E-032

合成波形の振幅誤差を最小化する回帰型 WaveNet の提案
Proposal of Regression WaveNet to Minimize Amplitude Error of Composite Waveform

那須野 僚輔 †
Ryosuke Nasuno

荒井 秀一 †
Shuichi Arai

1. まえがき

音声信号処理では音声を周波数領域でモデリングするのが一般的だが [1][2], 近年, 音声波形を直接モデリングする WaveNet[3] が提案され, 様々な分野で利用されつつある. WaveNet は音声波形の振幅値をクラスとして扱い、ネットワークをクラス分類器として構成している。[3] で実験している text-to-speech では, シンボルであるテキストから実世界の値の分布を求めることが必要となるため, ネットワークの出力を確率で表現することが適している. WaveNet を利用した研究には信号の帯域拡張 [4] や音声強調 [5] を目的としたものがあるが、ネットワークの入力と教師が実数値であるため, 分布を求めずに音声波形サンプルの振幅値をそのまま出力することが適していると考えた. また, クラス分類では振幅値を最小化することを目的としておらず人間の聴感に悪影響を及ぼす可能性がある. [4][5] に対しては振幅の誤差を最小化する回帰モデルを用いることで人間の聴感を改善できると考えた. 本稿では、WaveNet を基とした回帰モデルアーキテクチャを提案し、音声信号の帯域拡張を例に、従来法と提案法を比較する。

2.WaveNet

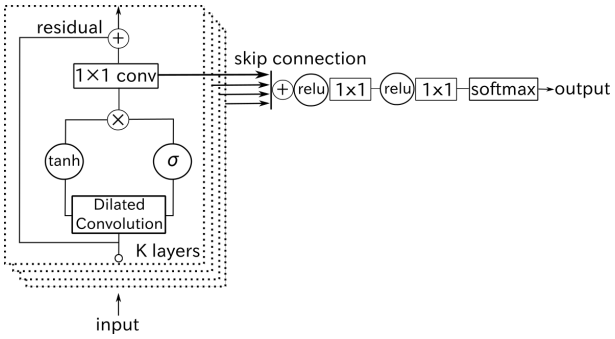

図 1: 従来の WaveNet

図 1 に本研究で用いた従来の WaveNet[3] のアーキテクチャの全体図を示す. WaveNet は図 2 に示す音声波形のサンプルを周期的に畳み込む dilated convolution を積み重ねることで構成される. 入力は dilated convolution で畳み込みをした後に式 (1) で表されるゲート付き活性化関数を通り, 入力と残差をとることで次の層への入力となる.

$$z = tanh(W_{f,k} * x) \odot \sigma(W_{g,k} * x) \tag{1}$$

†東京都市大学

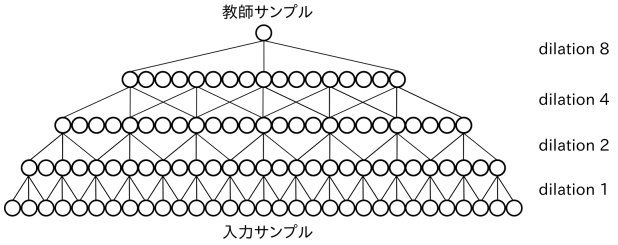

図 2: dilated convolution の概要

ただし, *は畳み込み演算を, $\odot$ は要素積演算を表し σ はシグモイド関数を表す. W は畳み込みの重み, f はフィルタで g はゲートを表し k はレイヤの番号を表す添字である. また, ゲート付き活性化関数を通した出力を各層から skip connection をして総和をとり, 最終的にソフトマックス関数を通すことにより信号波形サンプルの事後確率を出力する. 入力の音声波形サンプルを $\mathbf{x}$=$\{x_1, x_2, x_3, ..., x_n\}$ と定義したときネットワークの出力である事後確率は式 (2) と定義できる.

$$P(x) = \prod_{n=1}^{N} P(x_n | x_1, x_2, ..., x_{n-1}) \tag{2}$$

WaveNet では事後確率の対象を式 (3) で表される μ-law 変換によって 16bit の整数値で表される音声サンプルを 8bit の整数値に非線形圧縮をかけることでクラス数を 65536 クラスから 256 クラスに削減し, 事後確率を計算的に扱いやすくする.

$$f(x_n) = sign(x_n) \frac{\ln(1 + \mu|x_n|)}{\ln(1 + \mu)} \tag{3}$$

3. 回帰型 WaveNet

従来の WaveNet ではネットワークの出力を音声波形サンプルの事後確率として扱っている. 事後確率によってクラス分類することで生成された音声波形サンプルは, 教師データとの振幅の誤差を最小化せずに softmax 関数により算出された事後確率を最小化することで学習する. 複数のモデルを用いて信号波形をモデリングする場合において, それぞれのネットワークの出力をどの程度の確信度で活用するかを定義することできないため, 音声波形の振幅を回帰モデルで推定することは困難であると考えられるが, 複数モデルを用いずにモデリングすることが可能なタスクにおいては振幅の誤差を最小化する回帰モデルで学習を行うことが適していると考えられる. 図 3 に提案する振幅の誤差を最

