一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EA2019-124,SIP2019-126,SP2019-73(2020-03)

スペクトル特徴量を利用した深層学習による歪みエフェクタの高精度モデリング

吉本　健人†　　北原　大地†　　平林　晃†

† 立命館大学情報理工学部　〒525–8577 滋賀県草津市野路東 1–1–1
E-mail: † is0254iv@ed.ritsumei.ac.jp, d-kita@fc.ritsumei.ac.jp, akirahrb@media.ritsumei.ac.jp

あらまし　歪みエフェクタのモデリングを深層学習を用いて高精度に行う手法を提案する．WaveNet を用いた従来手法では，時間信号の誤差を損失関数に用いて学習を行っていたが，高周波成分が十分に再現されていなかった．そこで本研究では，損失関数にスペクトル特徴量の誤差を追加することで，より正確な高周波成分を再現する．スペクトル特徴量には短時間フーリエ変換およびメル周波数スペクトログラムを用いた．Ibanez 社製 SD9 を用いたシミュレーションにより，提案手法が高周波成分をより忠実に再現したモデリング音を生成できることを示す．
キーワード　歪みエフェクタ，ブラックボックスモデリング，WaveNet，損失関数，スペクトル特徴量

Deep Learning Modeling of Distortion Stomp Box Using Spectral Features

Kento YOSHIMOTO†, Daichi KITAHARA†, and Akira HIRABAYASHI†

† College of Information Science and Engineering, Ritsumeikan University
Nojihigashi 1–1–1, Kusatsu, Shiga, 525–8577 Japan
E-mail: † is0383ip@ed.ritsumei.ac.jp, d-kita@fc.ritsumei.ac.jp, akirahrb@media.ritsumei.ac.jp

Abstract　We propose a method for modeling distortion stomp box with high accuracy using a deep neural network, WaveNet. The conventional method using the WaveNet adopted the error-to-signal ratio (ESR) defined in time domain as the loss function. Then, the high-frequency components were not sufficiently reproduced. To reproduce more accurate high-frequency components, we modify the loss function by adding the error of the spectral feature. We use a short-time Fourier transform and a mel frequency spectrogram as the spectral feature. Numerical experiments using an Ibanez SD9 show that the proposed method can generate modeling sounds with more accurate high-frequency components.
Key words　Distrotion stomp box, black-box modeling, WaveNet, loss function, spectral features

1. はじめに

エレキギターを演奏する際，エフェクタやアンプの増幅回路を用いて信号を歪ませることがある．エフェクタやアンプには様々な種類があり，ギタリストやアーティストは好みの音色を作り出すために繊細に使い分ける．とりわけ，ヴィンテージと呼ばれる個体によってしか再現できない音があり，需要が大きい．あるいは既に生産が終了した製品を利用したいという需要もある．こうしたデバイスの音をソフトウェア的に再現するためのデジタルモデリング技術が研究されている [1].

デジタルモデリング技術は 2 種に大別できる．第 1 は，エフェクタの電子回路を数学的モデルに変換する方法である [1]. この手法では，回路とそれを構成する部品の全てをモデリングするために作業コストが高く，良質にデジタルモデリングされた商品は，非常に高価である．

第 2 は，エフェクタの内部構造に配慮することなく，入出力特性のみを再現するアプローチで，ブラックボックスアプローチと呼ばれている．この手法では，入力音とそれに対するエフェクタやアンプの出力音さえ取得しておけばよく，内部構造のモデリングと比較して作業コストを大幅に低減できる．本論文では後者のアプローチを採用する．

ブラックボックスアプローチの従来手法では深層学習を用いた手法が提案されている [4]．再帰型ニューラルネットワーク (Recurrent Neural Network, RNN) の一種である Long Short-Term Memory (LSTM) ネットワーク [5] を用いた歪みエフェクタのモデリング手法では，入力がクリーン音，出力が歪みモデリング音であるネットワークを学習する [6]．学習済みネットワークにクリーン音を入力すれば，歪みモデリング音が出力

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2020 by IEICE

される．LSTM を用いたネットワークでは高精度なモデリングを行えるが，再帰構造を持つので，学習に時間がかかる．

学習を高速に行うために，Damskägg らは WaveNet [7] を用いたモデリング手法を提案した [8], [9]．WaveNet では，畳み込み層を工夫することにより，計算量の増大を抑制しつつ，より広範囲の入力信号を用いて出力音を高精度に再現することができる [7]．この手法では，損失関数に目的音とモデリング音の高域強調を行った時間信号の誤差を用いている．しかし，ここで用いた高域強調フィルタの副作用で低周波成分が減衰するだけでなく, 高周波成分も依然として十分に再現されないという問題があった．

