性を検証するため，合計 8 個の比較手法を設定した．実データ以外の評価で用いた真の高品質音声を (1) **Ground-truth** と表記し，入力の劣化音声を (2) **Input (raw)** とし，入力の劣化音声からメルケプストラム・F0 を抽出して HiFi-GAN で再合成した音声を (3) **Input (synthesized)** と表記する．また，**MelSpec** として特徴量 $\hat{\boldsymbol{z}}_{\mathrm{rem}}$ にメルスペクトログラムを用いる手法を構成し，**SourceFilter** としてメルケプストラムおよび F0 を用いる手法を設定した．特徴量 $\hat{\boldsymbol{z}}_{\mathrm{rem}}$ にメルスペクトログラムを用いる手法のうち，擬似データで教師あり事前学習したモデルを適用する手法を (4) **MelSpec (SL-adaptation)** として設定した．この手法は，先行研究 [5] と同様の教師あり学習に基づく手法を異なるドメインの劣化音声データに適用した場合を評価するために設定した．さらに，3.2 節の手法を (5) **MelSpec (SSL-dual)**，3.3 節の手法を (6) **MelSpec (SSL-pretrain)** として設定した．これらの表記方法は，**SourceFilter** についても同様であり，(7) **SourceFilter (SL-adaptation)**，(8) **SourceFilter (SSL-dual)**，(9) **SourceFilter (SSL-pretrain)** のそれぞれを設定した．(5), (6), (8), (9) が提案手法である．

[*1] https://github.com/jik876/hifi-gan

表 1　教師あり事前学習での主観評価結果

	MOS
Ground-truth	4.49 ± 0.092
Input (raw)	1.38 ± 0.074
MelSpec (SL)	3.27 ± 0.11
SourceFilter (SL)	3.76 ± 0.12

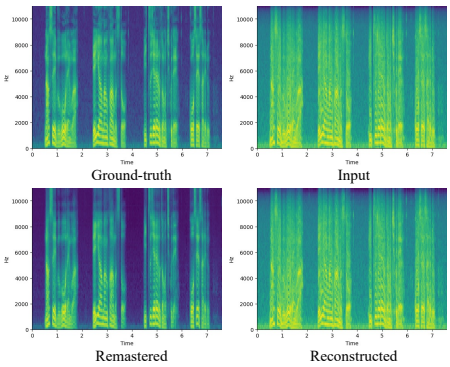


図 3　教師あり事前学習時のメルスペクトログラム

表 2　シミュレーションデータでの客観評価・主観評価結果

	MCD	MOS	SMOS
Ground-truth	-	4.59 ± 0.080	4.76 ± 0.064
Input (raw)	17.63	3.37 ± 0.097	3.43 ± 0.100
Input (synthesized)	12.59	2.56 ± 0.094	2.60 ± 0.090
MelSpec (SL-adaptation)	11.85	2.21 ± 0.099	2.85 ± 0.111
MelSpec (SSL-pretrain)	12.22	3.33 ± 0.095	3.38 ± 0.093
MelSpec (SSL-dual)	13.90	3.30 ± 0.092	3.37 ± 0.099
SourceFilter (SL-adaptation)	11.41	1.72 ± 0.095	1.90 ± 0.093
SourceFilter (SSL-pretrain)	8.96	2.92 ± 0.097	2.95 ± 0.106
SourceFilter (SSL-dual)	9.28	3.13 ± 0.089	3.18 ± 0.098

4.3　教師あり事前学習の評価

MP3 データ・実データの評価の前に，教師あり事前学習について検討した．主観評価として，Mean Opinion Score (MOS) による評価を行なった．クラウドソーシングで 60 人の評価者が参加し，音声サンプルの音質についての評価を行なった．結果を表 1 に示す．ただし，この教師あり手法の評価では，$\hat{\boldsymbol{z}}_{\text{rem}}$ にメルケプストラム・F0 を用いる手法を **SourceFilter (SL)** とし，メルスペクトログラムを用いる手法を **MelSpec (SL)** と表記する．教師ありの条件では，$\hat{\boldsymbol{z}}_{\text{rem}}$ にボコーダ特徴量・メルスペクトログラムを用いた両方の場合で，本手法のモデルで “Input (raw)” より有意に音質の高い音声に復元できていることがわかる．また，“Melspec” よりも “SourceFilter” が有意に高い性能を示している要因としては，$\mathcal{L}_{\text{feature}}$ の重みなど複数の要因が考えられるが，今後の調査が必要である．図 3 のメルスペクトログラムの可視化結果を示す．“Remastered” は復元音声，“Reconstructed” は再構成された音声を表す．結果として，ground-truth に近い歪みの小さな音声が復元できていることが見て取れる．また，再構成された劣化音声のスペクトログラムが，入力の劣化音声のスペクトログラムと非常に類似していることが確認できる．