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

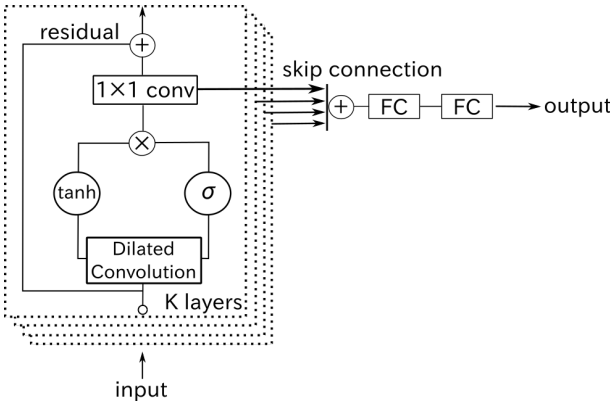

図 3: 提案する WaveNet

小化することを目的とした回帰モデルの WaveNet を示す．ただし FC(Full Connect) は全結合層を表す．回帰モデルでは出力層に全結合層を用いることが一般的であり [6][7], dilated convolution で抽出した特徴を効率的に扱うことができると考えたため出力層に全結合層を採用した．また提案手法の WaveNet を表 1 として構成した．

表 1: ネットワークの構成

	従来手法	提案手法
層数	16	16
最大 dilation	128	128
residual channels	256	256
skip out channels	512	128

4. 実験と評価

本章では従来の WaveNet と提案する WaveNet の性能比較をするために音声信号の帯域拡張のタスクに対して実験する．

4.1. 実験条件

本実験では, VCTK-corpus[8] に含まれる 6ヶ国語の合計 18 人のデータを使用した．サンプリングレートは 16kHz で学習に 3512 発話を用い, 残りの 192 発話を評価に用いた．また帯域拡張の実験方法を図 4 に示す．

4.2. 実験結果

定量評価に信号の歪みを測る SNR(signal to noise rate)[9], LSD(log spectral distance)[10] を用いた．評価結果を表 2 に示す．音声信号の帯域拡張の実験において, 従来の音声波形の振幅値をクラス分類する WaveNet より提案手法が SNR において 1.89[dB] 向上し LSD においては 2.55[dB] 向上した．また, 6 人の被験者にプリファレンステストを実施した．20 種類のテストセットを用意し AB の順番入れ替えた 40 セットに対して実験した．結果として分類モデルが 36.25%, 回帰モデルが 63.75%であった．

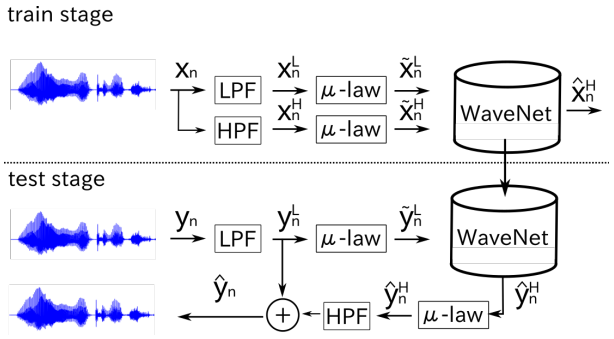

図 4: 実験方法

表 2: 実験結果

	従来の WaveNet	提案した WaveNet
SNR[dB]	17.59	19.38
LSD[dB]	12.02	9.47

5. 結論

本稿では, 音声波形の振幅値の誤差を最小化する回帰型 WaveNet の提案をした．また, 従来の音声波形の振幅値をクラス分類する WaveNet と提案手法の性能を比較するために音声信号の帯域拡張の実験をした．原音に対する合成音声の歪みを SNR と LSD の二つの指標により評価した．実験結果より音声信号の帯域拡張において, 従来の音声波形の振幅値をクラス分類する WaveNet より提案した音声波形の振幅値の誤差を最小化する回帰型 WaveNet が適していることを示した．

参考文献

[1] Volodymyr Kuleshov et al. ”audio super resolution using neural nets”. *ICLR workshop*, 2017.

[2] zu-Wei Fu et al. ”snr-aware convolutional neural network modeling for speech enhancement”. *InterSpeech*, 2016.

[3] A.van et al. ”wavenet:a generative model for raw audio”. *http://arxiv.org/abs/1609.03499*, 2016.

[4] Y.Gu et al. ”waveform modeling using stacked dilated convolutional neural networks for speech bandwidth extension”. *InterSpeech*, 2017.

[5] kaizhi Qian et al. ”speech enhancement using bayesian wavenet”. *InterSpeech*, 2017.

[6] daniel De Tone et al. ”deep image homography estimation”. *RSS Workshop on Limits and Potentials of Deep Learning in Robotics*, 2016.

[7] Peng Zhang et al. ”audio source separation from a monaural mixture using convolutional neural network in the time domain”. *ISNN 2017. LNCS, vol. 10262, pp. 388–395. Springer, Cham*, 2017.

[8] C.Veaux et al. ”cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,”. *http://dx.doi.org/10.7488/ds/1994*, 2017.

[9] T.Hayashi et al. ”an investigation of multi-speaker training for wavenet vocoder”. *Proc. ASRU*, 2017.

[10] Jun Du et al. ”a speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions”. *InterSpeech*, 2008.

Copyright © 2018 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.