この問題を解決するために本研究では，時間領域において高周波成分を強調することなく誤差を評価し，更に周波数領域におけるパワーの誤差を損失関数に加えることにより，全帯域に対して再現誤差を小さくする手法を提案する．信号の周波数領域における表現には，短時間フーリエ変換（Short-Time Fourier Transform, STFT）のパワースペクトログラムや，メルフィルタを適用したメル周波数スペクトログラムを利用する．歪みの強いエフェクタである Ibanez 社製 SD9 を用いた数値実験により，提案法の有効性を示す．

2. WaveNet を利用した手法

WaveNet を用いた歪みエフェクタのモデリング手法について説明する．WaveNet [7] とは，波形生成のためのディープニューラルネットワーク（DNN）である．WaveNet は本来，過去 R 個のサンプルから次のサンプルを予測する自己回帰モデルである．これを文献 [8], [9] では，過去 R 個の入力を用いて 1 個の出力を予測する順伝播型ニューラルネットワーク (FNN) として用いている．通常，過去の入力を用いる際には，causal convolution が用いられることが多い．これは，図 1 に示すように，現在の出力値を計算するために，その時点および過去の入力値から計算される全ての特徴を用いて畳み込みを計算する方法である．図 1 は，畳み込みのフィルタ長が $M = 2$ の場合

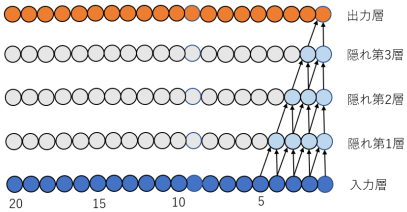


図 1: causal convolution

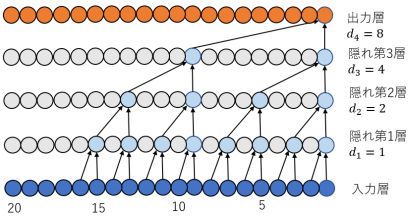


図 2: dilated causal convolution

を示している．しかしながら，全ての特徴を用いて畳み込みを行うと，入力信号の個数 R を拡大した時に，層数が増大し，積和演算に時間がかかってしまう．そこで WaveNet では，図 2 に示すように，一定間隔で間引いた特徴のみを用いて畳み込み演算を行う．この演算は dilated causal (DC) convolution と呼ばれている．まず，入力系列に対しては間引きを行わない通常（$d_1 = 1$）の畳み込み演算を行い，隠れ第 1 層の系列が求まる．信号は間引かれないが畳み込みの回数は間引かれる．この系列を $d_2 = 2$ 信号ごとに抽出した系列に対して畳み込みを行い，隠れ第 2 層の系列が求まる．この系列を $d_3 = 4$ 信号ごとに抽出した系列に対して畳み込みを行い，隠れ第 3 層の系列が求まる．このように計算を行うことにより，積和演算回数を抑制しつつ，出力計算に用いる入力信号数 R を効率よく拡大する．この値は受容野と呼ばれており，一般に

$$R = (M-1)\sum_{k=1}^{K} d_k + 1 \tag{1}$$

と表される．ここで，K は畳み込み層の数である．各層での dilation の大きさは $d_k = \{1, 2, 4, ..., 512, 1, 2, 4, ..., 512\}$ のように層が進むごとに 2 倍され，512 を越えると再び 1 に戻って繰り返される．図 2 では，$K = 4$ の各フィルタに対して dilation 係数 $d_k = \{1, 2, 4, 8\}$, および, フィルタサイズ $M = 2$ の 4 つの畳み込み層を持つ．この場合の受容野は $R = 16$ である．計算量削減のためによく利用される入力系列のダウンサンプリングでは，エイリアシングが問題となる [10]．これに対して，DC convolution を用いることにより，入力系列をダウンサンプリングすることなく，計算効率を維持できる．

