Vol.2018-SLP-123 No.1
2018/7/26

非線形ひずみ除去のための敵対的denoising autoencoder

俵 直弘[1]　小林 哲則[1]　藤枝 大[2]　片桐 一浩[2]　矢頭 隆[2]　小川 哲司[1]

概要：敵対的 denoising autoencoder (DAE) を用いて非線形ひずみを補正する手法を提案する．時間・周波数マスキングは目的音源と妨害音源を高精度に分離できるが，非線形信号処理特有の耳障りなひずみが生じることが知られている．そこで，シングルチャネル音声を対象とした音声強調において高い性能を達成している敵対的 DAE を用いて，非線形ひずみを含む音声からひずみを含まないクリーンな音声へのマッピングを学習することで，非線形ひずみを低減するフィルタを構築することを試みる．このとき，時間・周波数マスキングにより得られた信号では，妨害音源成分のみならず目的音源成分も抑圧されており，ひずみの低減のためには消失した目的音源成分を復元することが必要となる．そこで，音源分離前の観測信号や妨害音源情報を補助情報として敵対的 DAE に入力することで，消失した目的音源成分の復元を行うことを試みる．マルチチャネル音源分離実験により，提案するポストフィルタの有効性を評価したところ，時間・周波数マスキングの出力信号の非線形ひずみが低減され，音質が改善されることが示された．

キーワード：非線形ひずみ除去， 敵対的生成ネットワーク， 時間・周波数マスキング， 音声強調

Adversarial denoising autoencoder for non-linear distortion reduction

Tawara Naohiro[1]　Kobayashi Tetsunori[1]　Fujieda Masaru[2]　Katagiri Kazuhiro[2]
Yazu Takashi[2]　Ogawa Tetsuji[1]

Abstract: A novel post-filtering method using generative adversarial networks (GANs) is proposed to correct the effect of a nonlinear distortion caused by time-frequency (TF) masking. TF masking is a powerful framework for attenuating interfering sounds, but it can yield an unpleasant distortion of speech (e.g., a musical noise). A GAN-based autoencoder was recently shown to be effective for single-channel speech enhancement, however, using this technique for the post-processing of TF masking cannot help in nonlinear distortion reduction because some TF components are missing after TF-masking. Furthermore, the missing information is difficult to recover using an autoencoder. In order to recover such missing components, a reference signal that includes the target source components is concatenated with an enhanced signal, is then used as the input to the GAN-based autoencoder. Experimental comparisons show that the proposed post-filtering yields improvements in speech quality over TF-masking.

Keywords: Nonliner distortion reduction, adversarial generative network, frequency-time masking, speech enhancement

1. はじめに

複数音源が混合された観測信号から特定の音源成分のみを分離・強調する技術は，雑音下音声認識など多くの音声処理における前処理として欠かせない要素技術となっている [1]．このとき，最終的に得られる強調音声は，妨害音源の影響が効果的に除去されていることに加え，ひずみの少ない高い品質であることが望ましい．

音声強調技術は，線形処理に基づくアプローチと非線形処理に基づくアプローチに大別される．非線形処理に基づくアプローチの一つである時間・周波数マスキング [2] では，目的音源の周波数成分のみを通過させる非線形フィルタを観測信号に畳み込むことで妨害音源の影響を抑圧す

[1] 早稲田大学
Waseda University
[2] 沖電気工業株式会社
OKI Electric Industry Co., Ltd.

る．このような非線形フィルタは妨害音源成分を効果的に抑圧できるものの，ミュージカルノイズのような耳障りなひずみが生じることが知られている．それに対し，ケプストラム領域において時間平滑化を行うことで非線形ひずみを抑制する手法 [3] などが提案されているが，残響に似た別のひずみが生じるという問題がある．

また，観測信号から目的信号への非線形な変換を Denoising autoencoder (DAE) の枠組みで直接推定する試みがなされている．DAE に基づく手法は従来の非線形処理に基づく手法を上回る性能を達成しているが，損失関数として平均二乗誤差を用いることによる過剰な平滑化が行われたり，音声のクリッピングが発生するなどの問題が指摘されている [4], [5], [6]．これらの問題を解決するため，敵対的学習 [7] により，雑音を含まない音声信号と区別できないような高音質の信号を生成するように DAE を学習することで，自然な強調音声を生成する手法が提案されている [8], [9]．特に，敵対的 DAE により，強調音声波形を時間領域で end-to-end に生成する Speech enhancement generative adversarial network (SEGAN) は，高品質な強調音声を生成できることが示されている [8]．また，時間周波数領域において SEGAN を適用することで，残響環境下における音声認識性能を改善できることが示されている [9]．

