人工知能学会研究会資料
SIG-KST-035-06(2018-11-22)

敵対的生成ネットワークを用いた機械音の生成

Generation of Mechanical Sound Using Generative Adversarial Networks

田添 康平 [1,2]　河合 継 [1]　渡邊 滉大 [1,3]　光明 誠 [4]

[1] クリスタルメソッド株式会社
[2] 東京工業大学大学院情報理工学院
[3] 早稲田大学基幹理工学部情報通信学科
[4] トヨタ紡織株式会社

Abstract: 本稿では敵対的生成ネットワークを用いた機械接続音生成手法を提案する．機械製品の製造では，作業者がわずかな異音を聞き取り不良品を検知する工程が存在し，その自動化が期待されている．そのために不可欠な雑音抑圧技術では深層学習を使用した方式が高い性能を誇る一方で，多彩かつ大量の学習用機械音の収集は，非常に高コストである．そこで，既存データを利用し新たな音源を生成する手法として DCGAN と pix2pix に着目し、それらを用いて機械接続音を生成することを試みた．機械接続音を用いた実験の結果，pix2pix では元の機械接続音と質的に近いデータの生成に成功した．

1　はじめに

生産設備、工場の内部では、実に様々な音が発生している．サイレン音、金属音、人の声など様々である．生産工程で作業の正しさを証明するために一定の検査が必要になるが、音による検査の場合、周りの雑音が影響して正しい検査が行えないことがある．現状，このような作業は熟練した専門家が実施しており，特に機械接続不良検知の自動化が望まれている．

近年，機械学習に基づく雑音抑圧技術の有効性が実証されている [16]．一般的に，機械学習に基づく雑音抑圧システムは，静かな環境で収録した音源（ドライソース）と，対象環境で収録した種々の騒音をドライソースに重畳することで得た模擬雑音下音源の対を用いて構築されることが多い．

特に，高い性能を与えることが実証されている深層学習に基づく方式は，大量の学習データを必要とすることが知られている．

このとき，人手による作業である機械接続の際に生じる接続音は多様であり，それらを静かな環境において網羅的に収集することは非常に高コストである．そこで本研究では，所望の多様な機械接続音のドライソースを，既存のデータを用いて生成することを試みる．

学習データの拡張を目的に既存のデータを加工する試みは多く行われている [13, 14, 15]．近年，学習データと類似のデータを生成する技術として，変分自己符号化器 (variational autoencoder; VAE) [2] や，敵対的生成ネットワーク (generative adversarial network; GAN) [1] が盛んに研究されている．特に，GAN に基づくアプローチにより，明瞭な画像の生成 [8, 9, 10, 11] や音声の生成 [12] に成功しており，学習データ拡張への利用も現実性を帯びてきている．

そこで本研究では，GAN に基づくアプローチに焦点を当て，機械接続音のドライソース生成を試みる．特に，deep convolutional GAN (DCGAN) [3] と pix2pix [4] を適用し，機械接続音生成性能を比較する．

本論文の構成は以下の通りである．まず，2 において，本研究の基本技術である GAN について概観した後，DCGAN と pix2pix を用いた機械音の生成について述べる．続いて 3 において，それらを用いた機械接続音生成実験について述べる．最後に 4 でまとめと今後の課題を述べる．

2　GAN に基づく機械音生成

2.1　GAN

学習したデータと類似したデータを生成することを目的とする深層学習モデルを，生成モデルと呼ぶ．GAN はこの生成モデルの 1 種であり，生成器（Generator）と鑑別器（Discriminator）と呼ばれる 2 つのネットワークから構成される（図 1）．Generator G では学習データと似たデータの生成を試み、Discriminator D では入力されたデータが Generator から生成されたデータであるか否かを識別する．生成モデルでは，学習デー

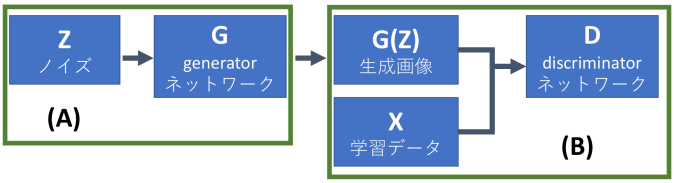


図 1: GAN の基本構造．(A) 入力から，鑑別機 D に本物と認識させる画像を生成する (B) 入力された画像が本物か偽物かを判別する

タに非常に近い出力を行う生成機 G を得ることを目的とする．

GAN の目的関数は式 (1) で与えられる．ただし，ノイズ $z \in Z$ を既知の任意の分布 P_Z から得られる独立なベクトル，X を自然なデータ空間とすることで，Generator $G : Z \to X$，Discriminator $D : X \to [0, 1]$ なる写像として定義される．

