論文 / 著書情報
Article / Book Information

論題(和文)	音声強調のための拡散モデルにおける計算量の削減
Title(English)	
著者(和文)	西 悠希, 岩野 公司, 篠田 浩一
Authors(English)	Yuki Nishi, Koichi Shinoda
出典(和文)	日本音響学会第151回(2024年春季)研究発表会 講演論文集, , ,
Citation(English)	, , ,
発行日 / Pub. date	2024, 3

音声強調のための拡散モデルにおける計算量の削減 *

○西悠希 (東京工業大), 岩野公司 (東京都市大), 篠田浩一 (東京工業大)

1 まえがき

雑音除去を目的とした音声強調で近年はニューラルネットワークを用いた技術が広く使われるようになり, より高い精度が実現されている [1]. 一方で, 近年は拡散モデルと呼ばれる手法が, データの生成に関し高精度かつ訓練が安定しているということで注目されている. 音声強調においても, この拡散モデルを用いる研究がなされている [2,3]. しかしこの手法は, 生成段階で繰り返しデータをネットワークに通す必要があるために計算量が多いという欠点を持つ.

本研究は, 拡散モデルにおける生成段階の計算コストを, なるべく精度を落とさずに削減することを目指す. AutoEncoder により, 入力される音声データを圧縮し, その中で拡散モデルの生成処理プロセスを実行することで, 計算コストの削減を精度を落とさずに実現した.

2 拡散モデル

2.1 拡散モデルの概要

生成モデルによく使われる手法に, 拡散モデルがある. 拡散モデルとは, 対象とするデータがノイズにより拡散していく過程をモデル化したもので, その逆過程を辿ることで, ノイズが混入したデータから元データを生成する方法である. この逆過程にガウスノイズなどを初期のノイズとして入力することで, ノイズがない何らかの意味のあるデータを出力させることができる. 本稿ではスコアを用いた拡散モデル [4] を利用する. これは, 確率的微分方程式 (Stochastic differential equation, 以下 SDE) を元に定式化がなされる. まず, ノイズの拡散方向である順方向を以下のように定式化する.

$$\mathrm{d}x_t = f(x_t)\mathrm{d}t + g(t)\mathrm{d}w, \quad 0 \leqq t \leqq T \tag{1}$$

x_t は t におけるデータの状態で, x_0 がノイズが付加されていないデータに相当する. f はドリフト係数と呼ばれ, ベクトルを返す関数である. g は拡散係数と呼ばれ, スカラー値を返す関数で, 加えるノイズの強さを表す. w はウィーナー過程である. f, g は, 実験設定により様々な定義がなされる. これの逆方向の SDE は, [5] により次のように定式化できる.

$$\mathrm{d}x_t = \left[-f(x_t) + g(t)^2 \nabla_{x_t} \log p_t(x_t)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{w} \tag{2}$$

$\nabla_{x_t} \log p_t(x_t)$ がスコアと呼ばれる. $p_t(x_t)$ は t における x_t の確率分布である。このスコアを推定することが, NN の目的となる. また, $\bar{w}$ はステップ t を T から 0 まで逆向きにたどった場合の標準ウィーナー過程である.

この逆過程をシミュレートする際、何度も NN にスコアを推定させるので、その反復回数分計算コストがかかることが、拡散モデルに共通する問題点と言える.

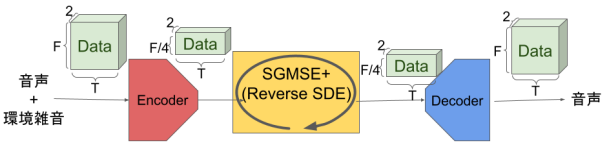

Fig. 1 圧縮率 $r = 1/4$ における提案手法の推論プロセス. F はフーリエ係数の数, T は周波数領域でのフレーム数である.

2.2 拡散モデルを用いた音声強調

スコアによる拡散モデルを用いた音声強調の例は, 本実験でベースラインとした SGMSE+ [3] がある. これは, 式 (1) においての f, g を, 以下のように定義したものである.

$$f(x_t, y) = \gamma(y - x_t) \tag{3}$$

$$g(t) = \sigma_{\min}\sigma_r^t\sqrt{2\log(\sigma_r)} \tag{4}$$

$$\sigma_r = \sigma_{\max}/\sigma_{\min} \tag{5}$$

ここで, y とは環境雑音が重畳した音声である. これらにより, x_0 が与えられたときの条件付き確率 p_{0t} が以下のように導出できる.