本研究では，時間周波数マスキングと敵対的 DAE を統合することで，目的音源のひずみを低減しながら，妨害音源成分を高精度に抑圧する手法を開発することを試みる．いま，時間・周波数マスキングにより得られる音声は，目的音源成分の消失により，音声の明瞭性や音声認識性能が著しく低下する．このとき，欠落した目的音源成分を補助情報なしに復元することは困難である．本研究では，noise-aware 学習 [10] に着想を得て，この問題の解決を試みた．マイクロホンの観測信号は欠落した目的音源成分を含むことに着目し，この観測信号を補助情報として敵対的 DAE の学習に導入することで，欠落した目的音源成分を頑健に復元することを試みる．また，非線形ひずみは妨害音源に依存することが予想されるため，補助情報として妨害音源信号を与えることも試みる．

本研究の主な成果は以下の 2 点である．まず，1) SEGAN の入力として，目的音源や妨害音源情報を含んだ参照信号を与えることで，強調音声の品質を大きく改善できることを示した．これによりシングルチャネル信号を対象とした音声強調でも従来の SEGAN よりも高い性能を達成できることを示した．さらに，2) 時間・周波数マスキングにより強調した信号に参照信号を用いた SEGAN を適用することで，強調信号に含まれる非線形歪みを効果的に低減できることを示した．

本稿の構成は以下の通りである．2 では，敵対的生成ネットワークに基づく音声強調法について述べる．3 では，提案する敵対的生成ネットワークに基づくひずみ除去フィルタの構築法について述べる．4 では，マルチチャネル音声強調実験により提案法の有効性を示す．最後に，5 で，本稿のまとめと今後の課題を述べる．

2. 敵対的生成ネットワークによる音声強調

SEGAN [8] は，雑音を含む観測信号から雑音が除去された強調信号への時間領域でのマッピングを DAE により実現する．時間領域における DAE は，従来の時間・周波数領域を対象とした DAE と異なり直接信号波形を推定する．そのため，位相を考慮する必要がないという利点がある一方，音声波形の自由度が大きいため学習が困難であるという問題があった．そこで，SEGAN では敵対的学習の枠組みを導入することで，クリーンな音声信号と区別できないような高品質な音声信号を生成するように DAE を学習する．

SEGAN は生成器 G と識別器 D から構成される．生成器として以下の構造の DAE を用いる．エンコーダは窓幅を 31，ストライド幅を 2 とした 11 層の 1 次元畳み込み層により構成され，後半の層ほどチャネル数が多くなるように設計される．エンコーダに約 1 秒（厳密には 16kHZ サンプリングで 16384 サンプル）の信号を入力すると，11 層の畳み込みによるダウンサンプリングを経て，最終的に系列数 8，チャネル数 1024 の中間出力が得られる．このとき，各層からは系列数 $\times$ チャネルとして，$16384 \times 1, 8192 \times 16, 4096 \times 32, 2048 \times 32, 1024 \times 64, 512 \times 64, 256 \times 128, 128 \times 128, 64 \times 256, 32 \times 256, 16 \times 512, 8 \times 1024$ 次元の出力が得られる．エンコーダから得られた中間出力は，正規分布からサンプリングしたランダムノイズが連結された後にデコーダに入力される．デコーダはエンコーダと同一の窓幅とストライド幅を持つ 11 層の逆畳み込みにより構成され，後半の層ほどフィルタ数が少なくなるように設計されている．11 層の逆畳み込みによるアップサンプリングを経て，デコーダからは最終的に観測信号と同じ系列数 16384，チャネル数 1 の波形信号が得られる．また，各逆畳み込み層の入力として，1 つ前の逆畳み込み層の出力と共にエンコーダの対応する深さの層の出力を用いる skip connection [11] を導入することで，より細かな解像度でのデコードが可能になる．このとき，skip connection によりバックプロパゲーション時にデコーダの各層に置ける損失がエンコーダへ直接流れるため勾配消失の影響を低減できることが期待される [12]．各層における活性化関数として，parametric rectified linear unit (PReLU) [13] を用いる．生成器の損失関数として，DAE の出力波形とクリーン信号との L_1 誤差を用いる．

DAE から得られた強調信号は観測信号とともに識別器 D に入力される．識別器は，入力された信号がクリーン信号と DAE により生成された強調信号のどちらであるか