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(x)))] \quad (1)$$

この手法を用いて，Generator はデータ生成について学習し，Discriminator はより精度の高い識別について学習する．このような手法はミニマックスゲームと呼ばれる．両者の関係は，紙幣の偽造者とそれを判別する警察の関係に例えられることが多い．

本章では，GAN の一種である DCGAN と pix2pix の 2 つの手法を使用した機械音の生成手法について述べる．

2.2 データセットの構築

機械音の一つである，機械を接続する際に発生した機械接続音をドライソースとして録音し，学習に使用する．ただし，録音信号をパワースペクトログラムに変換する処理を行い，画像データを入力データとして使用することで，音データの生成に応用する．パワースペクトログラムとは，音データに短時間フーリエ変換を適用して周波数，時間，振幅分布の三次元で表示した記録図である．縦軸，横軸，色味はそれぞれ，線形周波数，時間，周波数成分の強さを表す．訓練用に録音した 10 種類の機械接続音を 512×32 pixel のパワースペクトログラムに変換した（図 2 左）．

これらの録音信号から変換された 10 の画像データを学習用素材として用いる．

2.3 DCGAN を用いた機械音生成

DCGAN [3] は，Radford らによって提案された生成モデルである．GAN の Generator, Discriminator に対

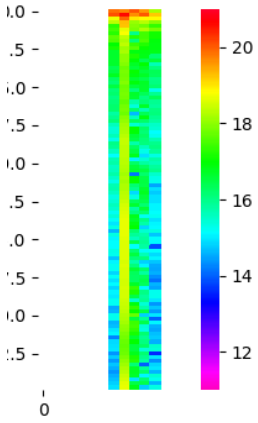


図 2: 機械接続音のパワースペクトログラム

して CNN（Convolutional Neural Network）を応用することで，高解像度な画像の生成を可能とする．図 3, 4 は今回使用した Generator, Discriminator の構造を表す．活性化関数は ReLU を用い，Generator の出力層にのみ Sigmoid 関数を用いた．なお，パラメータは実験的に設定している．

パワースペクトログラムから機械音発生箇所のみを抽出し，512×8 pixel のデータに変換し，学習データとして使用した（図 5 右）．

2.4 pix2pix を用いた機械音生成

pix2pix[4] とは，Isola らによって提案された生成モデルであり，入力データと正解データの組から，両者の関係性を学習する．任意の入力を与えることで，学習した関係性を反映した出力を行うことが出来る．pix2pix の Generator には，画像セグメンテーションのための U-Net[6] が使われている．U-Net は，データから抽出された局所的な特徴のみでなく，位置に関する情報も保持できるという特徴を持つ．浅い層で獲得される特徴も取得されるため，質の高い画像を得ることができる．

図 6，7 はそれぞれ, pix2pix の Generator, Discriminator の構造を表す．活性化関数は LeaklyReLU を用い，パラメータは実験的に設定した．

今回の学習データとして，機械音発生箇所のみを切り取った 512×8 pixel のパワースペクトログラム画像から，ランダムに 8×8 pixel の領域を選択し，選択箇所を切り取った．8×8 pixel の領域を抽出した画像を入力データ（図 5 左），抽出前の画像（図 5 右）を正解データとして与えることで，抜き取られた領域を補完したパワースペクトログラムを生成するように学習させた（図 5）．pix2pix では，録音信号 1 つにつき 1000 個の入力データを生成した．録音信号数が 10 であったので，合計 10000 の学習用機械音を生成した．