DC convolution を用いて入力 $x[n]$ から出力 $\hat{y}[n]$ を計算するネットワークの詳細を図 3 に示す．このネットワークの全パラメータを θ で表し，入出力関係を f_θ で表す：

$$\hat{y}[n] = f_\theta(x[n-R+1], x[n-R+2], ..., x[n]) \tag{2}$$

まず，入力 $x[n]$（モノラル音源）に対して複数の 1x1 畳み込み層を通ることで多チャンネルの系列 $\boldsymbol{x_0}[n]$ を出力する:

$$\boldsymbol{x_0}[n] = (\boldsymbol{W_0} * x)[n] \tag{3}$$

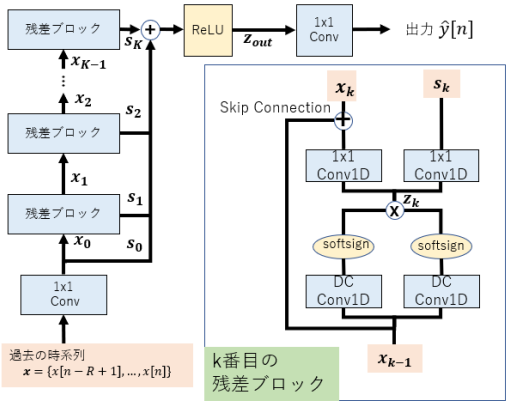


図 3: 歪みエフェクタのモデリングで用いる WaveNet [9]

ここで，$*$ は畳み込み演算である．残差ブロックは，2 つの一次元 DC convolution 層を含み，それらの出力を 2 つの活性化関数を経て統合し出力する．すなわち，k 番目の残差ブロックの DC convolution と活性化関数の出力 $\boldsymbol{z_k}(k = 1, 2 \ldots, K)$ は

$$\boldsymbol{z_k}[n] = g\left((\boldsymbol{W_{1,k}} * \boldsymbol{x_{k-1}})[n]\right) \odot g((\boldsymbol{W_{2,k}} * \boldsymbol{x_{k-1}})[n]) \quad (4)$$

と表される．ここで，$\odot$ は要素積，$\boldsymbol{W_{1,k}}$ および $\boldsymbol{W_{2,k}}$ はそれぞれ k 番目の DC convolution 層におけるフィルタサイズ $M = 3$ の畳み込みフィルタである．$g(\cdot)$ は活性化関数の softsign 関数 $g(x) = \frac{x}{1+|x|}$ である．次の残差ブロックの入力 $\boldsymbol{x_k}[n]$ は，1x1 畳み込み $\boldsymbol{W_{x,k}}$ と残差ブロックの入力 $\boldsymbol{x_{k-1}}[n]$ から求めることができる:

$$\boldsymbol{x_k}[n] = (\boldsymbol{W_{x,k}} * \boldsymbol{z_k})[n] + \boldsymbol{x_{k-1}}[n] \quad (5)$$

ここで，1x1 畳み込みフィルタ $\boldsymbol{W_{x,k}}$ は層の入力 $\boldsymbol{x_{k-1}}[n]$ と出力 $\boldsymbol{z_k}[n]$ の混合制御を行う．残差ブロックの出力 $\boldsymbol{s_k}[n]$ は 1x1 畳み込みフィルタ $\boldsymbol{W_{s,k}}$ から求めることができる:

$$\boldsymbol{s_k}[n] = (\boldsymbol{W_{s,k}} * \boldsymbol{z_k})[n] \quad (6)$$

残差ブロックの出力 $\boldsymbol{s_k}[n]$ は skip connection によって統合され，活性化関数の ReLU を用いて多チャンネル出力 $\boldsymbol{z_{out}}[n]$ を次式により求める:

$$\boldsymbol{z_{out}}[n] = \mathrm{ReLU}\left(\sum_{k=1}^{K} \boldsymbol{s_k}[n]\right) \quad (7)$$

多チャンネル出力 $\boldsymbol{z_{out}}[n]$ は，1 層の 1x1 畳み込み層のフィルタ $\boldsymbol{W_{out}}$ を通り，1 チャンネルのモデリング音 $\hat{y}[n]$ を出力する．

$$\hat{y}[n] = (\boldsymbol{W_{out}} * \boldsymbol{z_{out}})[n] \quad (8)$$

こうして出力されたモデリング音 $\hat{y}[n]$ が目的音 $y[n]$ に近くなるように学習を行う．文献 [8], [9] の手法では更に，高周波成分の再現性を向上させるために，高域強調フィルタ $H(z) = 1 - 0.95z^{-1}$ を適用した出力 $H(\hat{y})$ が $H(y)$ に近くなるように，すなわち

$$\mathrm{ESR} = \frac{\sum_{n=1}^{N} \{H(\hat{y})[n] - H(y)[n]\}^2}{\sum_{n=1}^{N} \{H(y)[n]\}^2} \quad (9)$$