の識別を行う．識別器として，生成器のエンコーダと同じ構造の 11 層の 1 次元畳み込み層からなる畳み込みニューラルネットを用いる．ただし，活性化関数には PReLU の代わりに leaky ReLU [14] を用い，各層の出力には virtual batch normalization [15] を適用する．識別器より得られる 8×1024 次元の出力は 1×1 畳み込みにより 8×1 次元のベクトルに変換された後，全接続の線形変換により最終的に 1 次元のスカラー値へ射影される．識別器の損失関数として，入力された信号がクリーン信号であるか，または DAE により生成された強調信号であるかを示すバイナリ値と識別器による識別結果との平均二乗誤差を用いる．

学習時には以下に示す敵対的学習の枠組みにより生成器と識別器を交互に最適化する．まず識別器 D のパラメタを固定した上で，生成器 G について以下の目的関数を最小化するようにパラメタを更新することで，生成器の損失が減少するとともに識別器の損失が増大するように生成器を最適化する．

$$\mathcal{L}_{\mathrm{cGAN}}(G) = \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\mathbf{x}_c\sim p_{\mathrm{data}}(\mathbf{x}_c)}[1 - D(\mathbf{x}, G(\mathbf{z}, \mathbf{x}_c)))^2 + \lambda||\mathbf{x} - G(\mathbf{z}, \mathbf{x}_c)||_1] \quad (1)$$

ただし，$\mathbf{z}$，$\mathbf{x}_c$, $\mathbf{x}$ はそれぞれランダムノイズ，観測信号，クリーン信号で，$p_{\mathbf{z}}(\mathbf{z})$，$p_{\mathrm{data}}(\mathbf{x}_c)$ は，ノイズベクトル $\mathbf{z}$ を生成するガウス分布と観測信号 $\mathbf{x}_c$ の経験分布である．また，λ は敵対損失に対する復元損失の重みである．式（1）を最小化することで，真のクリーン信号と区別できないような強調信号を出力するように生成器が最適化される．次に，生成器 G のパラメタを固定した上で，識別器 D について以下の目的関数を最小化するようにパラメタを更新することで，識別器の平均二乗損失が小さくなるように識別器を最適化する．

$$\mathcal{L}_{\mathrm{cGAN}}(D) = \mathbb{E}_{\mathbf{x},\mathbf{x}_c\sim p_{\mathrm{data}}(\mathbf{x},\mathbf{x}_c)}[(1 - D(\mathbf{x}, \mathbf{x}_c))^2] + \mathbb{E}_{\mathbf{x},\mathbf{x}_c\sim p_{\mathrm{data}}(\mathbf{x},\mathbf{x}_c)}[(D(\mathbf{x}, G(\mathbf{x}_c)))^2] \quad (2)$$

ただし，$p_{\mathrm{data}}(\mathbf{x}, \mathbf{x}_c)$ は観測信号 $\mathbf{x}$ と対応するクリーン信号 $\mathbf{x}_c$ のペアの経験分布である．式（2）を最小化することで，真のクリーン信号と DAE による強調信号とを正しく区別できるように識別器が最適化される．以上の生成器と識別器の最適化をミニバッチ単位で交互に行うことで生成器と識別器は互いに騙すように学習され，最終的に生成器はクリーン信号と区別できない高い品質の強調音声を生成することが可能となる．

3. 参照信号を用いた SEGAN による非線形ひずみ補正

時間・周波数マスキングにより生じたひずみを SEGAN により補正する枠組みを提案する．このとき，時間・周波数マスキングにより強調された信号は，妨害音源の成分のみならず目的音源の成分も抑圧されていることに留意する必要がある．そのため，単純に SEGAN の入力として時間・周波数マスキングによる強調信号を直接入力しても失われた目的信号の情報を補完することが困難であると考えられる．そこで，失われた情報を補完するため，SEGAN の入力として，強調音源に加えて観測信号や雑音に関する情報を参照信号として与えることを考える．観測信号を参照信号として与えた場合，過度に削減された音源情報が補完できることが期待される．また，非線形ひずみは重畳雑音の性質に依存すると考えられるため，雑音に関する情報が参照信号として有効であることが期待される．そこで，非発話区間から抽出した雑音信号や，時間・周波数マスキングにより強調された強調雑音信号も参照信号として与える．図 1 に提案する参照信号を用いた SEGAN の構造を示す．また，提案する noise-aware 学習の枠組みは，従来の単チャネル信号を対象とした SEGAN においても有効であることが期待される．そこで次節では，従来の単チャネル信号を対象とした SEGAN に対し，同様の参照信号を与えた時の効果についても調査する．