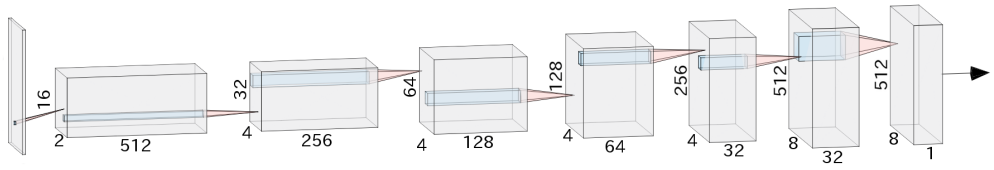


図 3: DCGAN Generator ネットワーク構造

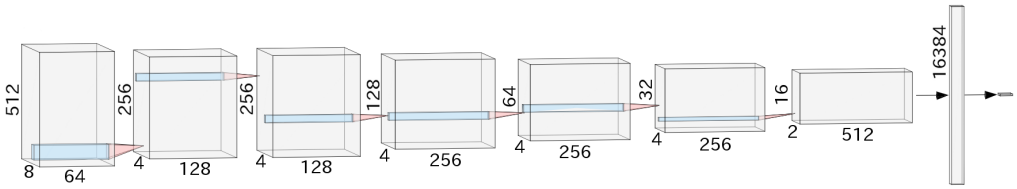


図 4: DCGAN Discriminator ネットワーク構造

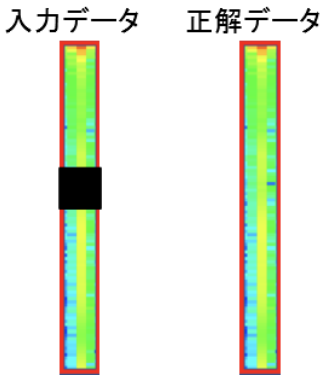


図 5: pix2pix のために切り出したパワースペクトログラム

3　機械音生成実験

ここでは 2 つの実験を行う。 1 つめの実験では全体を比較、 2 つめの実験では pix2pix での生成信号の妥当性について検討する．

DCGAN および pix2pix により生成された機械音のパワースペクトログラムの例を（図 8）に示す．録音信号の波形（図 9）と比較した場合，DCGAN では元の信号と大きく異なった波形（図 10）が生成された．一方で，pix2pix で生成した信号（図 11）では類似した波形が得られた．

DCGAN では，振幅の変動が開始･終了する時刻，振幅がピークを取る時刻はどのデータでもおおよそ一致したが，最大振幅および各時刻での振幅値についてはデータごとのばらつきが大きくなる結果が得られた．

pix2pix では、一部分を切り抜き、512×8 のうちの一部分のみを生成することで、各ドライデータにつき 1000 個データを生成し、計 10000 組の豊富な学習データを生成することができた．切り抜きの箇所を増やすほど、元のデータと類似度が低い機械音が生成されることが確認できた．

生成結果の評価のため，10 の録音信号と 10000 の生成信号について，dynamic time warping (DTW) [7] によって信号同士の信号間距離を計算した．録音信号同士の全ての組み合わせ，録音信号と DCGAN による生成信号の全ての組み合わせ，録音信号と pix2pix による生成信号の全ての組み合わせに対して計算した信号間距離の最大値，最小値，平均値を表 1 に示す．表を元に，二つの手法による生成結果の比較を行う．

DCGAN の生成信号について，録音信号同士の比較と比べた際，すべての項目について大きな差がそれぞれ見られた．反対に，pix2pix の生成信号では，すべての項目について，録音信号同士の比較と遜色ない値が得られた．以上より，pix2pix では十分に望ましい機械接続音の生成ができたと言える．

表 1: 信号 1 と信号 2 の全ての組み合わせにおける誤差の最大値, 最小値, 及び，平均値

信号 1 信号 2	録音 録音	録音 DCGAN 生成	録音 pix2pix 生成
最大値	0.016	0.068	0.016
最小値	0.006	0.067	0.004
平均値	0.011	0.067	0.005