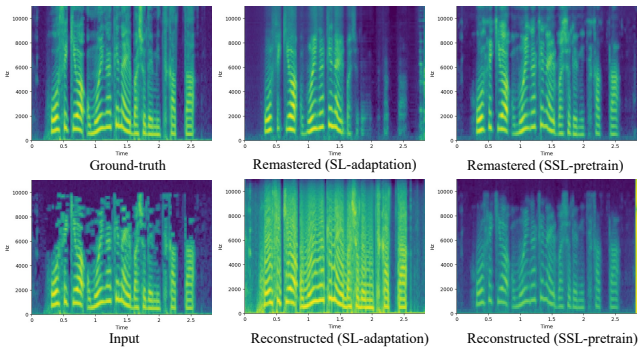


図 4　自己教師あり学習時のメルスペクトログラム可視化図

4.4　シミュレーションデータでの評価

真の高品質音声がある状況での評価を実施するため，MP3 圧縮によるシミュレーションデータを用いた評価を行なった．実験結果を表 2 に示す．まず客観評価として，“Ground-truth” と各手法での音声とのメルケプストラム歪み (MCD) [15] を算出した．このメルケストラムの次元は 24 次元とした．結果として，比較した音声復元手法の MCD は，いずれも “Input (raw)” の MCD よりも小さい値を示していることが確認できる．

また，主観評価実験として，音質に関する MOS テストと，話者類似性に関する SMOS (Similarity MOS) テストを実施した．いずれの評価でもクラウドソーシング経由で 60 人の聴取者が参加し，評価を行なった．結果を見ると，“MelSpec” および “SourceFilter” 共に，教師あり学習モデルを適用した場合よりも，自己教師あり学習した場合の方が MOS・SMOS ともに向上することがわかった．4.3 節に示したように，教師あり手法では擬似データに対しては高品質に音声復元できていた一方で，学習済みモデルをシミュレーションデータに適用すると大きく性能が下がることが確認できる．このことから，シミュレーションデータで自己教師あり学習を行うことの有効性が示唆される．また，“SSL-pretrain” と “SSL-dual” を比較すると，ほぼ全てのケースで両者の性能に有意差はないが，SMOS に関しては，“SourceFilter (SSL-pretrain)” よりも “SourceFilter (SSL-dual)” が優位に高いスコアを示した．“SSL-dual” では，分析モジュールに対する条件付けを行うため，より ground-truth 音声に近い音響特徴量を推定できる可能性が示唆される．さらに，提案手法のスコアと “Input (raw)” のスコアを比較すると，有意差がないか，場合によっては提案手法のスコアが優位に下がることが確認できる．一方で，各提案手法と “Input (synthesized)” を比較すると，MOS・SMOS 共に提案手法が有意に高いことが確認できる．このことから，提案手法により高品質な音響特徴量が推定できている一方で，合成モジュールが復元音声品質の低下の原因となっていることが示唆される．

図 4 は，復元音声および再構成音声のメルスペクトログラムである．“SourceFilter (SL-adaptation)” では，復元音

表 3　実データでの主観評価結果

	MOS
Input (raw)	2.71 ± 0.124
Input (syntheized)	2.40 ± 0.122
MelSpec (SL-adaptation)	2.08 ± 0.111
MelSpec (SSL-pretrain)	2.66 ± 0.123
MelSpec (SSL-dual)	2.63 ± 0.120
SourceFilter (SL-adaptation)	1.57 ± 0.090
SourceFilter (SSL-pretrain)	2.34 ± 0.117
SourceFilter (SSL-dual)	2.43 ± 0.122

声スペクトルの一部が欠損している一方で，“SourceFilter (SSL-step)” では，より “Ground-truth” に近いスペクトルを推定できていることが見て取れる．また，自己教師あり学習を行うことで，再構成音声のスペクトルが入力劣化音声により類似することがわかる．