$$p_{0t}(x_t|x_0, y) = N_{\mathbb{C}}(x_t; \mu(x_0, y, t), \sigma(t)^2\mathrm{I}) \tag{6}$$

$$\mu(x_0, y, t) = e^{-\gamma t}x_0 + (1 - e^{-\gamma t}y) \tag{7}$$

$$\sigma(t)^2 = \frac{\sigma_{\min}^2\left(\sigma_r^{2t} - e^{-2\gamma t}\right)\log(\sigma_r)}{\gamma + \log(\sigma_r)} \tag{8}$$

$\sigma_{\max}$, $\sigma_{\min}$, γ はハイパーパラメータで, SGMSE+では $\sigma_{\max} = 0.5$, $\sigma_{\min} = 0.05, \gamma = 1.5$ と設定された. このように定義することで, t が増加するにつれ, x_t の中心は y へ漸近する.

3 提案手法

3.1 潜在空間の利用

潜在空間の次元を, 元の次元よりも小さくする Encoder, Decoder(以降この 2 つを表現する際は EncDec と略す) を用意する. その潜在空間の中で逆拡散を実行する. 逆拡散における, 扱うデータの次元が小さければ, その分計算コストは軽くなる (図 1).

本実験では EncDec は VAE ではなく通常の AutoEncoder であるが, そのまま用いると潜在空間におけるデータの分散が, Encoder による変換前のデータに比べ著しく大きくなるという問題がある. そこで, Encoder の出力に tanh 関数を挟む.

4 実験

4.1 実験手順

実験は 2 つの段階に分けられる.
第一段階は, EncDec を訓練する. 雑音の無い音声を

Table 1　VB-DMD における, それぞれの圧縮率における強調処理の結果. Time は一発話の処理時間.

	PESQ↑	ESTOI↑	SI-SDR↑	Time[s]
SGMSE+	2.89	0.86	16.8	6.50
$r = 1/2$	3.12	0.87	17.4	4.33
$r = 1/4$	2.96	0.87	17.8	3.25
$r = 1/8$	2.62	0.80	12.0	2.81

Table 2　WSJ-C3 における, それぞれの圧縮率における強調処理の結果. Time は一発話の処理時間.

	PESQ↑	ESTOI↑	SI-SDR↑	Time[s]
SGMSE+	2.96	0.91	17.2	14.1
$r = 1/2$	3.05	0.91	17.9	8.25
$r = 1/4$	3.03	0.91	17.2	6.56
$r = 1/8$	2.60	0.85	11.3	3.88

EncDec に通し, 生成された音声と処理前の音声に関する二乗誤差を損失関数として訓練した.

第二段階は, 第一段階で訓練した EncDec を用いて, 拡散モデルの NN を訓練する. この際, EncDec のパラメータは固定する. 圧縮率は $r = 1/2, 1/4, 1/8$ を実行した.

4.2　実験設定

ベースとなる SGMSE+のコードには Web で公開されているものを利用した[1]. 拡散モデルの定義に関する各パラメータは [3] から変化はない. データセットは, VOICEBANK-DEMAND(VB-DMD と略す) [6], WSJ0-CHiME3(WSJ-C3 と略す) [3] を用いた. 評価指標には, PESQ [7], ESTOI [8], SI-SDR [9] を用いた. 第二段階にて使用する EncDec は, 雑音の無い音声を再構成するタスクに関して PESQ [7] が 4.0 以上を達成できることを確認した.

5　結果

表 1,2 は, 各データセットに対する音声強調の性能評価結果である. 演算処理に利用した GPU は RTX 5000 である. これらの結果により、$r = 1/4$ までは性能を落とさずに生成時間の短縮が可能であることが示された. 特に $r = 1/4$ の場合, 50% 以上の短縮に成功した.

6　結論

本研究では, SDE による拡散モデルにおいて, 潜在空間を用いることで計算コストを削減しつつ精度を保てることを示した.

今後は, 潜在空間におけるデータの標準偏差と性能の関係の検証, さらに高い圧縮率の実現, GAN や VAE の利用, 複素数として扱う機械学習の手法との組み合わせが研究課題である.

[1] https://github.com/sp-uhh/sgmse/tree/main

参考文献

[1] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, Vol. 26, No. 10, p. 1702–1726, 10 2018.

[2] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7402–7406, 2022.

[3] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 2351–2364, 2023.

[4] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[5] Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, Vol. 12, No. 3, pp. 313–326, 1982.

[6] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pp. 146–152, 2016.

[7] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, Vol. 2, pp. 749–752 vol.2, 2001.

[8] Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 11, pp. 2009–2022, 2016.

[9] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019.