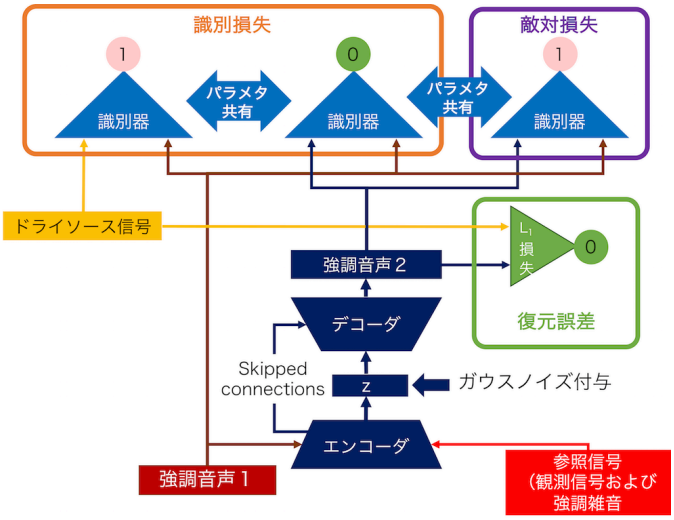


図 1 Architecture of a speech enhancement generative adversarial network (SEGAN) with an auxiliary reference input.

4. 音声強調実験

SEGAN の入力として参照信号を導入することの有効性と，SEGAN に基づく非線形ひずみの補正の有効性を評価するために，マルチチャネル音声を用いた音声強調実験を行った．

4.1 実験条件

実験に使用したマルチチャネル音声の収録環境を図 2 に示す．目的音源と妨害音源はそれぞれ 2 チャンネルマイクロホンの正面と左側面に設置した．妨害音源として，JEIDA 雑音コーパス [16] に含まれる 9 種類の雑音信号を

表 1 Models evaluated.

system	original input	auxiliary reference input
observation	noisy speech	——
SAFIA	noisy speech	——
SEGAN	noisy speech	——
SEGAN-oracle	noisy speech	matched (correct) noise
SEGAN-matched	noisy speech	matched (unsynchronized) noise
SEGAN-enhanced	noisy speech	enhanced noise (by SAFIA)
SAFIA-SEGAN	enhanced voice	——
SAFIA-SEGAN-oracle	enhanced voice	matched (correct) noise
SAFIA-SEGAN-matched	enhanced voice	matched (unsynchronized) noise
SAFIA-SEGAN-enhanced	enhanced voice	enhanced noise (by SAFIA)
SAFIA-SEGAN-obs	enhanced voice	microphone observation

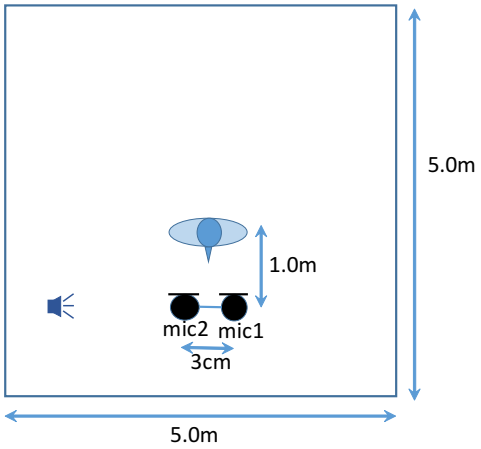


図 2 Experimental environment with two microphones, a target source, and an interference source.

用いた．表 2 に SEGAN の学習と評価に用いた雑音の種類を示す．マイクロホンの観測信号は，JNAS [17] からランダムに抽出した発話に目的音源とマイクロホン間のインパルス応答を畳み込むことで得た信号に対し，雑音信号を重畳することで作成した．学習データとして，8 種類の雑音を-10, -5, 0, 5, 10 dB で重畳して得られた計 8000 発話を用いた．同様に，テストデータとして，学習データとは異なる雑音を-10, -5, 0, 5, 10 dB で重畳して得られた計 50 発話を用いた．全ての信号について，係数 0.95 のプリエンファシスフィルタを適用した．

非線形な音声強調手法として，位相を基準とした SAFIA [18] を用いた．観測された 2 チャンネル信号の位相比が 0.1 以下の時間・周波数成分のみを抽出することで得た信号を強調音声とし，残りの成分を抽出することで得た信号を強調雑音とした．

推定した強調音声の品質を評価するため，強調音声とクリーン音声の Signal distortion rate (SDR) を算出した．SDR の算出には BSS Eval toolbox [19] を用いた．また知覚品質を評価するため，ITU standard P.862 [20] に基づき perceptual evaluation of speech quality (PESQ) を算出した．

SEGAN の学習には，学習率を 0.0002 とした RMSprop [21] を用いた．エポック数は 172，バッチサイズは 100 とし，式 (1) 内で定義した識別器に対する生成器の重み λ は 100 とした．

表 2 Interfering noise recorded. Noise signals are selected from the JEIDA noise database.