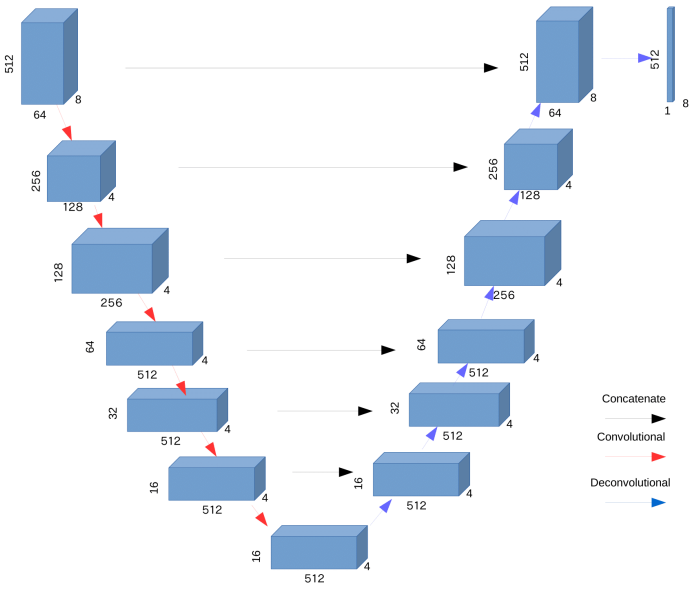


図 6: pix2pix Encoder-Decoder(Generator) ネットワーク構造

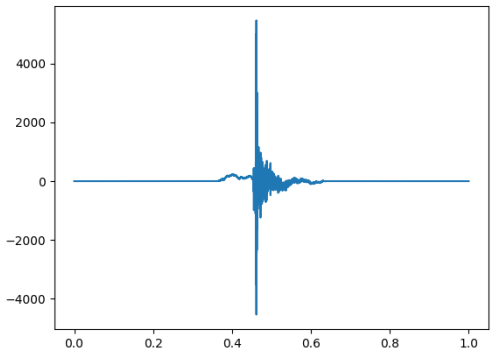


図 9: 録音された機械音波形の例

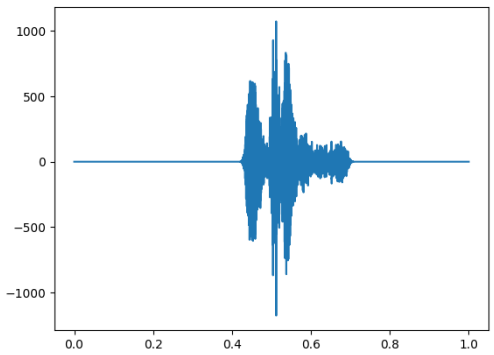


図 10: DCGAN による機械音波形の例

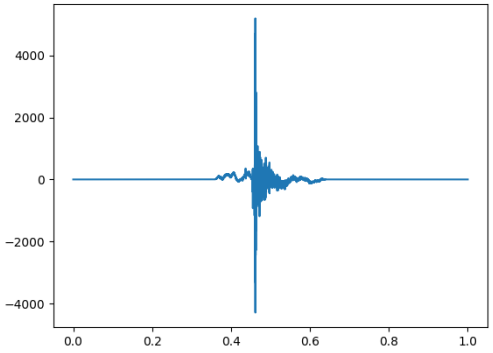


図 11: pix2pix による機械音波形の例

次に pix2pix にてパワースペクトログラムから 8 × 8pixcel の領域をそれぞれ 2 箇所、4 箇所、8 箇所、16 箇所、32 箇所切り取ったものを入力データとして学習用機械接続音を生成した. 録音信号数が 10 であったので、それぞれ 10000 の学習用機械接続音を生成した.

生成結果の評価のため，録音信号と pix2pix による生成信号の全ての組み合わせについて dynamic time warping (DTW) [7] によって信号同士の信号間距離の平均値を計算した. 信号間距離の平均値と信号生成の際に切り取った 8 × 8 領域数の関係を（図 12) に示す.

切り抜きの箇所を増やすほど，元のデータと類似度が低い機械音が生成されることが確認できる。

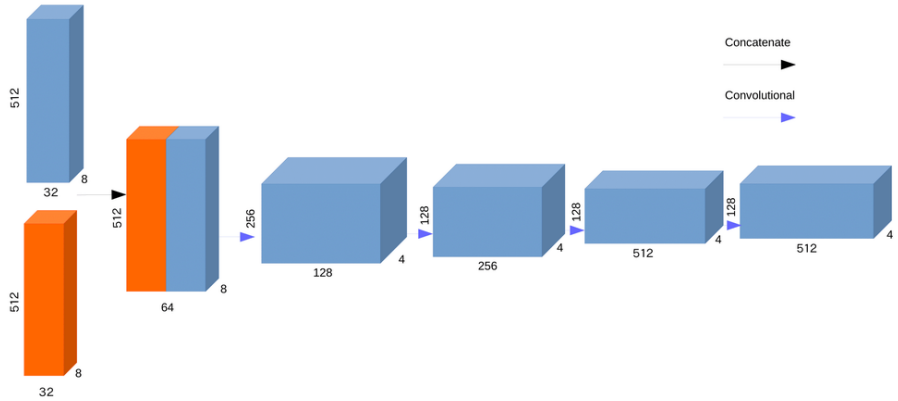
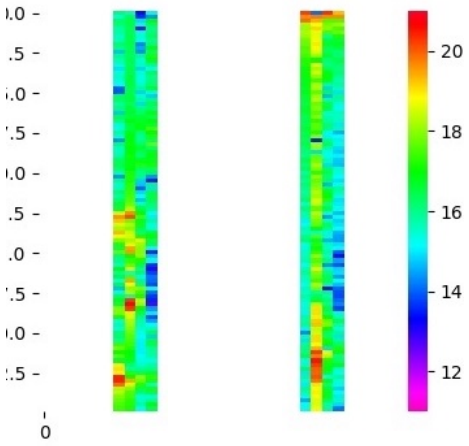