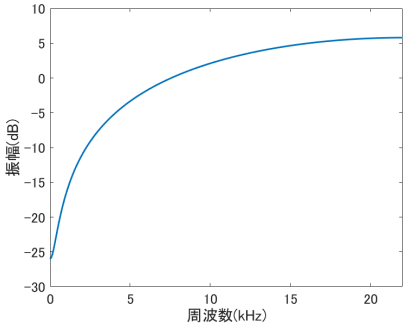


図 4: $H(z)$ の周波数特性 (サンプリング周波数 44.1kHz)

を最小化することによりネットワークを学習する．ここで，N は学習用音源のサンプル数を表す．フィルタ $H(z)$ の周波数特性を図 4 に示す．この手法では，高域強調フィルタの副作用で低周波成分が減衰してしまうだけでなく，高周波成分も依然として十分に再現されない．

3. 提 案 手 法

3.1 利用する周波数領域の特徴

一般に畳み込み層のチャンネル数やフィルタサイズ，層数を増やすことでモデルの非線形性や表現力が増加する．一方で，ネットワークのパラメータが増加し，計算量も増加する．ギターエフェクタのモデリングでは，低遅延が重要であり，畳み込み層のパラメータ数は少ない方が良い．

前章で説明した WaveNet の損失関数では，時間波形の目的音とモデリング音に対してそれぞれ高域強調フィルタを用いた際の二乗誤差を求めることで高周波成分の再現に寄与していた．しかし，高域強調フィルタを用いることで低周波成分の精度が低下しており，高周波成分の精度も不十分であった．

本手法では低周波成分の精度を低下させずに高周波成分をより忠実に再現するために，スペクトル特徴量としてパワースペクトログラム（Power Spectrogram, PS）とメル周波数スペクトログラム (Mel Frequency Spectrogram, MFS) を用いる．パワースペクトログラムは，目的音とモデリング音それぞれの時間信号 $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$ から，短時間フーリエ変換 (Short-Time Fourier Transform, STFT) により Y_{pow}, $\hat{Y}_{\mathrm{pow}}$ を求める．また，パワースペクトログラムにメルフィルタバンクを適用することで，それぞれのメル周波数スペクトログラム Y_{mel}, $\hat{Y}_{\mathrm{mel}}$ を求める．

3.2 提案する損失関数

音響信号処理の分野では，スペクトログラムの距離尺度として，主にユークリッド距離，一般化 Kullbuck–Leibler (KL) divergence，Itakura–Saito (IS) divergence などが用いられる．

- ユークリッド距離

$$\mathrm{EUC}(y||\hat{y}) = (\hat{y} - y)^2 \quad (10)$$

- 一般化 Kullbuck–Leibler (KL) divergence

$$\mathrm{KL}(y||\hat{y}) = y\log\frac{y}{\hat{y}} - (y - \hat{y}) \quad (11)$$

- Itakura–Saito (IS) divergence

$$\mathrm{IS}(y||\hat{y}) = \frac{y}{\hat{y}} - \log\frac{y}{\hat{y}} - 1 \quad (12)$$

1 次元の距離値を図 5 に示す．一般化 KL divergence と IS divergence では，ネットワーク出力 $\hat{y}$ が目的音 y より小さい場合には大きな値に，ネットワーク出力 $\hat{y}$ が目的音 y より大きい場合には小さな値になる．すなわち，出力値が大きいことは許容されるが，足りないことには敏感になる．

IS divergence は, スケールに対して不変な距離尺度である．歪みエフェクタの出力音は，スペクトルピーク以外の部分の周波数成分が増える．この部分の周波数成分やノイズ部分の誤差を IS divergence では大きく評価しすぎてしまい，スペクトル

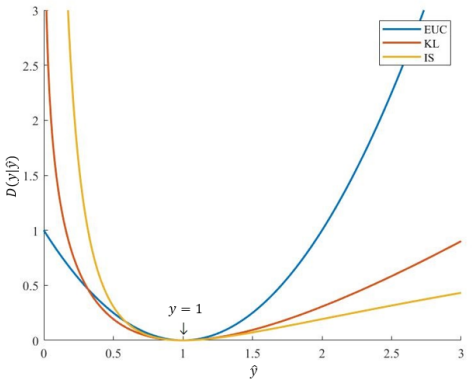


図 5: 距離値 ($y = 1$)