DB id	noise type	use
09	exhibition hall (booth)	training
11	exhibition hall (aisle)	training
13	station (concourse)	training
14	station (aisle)	training
18	factory (machine)	training
20	factory (metal)	training
26	street	training
28	intersection	training
30	crowd	testing

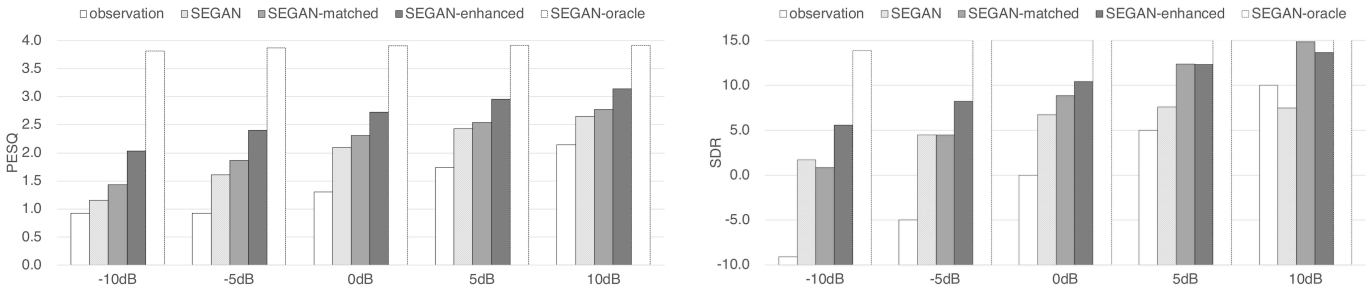
4.2 実験結果

4.2.1 補助情報の有効性

まず，SEGAN の補助的な入力として参照信号を用いることの有効性を評価した．ここでは以下の 4 つのモデルを評価した．

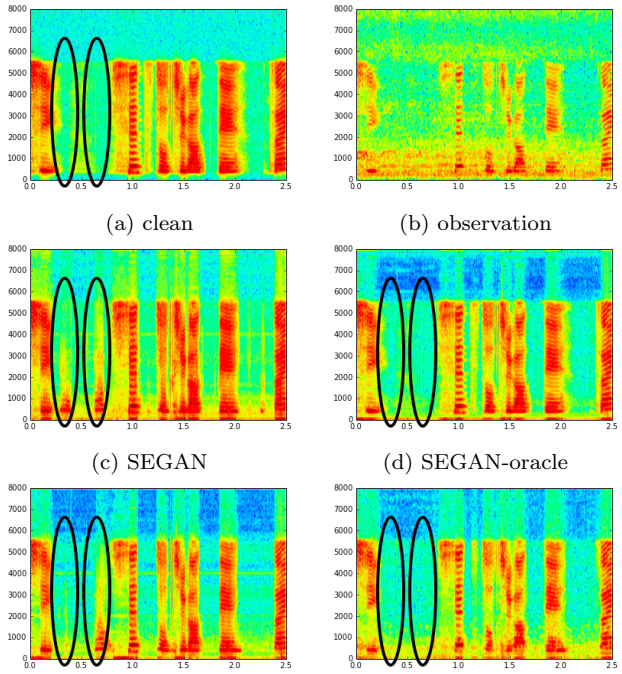
- **SEGAN:** 参照信号を用いない SEGAN [8].
- **SEGAN-oracle:** クリーン音声に重畳した雑音 (正解の雑音) でかつ時間的に同期が取れている信号を参照信号として用いた SEGAN.
- **SEGAN-matched:** 正解の雑音ではあるが, 時間同期が取れていない信号を参照信号として用いた SEGAN.
- **SEGAN-enhanced:** SAFIA により強調された雑音源の信号を参照信号として用いた SEGAN.