図 7: pix2pix Discriminator ネットワーク構造



(a) DCGAN (b) pix2pix

図 8: 生成信号のパワースペクトログラム

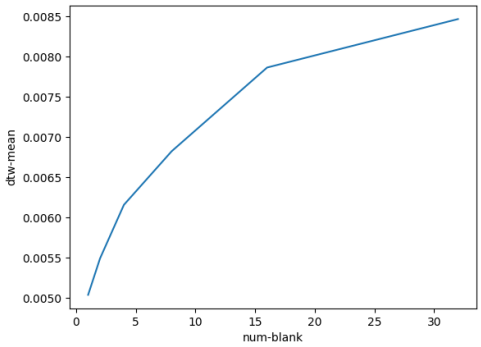


図 12: 切り取った領域数と DTW 値の関係

4 まとめ

機械装置の製造工程における，機械音を作業者が聴き取ることで不良品を検知する工程を，音による不良品検知技術を利用して自動化する方法を検討した．

音源中に存在する様々な雑音の除去を行うために，データ・ドリブンな深層学習を使用する場合，音源の人手による網羅的な収集には莫大なコストがかかる．

そこで本論文では，敵対的生成ネットワーク（GAN）の一種である DCGAN と pix2pix の 2 つの手法を用いた音源生成実験とその考察を行なった．

生成の結果，DCGAN では学習データの不足から所望する音源生成はできなかったが，pix2pix による生成ではオリジナルの機械接続音と質的に近い音源を得ることができた．本稿では機械接続音について取り扱ったが，今後は，その他の局所発生的な音声全般についても応用を試みたい，

参考文献

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets., *In Advances in neural information processing systems* , pp. 2672-2680 (2014)

[2] Kingma, D. P., Welling, M.: Auto-encoding variational bayes., *arXiv preprint*, arXiv:1312.6114 (2013)

[3] Radford, A., Metz, L., and Chintala, S.: .Unsupervised representation learning with deep convolutional generative adversarial networks., *arXiv preprint* arXiv:1511.06434. (2015)

[4] Isola, P., Zhu, J. Y., Zhou, T., et al.: Image-to-image translation with conditional adversarial networks., *arXiv preprint* (2017)

[5] Xi, X., Keogh, E., Shelton, C.: Fast time series classification using numerosity reduction., *In Proceedings of the 23rd international conference on Machine learning*, pp. 1033-1040 (2006)

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

[7] Bellman, R., Kalaba, R.: On adaptive control processes, *In IRE Transactions on Automatic Control*, pp. 1-9(1959)

[8] Denton. E, Chintala. S, Szlam. A, and Fergus. R; Deep Generative Image Models using a Laplacian Pyramid of Adversarial Networks, *arXiv preprint* (2015)

[9] Ledig. C, Theis. L, Huszar. F, Caballero. J, Cunningham. A, Acosta. A, Aitken. A, Tejini. A, Totz. J, Wang. Z, and Shi. W; Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network, *arXiv preprint* (2016)

[10] Zhang. H, Xu. T, Li. H, Zhang. S, Wang. X, Huang. X, and Metaxas. D; StackGan: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks, *arXiv preptint*

[11] Karras. T, Aila. T, Laine. S, Lethinen. J; Progressive Growing of GANs for Improved Quality, Stability, and Variation, *arXiv prepring*

[12] Donahue. C, McAuley. J, and Puckette. M; Synthesizing Audio with Generative Adversarial Networks, *arXiv preprint* (2018)

[13] Salamon. J, Bello. J; Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound, *In IEEE Single Processing Letters*, pp. 279-283(2017)

[14] McFee. B, Humphrey. E, Bello. J; A Software Frameworkfor Musical DataAugmentation *In The International Society for Music Information Retrieval* pp. 248-254(2015)

[15] Simard. P, Steinkraus. D, Platt. J; Best Practices forConvolutional Neural Networks Applied to-Visual Docment Analysis *In International Conference on Document Analysis and Recognition* pp.958-962 (2003)

[16] Feng.X, Zhang.Y, and Glass.J: Speech feature dinoising and dereverberation via deep autoencoders for noisy reverberant speech recognition *In International Conference on Acoustics, Speech and Signal Processing* pp.1759-1763(2014)