の特徴が適切に学習されない．

また，ユークリッド距離と一般化 KL divergence を比較実験した場合，一般化 KL divergence ではスペクトログラムの特徴をよく捉えることができた．以上より提案法では，時間信号の誤差 l_{wave} にはユークリッド距離の平均である平均二乗誤差，スペクトログラムの誤差 l_{freq} には，一般化 KL divergence を用いる．よって，最終的にネットワークの損失関数を

$$\begin{aligned}\mathrm{Loss} &= l_{\mathrm{wave}} + l_{\mathrm{freq}} \\ &= \frac{1}{N}\sum_{n=1}^{N}(\hat{y}[n] - y[n])^2 \\ &+ \frac{\lambda}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(Y_{i,j}\log\frac{Y_{i,j}}{\hat{Y}_{i,j}} - (Y_{i,j} - \hat{Y}_{i,j})\right)\end{aligned} \tag{13}$$

と定義する．重みパラメータ $\lambda > 0$ を用いて波形とスペクトル特徴量の誤差のスケール調整を行う．時間波形を重視すれば，振幅の大きい低周波成分は考慮されるが，波形ではわずかな誤差である高周波成分は考慮されにくい．スペクトル特徴量を重視すれば，一般化 KL divergence の特性により，パワーの足りない高周波成分の誤差を大きく評価する．これにより，時間信号のみの損失関数の場合よりも，高周波領域が十分に考慮されるが位相情報が考慮されないため，位相が反転した波形を出力する場合がある．第 4 節の実験ではスペクトル特徴量を重視した重みづけを行うためにパワースペクトログラムを用いる場合では $\lambda = 1$，メル周波数スペクトログラムでは $\lambda = 0.1$ とする．

3.3 学習アルゴリズム

学習アルゴリズムについて説明する．Neural Network は D 個からなる訓練データ $\boldsymbol{x}_{\mathbf{train}} = \{x^{(1)}_{\mathrm{train}}, x^{(2)}_{\mathrm{train}}, \ldots, x^{(D)}_{\mathrm{train}}\}$ を用いて最適なパラメータを学習させる．しかし，訓練データは膨大であり，全てのデータを対象とした損失関数を計算するのは現実的ではない．そこで，データの中から一部を選び出し，その一部データを全体の近似とみなしてパラメータを更新する．このような手法をミニバッチ学習と言う．学習データ $\boldsymbol{x}_{\mathbf{train}}$ から N 個ランダムに取得した j 番目の $\boldsymbol{x}_{\mathbf{MB}j} = \{x^{(1)}_{\mathrm{MB}j}, x^{(2)}_{\mathrm{MB}j}, \ldots, x^{(N)}_{\mathrm{MB}j}\}$ をミニバッチとする．

ネットワークにクリーン音 $x^{(i)}_{\mathrm{MB}j}$ を入力することで，出力であるモデリング音 $\hat{y}^{(i)}_{\mathrm{MB}j}$ を取得する．また，教師データの目的音 $y^{(i)}_{\mathrm{MB}j}$ とモデリング音 $\hat{y}^{(i)}_{\mathrm{MB}j}$ から，STFT を用いてそれぞれのパワースペクトログラム，あるいは，パワースペクトログラムにメルフィルタバンクを適用したメル周波数スペクトログラム $Y^{(i)}_{\mathrm{MB}j}$ と $\hat{Y}^{(i)}_{\mathrm{MB}j}$ を取得する．そこから，式 (13) を用いて目的音とモデリング音の誤差を求め，ミニバッチに関する平均誤差を減少させるために，最適化手法である Adam [13] を用いて重みパラメータ θ を更新する．これを学習データ全てを用いて行う．このプロセスを ‘epoch’ と呼び，これを一定回数繰り返す．このように学習したネットワークにテスト音源 $\boldsymbol{x}_{\mathbf{test}}$ を入力することで，モデリング音源 $\hat{\boldsymbol{y}}_{\mathbf{test}}$ を取得する．

4. 実験概要

4.1 学習データ

本実験では，エフェクタには歪みの強い製品である Ibanez 社製 SD9 を用いた．つまみは，全て中心部分の 12 時方向に合わせた．学習データセットに IDMT データセット [11], [12] を使用し，サンプリング周波数は 44.1 kHz とした．このデータセットからランダムに選んだギター音 2 分 30 秒，ベース音 2 分 30 秒の合計 5 分のデータを学習に用いた．これらのデータを 0.1 秒毎に切り取り，3,000 個のデータとしてネットワークの学習に用いた．また，early stopping の検証にはギターとベースそれぞれ 30 秒の合計 1 分の 600 個のデータを用いた．プログラムは Python 3.7.3 で実装した．実行環境は，OS: Windows10 Pro, CPU: Core i9-7980X，メインメモリ: 128GB，GPU: GeForce GTX1080Ti である．実験条件を表 1 に示す．