各手法により得られた強調音声を SDR と PESQ により評価した結果を図 3 に示す．この結果から，SEGAN により強調された信号は，観測信号に比べて SDR，PESQ ともに性能が大幅に改善されたことがわかる (**SEGAN**)．さらに，SEGAN の入力として観測信号とともに参照信号を用いた場合，いずれの参照信号を用いた場合でも性能が向上した．時間同期した正解の雑音信号を参照信号として用いた際に最高性能が得られている (**SEGAN-oracle**) が，



(a) PESQ (b) SDR

図 3 Speech enhancement performance of SEGANs with and without auxiliary reference inputs, where PESQ and SDR were averaged over 10 utterances for each condition.



(a) clean (b) observation

(c) SEGAN (d) SEGAN-oracle

(e) SEGAN-matched (e) SEGAN-enhanced

図 4 Spectrograms of (a) a clean signal and (b) an observed noise-corrupted signal, and enhanced signals obtained by (c) an original SEGAN, (d) SEGAN-oracle, (e) SEGAN-matched, and (f) SEGAN-enhanced.

観測信号と時間同期した正解の雑音信号間の差分を算出するネットワークが学習されたためと考えられる．しかし，実際には正解の雑音信号は利用できないため代替手段が必要となる．実験結果から，雑音が正確に推定できれば，時間同期が取れていなくても効果があることが見て取れる (**SEGAN-matched**)．また，時間同期が取れた雑音信号を時間・周波数マスキング等により推定することで，さらなる性能改善が得られることがわかる (**SEGAN-enhanced**)．

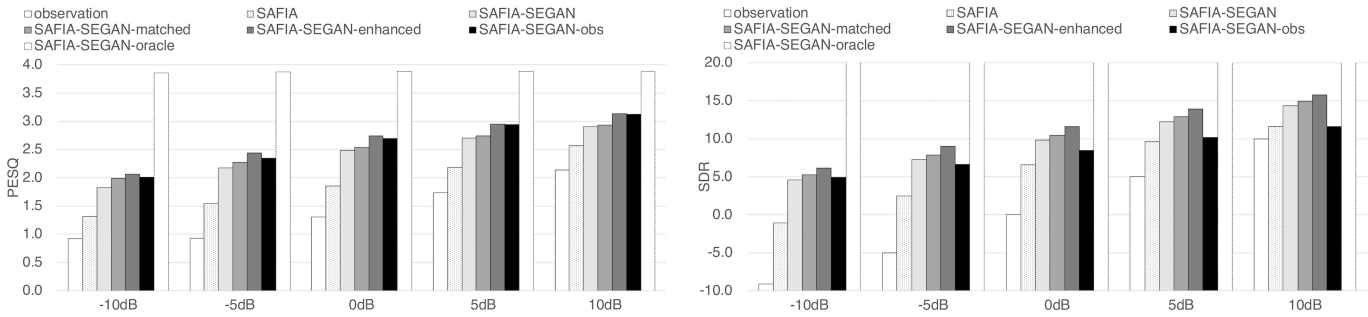
各手法により得られた強調音声のスペクトログラムを図 4 に示す．これより，SEGAN では円で囲まれた領域においてクリーン信号に存在しない冗長な成分が発生していることがわかる．一方， 時間同期が取れていない雑音信号を参照信号として用いることでこの成分が抑圧され，さらに雑音の推定値を参照信号として用いた場合では完全に消失していることがわかる．この結果から，雑音信号を参照信号として用いることで SEGAN の性能が改善できると言える．

4.2.2 非線形ひずみの補正における有効性

SAFIA により得られた強調音声に対し SEGAN を適用することで，SEGAN による非線形ひずみの補正性能を検証した．ここでは，以下の 5 つのモデルを評価した．