4.5　実データでの評価

実データで音質に関する MOS テストを実施した．結果を表 3 に示す．本評価で用いた実データは，複数話者が混合した合計 20 分程度のデータであるにもかかわらず，教師あり学習によるモデルを適用した場合よりも，実データで自己教師あり学習した場合の方が高い MOS を示すことが確認できる．また，提案手法のうち，“MelSpec” の結果を見ると，“Input (synthesized)” よりも有意に高い MOS を示しており，高品質な特徴量を推定できていることがわかる．一方で，“SourceFilter” のケースでは，“Input (synthesized)” に対して有意差はなく，よりデータ量を増やす必要性が示唆される．

4.6　チャネルマッチングの検討

学習データから得たチャネル特徴を，別の高品質音声データに付与する実験を行った．学習データは 4.4 節で用いたシミュレーションデータであり，MP3 変換による聴こえの歪みを提案手法がどの程度再現できるかを評価した．提案手法による音声（“Chmatch”）は，JVS コーパスの高品質音声（“High-Quality”）にシミュレーションデータのチャネル特徴を付与したものである．併せて，同高品質音声を実際に MP3 変換して得られるリファレンス音声（“Reference”），同高品質音声の振幅スペクトルに，シミュレーションデータの時間平均振幅スペクトルをかけた音声（“Baseline”）を用意した．これらの音声の聴こえが，シミュレーションデータの聴こえにどの程度類似するかを，5 段階 SMOS テストで評価した．評価人数は 60 人である．

結果的に，提案手法でチャネルマッチングした場合は，高品質音声よりも有意に高い SMOS を示しており，シミュレーションデータにより近い音質に変換できていることがわかる．“Chmatch” と “Baseline” との間に SMOS の有意差はなく，また，この両者の間で XAB テストによる相対評価を実施した場合も有意差は確認できなかった．この結果より，チャネルマッチングによって Baseline 手法と同程度にチャネル特徴の付与ができることがわかるが，今後より精緻なチャネル特徴の付与が課題となる．

表 4　チャネルマッチングの主観評価結果

	SMOS
Reference	3.40 ± 0.127
High-Quality	2.03 ± 0.125
Baseline	2.53 ± 0.124
Chmatch	2.54 ± 0.116

5.　おわりに

本研究では，自己教師あり学習に基づく音声復元手法を提案した．今後の課題として，合成モジュールの品質改善や，チャネル特徴をより精緻に表現することが挙げられる．

謝辞　本研究は，JSPS 科研費 21H04900，総務省 SCOPE (受付番号 182103104) の委託を受けた．

参考文献

[1] O. Ernst et al., “Speech dereverberation using fully convolutional networks,” in *Proc. EUSIPCO*, Rome, Italy, Sep. 2018, pp. 390–394.
[2] P. Závika et al., “A survey and an extensive evaluation of popular audio declipping methods,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, pp. 5–24, 2021.
[3] V. Kuleshov et al., “Audio super resolution using neural networks,” *arXiv*, vol. abs/1708.00853, 2017.
[4] K. Han et al., “Learning spectral mapping for speech dereverberation and denoising,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.
[5] H. Liu et al., “Voicefixer: Toward general speech restoration with neural vocoder,” *arXiv*, vol. abs/2109.13731, 2021.
[6] A. Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv*, vol. abs/2006.11477, 2020.
[7] W.-N. Hsu et al., “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” 2021.
[8] C. Yi et al., “Applying wav2vec2.0 to speech recognition in various low-resource languages,” *arXiv*, vol. abs/2012.12121, 2020.
[9] J. Engel et al., “Ddsp: Differentiable digital signal processing,” *arXiv*, vol. abs/2001.04643, 2020.
[10] J. Kong et al., “HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis,” *arXiv*, vol. abs/2010.05646, 2020.
[11] S. Takamichi et al., “Jsut and jvs: free japanese voice corpora for accelerating speech synthesis research,” *Acoustical Science and Technology*, vol. 41, pp. 761–768, 2020.
[12] 佐々木徳夫, **遠野の昔話 / 佐々木徳夫編**. 桜楓社, 1985.
[13] T. Sainburg, “timsainb/noisereduce: v1.0,” June 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139
[14] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” in *INTERSPEECH*, 2017, pp. 2321–2325.
[15] R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in *Proc. PACRIM*, 1993, pp. 125–128.