表 1: 実験条件

サンプリング周波数	44.1kHz
入力タイムステップ数	100ms（4410 samples）
出力タイムステップ数	10ms（440 samples）
窓長	23ms（1024 samples）
シフト長	5ms（256 samples）
DFT 点数	8192
MFS 周波数帯	60 Hz-22000 Hz
MFS チャネル数	300
batch size	16
最適化法	Adam Optimizer
epoch 数	1000 epoch
比較手法（損失関数）	従来法: ESR（高域強調） 提案法 1: MSE（波形）+ KL（PS） 提案法 2: MSE（波形）+ KL（MFS）

4.2 実験結果

本実験では 4 種類のテスト用音源を用いた(注1) それぞれの特徴は，音源 1 はアタックが強く音域の高い和音，音源 2 はアタックが弱く音域の低い和音，音源 3 はアタックが強く音域の低い単音，音源 4 はアタックが弱い中音域のオクターブの 2 和音である．図 6 の (a), (b) は音源 1 のクリーン音と，それ

(注1)：実際に使用した目的音とモデリング音は以下の Web ページから聞ける．
http://www.ms.is.ritsumei.ac.jp/audio.html

表 2: 各モデリング音の ESR

手法	音源 1	音源 2	音源 3	音源 4	平均
従来法	1.62%	0.69%	0.95%	0.55%	0.95%
提案法 1（PS）	1.41%	**0.60%**	0.67%	**0.38%**	0.76%
提案法 2（MFS）	**1.21%**	**0.60%**	**0.53%**	0.45%	**0.70%**

表 3: 各モデリング音の ESR (高域強調)

手法	音源 1	音源 2	音源 3	音源 4	平均
従来法	16.86%	7.29%	5.03%	7.15%	9.08%
提案法 1（PS）	11.68%	7.36%	5.19%	9.86%	8.52%
提案法 2（MFS）	**9.01%**	**4.51%**	**2.57%**	**6.65%**	**5.69%**

表 4: 各モデリング音の MSE (PS)

手法	音源 1	音源 2	音源 3	音源 4	平均
従来法	28.86	26.11	77.77	59.82	48.14
提案法 1（PS）	14.44	8.89	28.12	16.24	19.92
提案法 2（MFS）	**12.18**	**8.21**	**20.91**	**14.95**	**14.06**

を SD9 で歪ませた目的音のパワースペクトログラム，図 6 の (c), (e), (g) はそれぞれの手法で生成したモデリング音と目的音の波形の比較，図 6 の (d), (f), (h) はそれぞれの手法で生成したモデリング音のパワースペクトログラムである．従来手法では，波形の概形は再現できているが，低周波成分が減衰したり，高周波成分が十分に再現されていない部分がある．そのため振幅が足りない部分や小さなピークの部分の精度が低い．提案手法では，いずれのスペクトログラムでも低周波成分と高周波成分の両者がより高精度に再現されているため，波形の微細な特徴まで表現できていることが確認できる．スペクトル特徴量に関して，パワースペクトログラムとメル周波数スペクトログラムの場合を比較すると，パワースペクトログラムの場合では波形のピークの細かい特徴での誤差が確認できるが，メル周波数スペクトログラムの場合では，その部分が高精度に再現できていることを確認できる．ESR に高域強調フィルタを用いた従来法と，提案法である時間波形の平均二乗誤差（MSE）とスペクトログラムの一般化 KL divergence の誤差の性能比較を行う．評価指標には波形の ESR，高域強調を用いた波形の ESR，パワースペクトログラムの MSE を用いた．それぞれの結果を表 2, 3, 4 に示す．従来手法と比較して，歪みの強いモデルである Ibanez SD9 では，波形領域の ESR は提案法 1（PS）では 20.0%，提案法 2（MFS）では 26.0% 向上した．周波数領域の MSE では，提案法 1（PS）では 58.6%，提案法 2（MFS）では 70.8% 向上した．また，高域強調を行った ESR の結果から，Ibanez SD9 のような強い歪みのモデルで生成される強いパワーの高周波数成分に対しては，従来手法では再現が十分でない．一方，提案手法ではスペクトログラムの誤差を一般化 KL divergence により評価することで，高周波成分の特徴を適切に学習できていた．以上により，提案手法は時間領域においても周波数領域においても高精度なモデリングを達成することを示した．