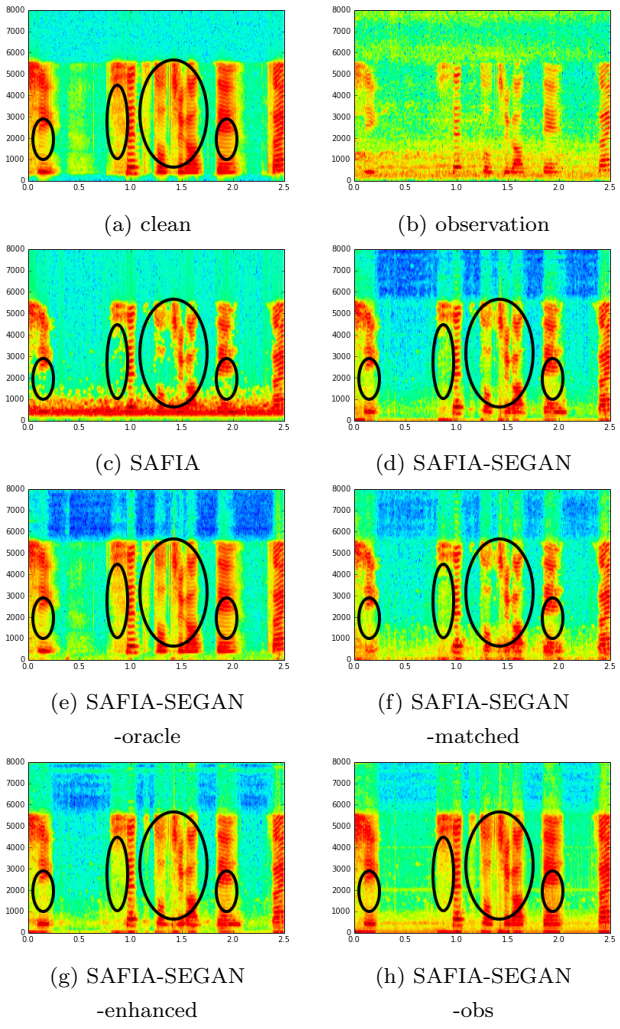
- **SAFIA-SEGAN:** SAFIA により得られた強調音声を入力とする SEGAN.
- **SAFIA-SEGAN-oracle:** クリーン音声に重畳した雑音 (正解の雑音) でかつ時間的に同期が取れている信号を参照信号として用いた SAFIA-SEGAN.
- **SAFIA-SEGAN-matched:** 正解の雑音ではあるが，時間同期が取れていない信号を参照信号とした SAFIA-SEGAN.
- **SAFIA-SEGAN-enhanced:** SAFIA により強調された雑音源の信号を参照信号とした SAFIA-SEGAN.
- **SAFIA-SEGAN-obs:** SAFIA を適用する前の観測信号を参照信号とした SAFIA-SEGAN.

図 5 に，各手法により得られた強調音声の品質を SDR と PESQ により評価した結果を示す．これより, SEGAN により非線形ひずみを補正することで，SDR と PESQ のいずれの尺度においても性能が大幅に改善されたことがわかる (**SAFIA-SEGAN**)．また，前節の実験と同様に，同期が取れている正解の雑音を参照信号として与えた際に最高性能が得られ，同期が取れていない正解の雑音を参照信号として用いた場合でも音質が若干改善した (**SAFIA-oracle**, **SAFIA-matced**)．また，SAFIA により強調された雑音を参照信号として用いることで，更なる性能改善がみられた (**SAFIA-enhanced**)．これは，非線形処理により生じた非線形ひずみは雑音に依存しており，同期の取れた雑音情報を与えることで情報の補完ができたためと考えられる．一方，SAFIA 適用前の観測信号を参照信号として用いた場合，参照信号を用いない場合と比べて PESQ が向上した一方で SDR は低下した (**SAFIA-obs**)．これは，雑



(a) PESQ (b) SDR

図 5 Speech enhancement performance from SAFIA and SAFIA-SEGAN, with and without auxiliary reference signals, where PESQ and SDR were averaged over 10 utterances for each condition.



(a) clean (b) observation

(c) SAFIA (d) SAFIA-SEGAN

(e) SAFIA-SEGAN -oracle (f) SAFIA-SEGAN -matched

(g) SAFIA-SEGAN -enhanced (h) SAFIA-SEGAN -obs

図 6 Spectrogram of (a) a clean signal and (b) an observed noise-corrupted signal, and enhanced signal obtained by (c) SAFIA, (d) SAFIA-SEGAN, (e) SAFIA-SEGAN-oracle, (f) SAFIA-SEGAN-matched, (g) SAFIA-SEGAN-enhanced, and (h) SAFIA-SEGAN-obs, respectively. Note that band-pass filter with band 300–5500 Hz was applied for TF-masking.