5. おわりに

歪みエフェクタを高精度に再現するために WaveNet を利用したモデリング手法を提案した．従来の手法では，損失関数に高域強調フィルタを用いた波形領域の誤差のみを評価していた．そのため，波形の概形は再現できるが，低周波成分の減衰により部分的に振幅の誤差が大きくなることや，高周波成分の再現が不十分で音色が変化してしまう問題があった．これらの問題点を解決するため提案法では，損失関数に高域強調フィルタを用いるのではなく，スペクトル特徴量の誤差を追加した．これにより，低周波成分が再現されて十分な振幅を得られると共に，高域強調フィルタでは不十分であった高周波成分も再現し，従来法での問題点を解決することができた．数値実験により，提案法が従来法に比べてエフェクタを高精度モデリングできることを示した．

文　献

[1] D.T. Yeh, J. Abel, and J.O. Smith, “Simulation of the diode limiter in guitar distortion circuits by numerical solution of ordinary differential equations,” Proc. Int. Conf. Digital Audio Effects, pp.197–204, Bordeaux, France, Sept. 2007.

[2] F. Eichas and U. Zölzer, “Black-box modeling of distortion circuits with block-oriented models,” Proc. Int. Conf. Digital Audio Effects, pp.39–45, Brno, Czech Republic, Sept. 2016.

[3] F. Eichas, S. Möller, and U. Zölzer, “Block-oriented gray box modeling of guitar amplifiers,” Proc. Int. Conf. Digital Audio Effects, pp.5–9, Edinburgh, UK, Sept. 2017.

[4] Z. Zhang, E. Olbrych, J. Bruchalski, T.J. McCormick, and D.L. Livingston, “A vacuum-tube guitar amplifier model using long/short-term memory networks,” Proc. IEEE Southeast Conf., pp.1-5, St. Petersburg, FL, April 2018.

[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol.9, no.8, pp.1735–1780, Nov. 1997.

[6] Y. Matsunaga, N. Aoki, Y. Dobashi, and T. Yamamoto, “A digital modeling technique for distortion effect based on a machine learning approach,” Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp.1888–1892, Honolulu, HI, Nov. 2018.

[7] A. van den Oord, S. Dieleman, H. Zen, K. Shimonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WAVENET: A generative model for raw audio,” arXiv pre-print, 2016, arXiv:1609.03499.

[8] E.-P. Damskägg, L. Juvela, and V. Välimäki, “Deep learning for tube amplifier emulation,” Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP), pp.471–475, Brighton, UK, May 2019.

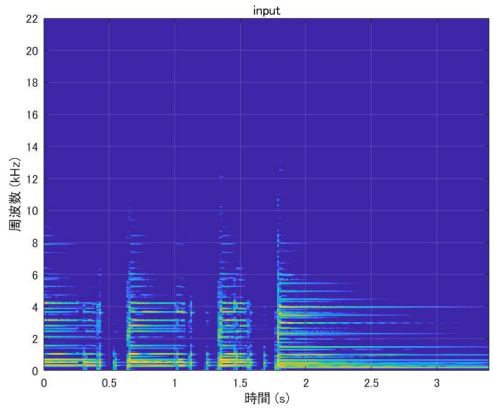
[9] E.-P. Damskägg, L. Juvela, and V. Välimäki, “Real-time modeling of audio distortion circuits with deep learning,” Proc. Sound and Music Computing Conf. (SMC-19), pp.332–339, Malaga, Spain, May 2019.

[10] 田中宏, 亀岡弘和, 金子卓弘, 北条伸克, “WaveCycleGAN2: 高品質音声合成のための時間領域ニューラルポストフィルタ,” 日本音響学会 2019 年秋季研究発表会講演論文集, pp.989–990, Sept. 2019.

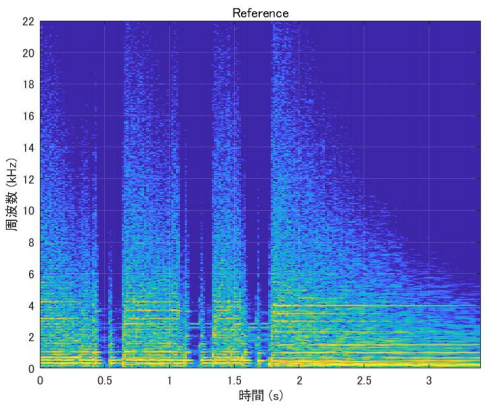
[11] `https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/guitar.html`, Jan. 20, 2020.

[12] `https://www.idmt.fraunhofer.de/en/business_units/m2d/smt/bass_lines.html`, Jan. 20, 2020.