音と目的音声の両方が含まれる観測信号を参照信号として与えることで，失われた目的信号の時間・周波数成分を補完できたため非線形ひずみは軽減されたが，雑音に起因する成分が再生成されてしまったため，結果として SDR が低下したと考えられる．

図 6 に各手法により得られた強調音声のスペクトログラムを示す．この結果から，時間・周波数マスキングは妨害音源成分のみならず目的信号成分も不当に除去しており，ミュージカルノイズが発生していることが見て取れる．このような非線形ひずみが含まれる信号に対し SEGAN を適用することで，円で囲まれた領域のように失われた目的成分が復元されることがわかる．また，強調音源や観測信号を参照信号として与えることで，失われた目的成分のさらなる復元が可能であることが見て取れる．

5. 結論と今後の課題

本研究では，時間周波数マスキングに基づく音声強調により生じた非線形ひずみを敵対的生成ネットワークに基づく DAE により除去する手法の検討を行った．このとき，雑音情報や観測信号を参照信号として与えることで，より高い精度で音声強調を行えることを示した．

今後の課題として，時間領域での音声強調は乗法性雑音に脆弱であることが知られているため，本手法を時間周波数領域に適用することを検討している．また，本手法により強調した音声信号の品質を音声認識性能により評価することを検討している．

参考文献

[1] Barker, J., Marxer, R., Vincent, E. and Watanabe, S.: The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines, *Proc. ASRU*, IEEE, pp. 504–511 (2015).
[2] Yilmaz, O. and Rickard, S.: Blind separation of speech mixtures via time-frequency masking, *IEEE trans. on signal processing*, Vol. 52, No. 7, pp. 1830–1847 (2004).
[3] Lu, X., Tsao, Y., Matsuda, S. and Hori, C.: Speech enhancement based on deep denoising autoencoder., *Proc. INTERSPEECH*, pp. 436–440 (2013).
[4] Xu, Y., Du, J., Dai, L.-R. and Lee, C.-H.: A regression approach to speech enhancement based on deep neural

networks, *IEEE trans. on Audio, Speech and Language Processing*, Vol. 23, No. 1, pp. 7–19 (2015).

[5] Shivakumar, P. G. and Georgiou, P. G.: Perception Optimized Deep Denoising AutoEncoders for Speech Enhancement., *Proc. INTERSPEECH*, pp. 3743–3747 (2016).

[6] Kang, T. G., Shin, J. W. and Kim, N. S.: DNN-based monaural speech enhancement with temporal and spectral variations equalization, *Digital Signal Processing*, Vol. 74, pp. 102–110 (2018).

[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.: Generative adversarial nets, *Proc. NIPS 2014*, pp. 2672–2680 (2014).

[8] Pascual, S., Bonafonte, A. and Serra, J.: SEGAN: Speech enhancement generative adversarial network, *arXiv preprint arXiv:1703.09452* (2017).

[9] Donahue, C., Li, B. and Prabhavalkar, R.: Exploring Speech Enhancement with Generative Adversarial Networks for Robust Speech Recognition, *arXiv preprint arXiv:1711.05747* (2017).

[10] Xu, Y., Du, J., Dai, L.-R. and Lee, C.-H.: Dynamic noise aware training for speech enhancement based on deep neural networks, *INTERSPEECH 2014*, pp. 2670–2674 (2014).

[11] Ronneberger, O., Fischer, P. and Brox, T.: U-net: Convolutional networks for biomedical image segmentation, *International Conference on Medical image computing and computer-assisted intervention*, Springer, pp. 234–241 (2015).

[12] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proc. CVPR 2016*, pp. 770–778 (2016).

[13] He, K., Zhang, X., Ren, S. and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034 (2015).

[14] Maas, A. L., Hannun, A. Y. and Ng, A. Y.: Rectifier nonlinearities improve neural network acoustic models, *Proc. ICML*, Vol. 30, No. 1, p. 3 (2013).

[15] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X.: Improved techniques for training gans, *Proc. NIPS 2016*, pp. 2234–2242 (2016).

[16] Itahashi, S.: A noise database and Japanese common speech data corpus, *The journal of the acoustical society of Japan*, Vol. 47, No. 12, pp. 951–953 (1991).

[17] Itou, K., Yamamoto, M., Takeda, K., Takezawa, T., Matsuoka, T., Kobayashi, T., Shikano, K. and Itahashi, S.: 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).

[18] Aoki, M., Okamoto, M., Aoki, S., Matsui, H., Sakurai, T. and Kaneda, Y.: Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones, *Acoustical science and technology*, Vol. 22, No. 2, pp. 149–157 (2001).

[19] Vincent, E., Gribonval, R. and Févotte, C.: Performance measurement in blind audio source separation, *IEEE trans. on audio, speech, and language processing*, Vol. 14, No. 4, pp. 1462–1469 (2006).

[20] Recommendation, I.-T.: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, *Rec. ITU-T P. 862* (2001).

[21] Hinton, G., Srivastava, N. and Swersky, K.: 2014. Lecture 6e: Rmsprop: Divide the gradient by a running average of its recent magnitude (CSC321 Winter 2014).