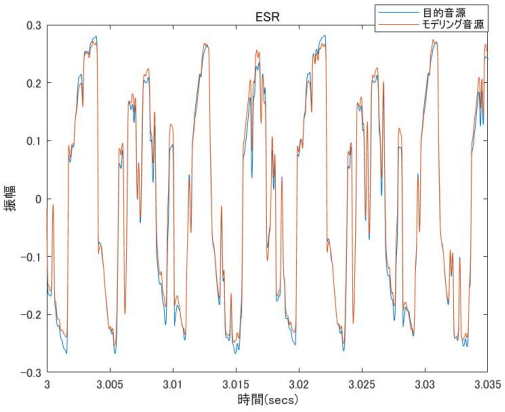
[13] D.P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” Proc. Int. Conf. Learning Representations (ICLR), 13 pages, San Diego, CA, May 2015.



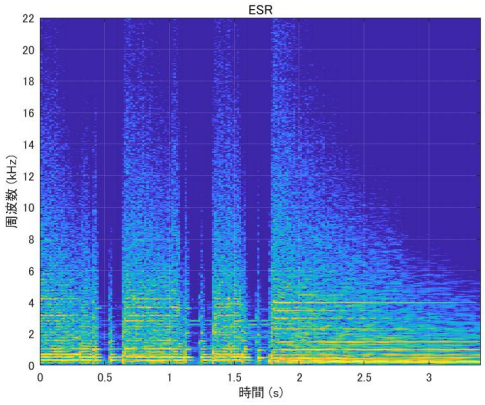
(a) 入力音源のパワースペクトログラム



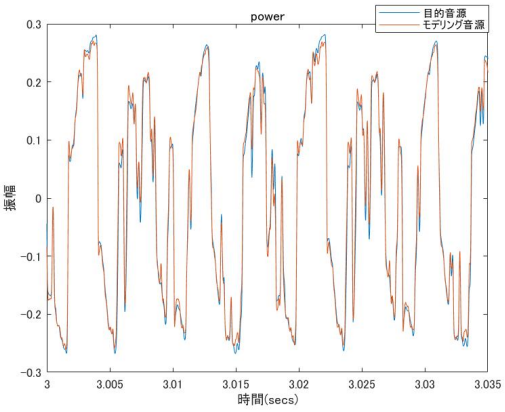
(b) 目的音源のパワースペクトログラム



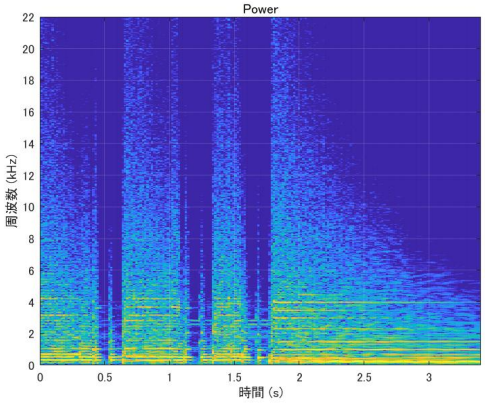
(c) 従来法 (ESR) の波形



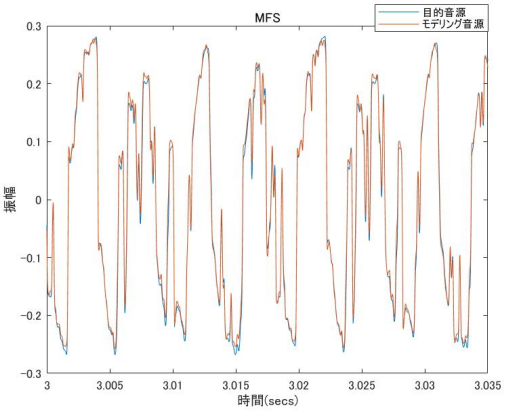
(d) 従来法 (ESR) のパワースペクトログラム



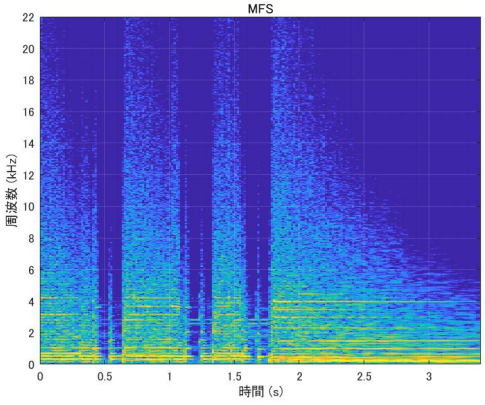
(e) 提案法 1 (PS) の波形



(f) 提案法 1 (PS) のパワースペクトログラム



(g) 提案法 2 (MFS) の波形



(h) 提案法 2 (MFS) のパワースペクトログラム

図 6: Ibanez 社製 SD9 の音源 1 に対するシミュレーションの結果