

ニューラルネットワークによる 音楽の自動変換

陶山大輝^{1 2 3}

2021 年 1 月 6 日

¹東京大学教養学部後期課程学際科学科総合情報学コース

²学籍番号：08-192021

³指導教員：金子知適准教授

目次

第 1 章	はじめに	2
1.1	Remix	2
1.2	音色	2
1.3	音色の変換	2
1.3.1	楽器の重ね合わせ	2
1.3.2	時間方向の音の繋ぎ方	2
1.3.3	音の重ね合わせ	3
1.4	既存研究	3
第 2 章	背景	4
2.1	多層パーセプトロン	4
2.2	GAN	4
2.3	CGAN	4
2.4	Pix2pix	5
第 3 章	データセット	6
3.1	音	6
3.2	データ形式	6
3.3	データセット	6
第 4 章	提案手法	7
第 5 章	まとめ	8

第1章 はじめに

1.1 Remix

音楽を作成する方法の一つとして remix がある。remix とは既存の曲の再構成及び加工を行うことでその曲の新しいバージョンとする方法のことである。

1.2 音色

音色とは、日本産業規格の定義によれば「聴覚に関する音の属性の一つで、物理的に異なる二つの音が、たとえ同じ音の大きさ及び高さであっても異なった感じに聞こえるとき、その相違に対応する属性」とある。

1.3 音色の変換

音色の変換を音楽で行う際に (音の特徴を学習する) が、以下の三つの点を解決するのが難しいと考えられる。また、以下の三つを解決することで、単音における音色の変換を音楽に適用することができる。

1.3.1 楽器の重ね合わせ

楽器ごとに音色が異なるので、楽器ごとの音波に分解して音色変換を行うことが良いと考えられる。なお、楽曲の作成時に楽器ごとに分離したデータ (パラデータ) で保存しておけば、直接楽器ごとの音波を利用できる。

1.3.2 時間方向の音の繋ぎ方

時間方向での音の繋ぎ方は都合の良いように分割していくことでなんとかなるのではないかな…?、分割する (1 つの音の判定を行う) のは難しい…?、自己回帰?

1.3.3 音の重ね合わせ

単位時間の楽器の音に注目した時、楽器ごとの音波に分離したとしても和音のようにその単位時間で複数の種類の音が鳴っている場合も難しいと考えられる。

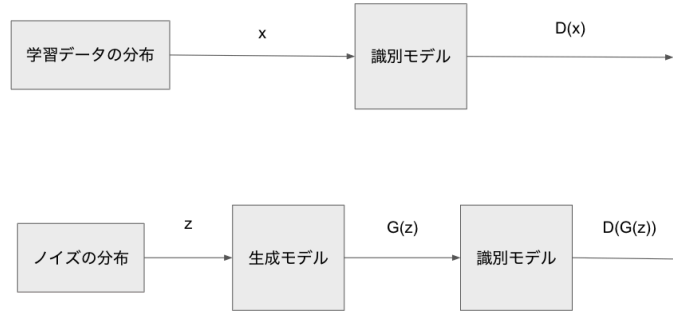
とりあえず試しても良いので実験を行いたい

1.4 既存研究

第2章 背景

2.1 多層パーセプトロン

2.2 GAN



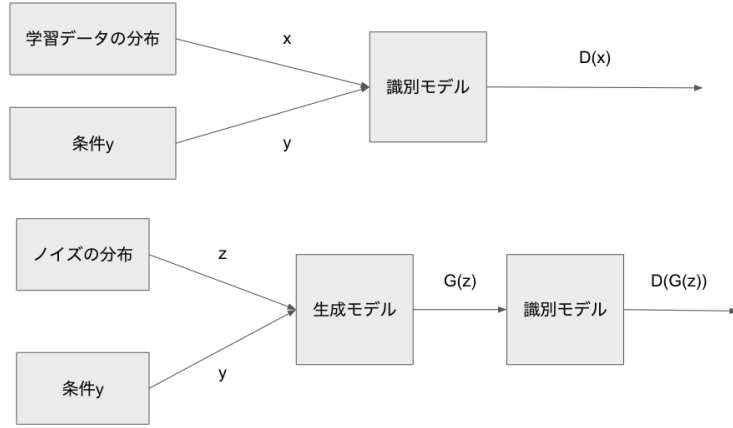
GAN (敵対的生成モデル) とは生成モデルと識別モデルが競合して学習を行うディープラーニングのモデルである。生成モデルは訓練データの分布を捉えようとし、識別モデルは訓練データである確率を推定する。

具体的には、訓練データ x の分布を p_{data} , 生成モデルの入力のノイズ z の分布を p_z , ノイズ z を元にデータを生成する関数を G , 生成モデルのデータではなく訓練データである確率を返す関数を D とした時、式 2.1 を生成モデルは最小化し識別モデルは最大化をすることを目標として学習を行う [1]。

$$\mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (2.1)$$

2.3 CGAN

CGAN(条件付き敵対的生成モデル) とは条件付きの GAN である。具体的には y という条件下で式 2.2 を生成モデルは最小化し識別モデルは最大化をすることを目標として学習を行う [2]。



原著論文では訓練時に MNIST の数字のラベルを条件として与えることで特定の数字のラベルの画像を生成するモデルを実装している。

$$\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x} \mid \mathbf{y})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z} \mid \mathbf{y})))] \quad (2.2)$$

2.4 Pix2pix

pix2pix とはある条件下で画像間の変換を行う GAN である。先程の CGAN を元に U-Net と PatchGAN を組み合わせたネットワークになっている。**U-net とは…, PatchGAN とは…。(ここは後々調べて書く)**

具体的には、画像 x を条件として式 2.5 を生成モデルは最小化し識別モデルは最大化をすることを目標として学習を行う [3]。

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y} [\log D(x, y)] + \mathbb{E}_{x,z} [\log(1 - D(x, G(x, z)))] \quad (2.3)$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z} [\|y - G(x, z)\|_1] \quad (2.4)$$

$$\mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \quad (2.5)$$

第3章 データセット

3.1 音

弾性体中を伝播する弾性波により起こされる音波が聴覚されるもののこと。この論文では弾性体として空気のみを想定する。

また、音は時間を定義域としたアナログ信号なので、離散化及び量子化を行ってデジタル信号へと変換することでデジタル機械の扱えるデータ形式となる。

3.2 データ形式

音のファイル形式としては WAV を用いる。一般的には MP3 や MP4 などの非可逆圧縮形式が広く用いられるが、WAV は波形データを直接保持しており本論文では WAV を扱いやすいと考えたためである。

音のデータそのものだけでなく量子化ビット, サンプルング周波数, チャンネル数, サンプルング数のメタ情報を得ることができる。

また、サンプルング周波数を 44100Hz, 量子化ビットを 16bit, チャンネル数を 1, サンプルング数を 44100 に固定して今回の実験を行った。

3.3 データセット

楽譜作成ソフトの MuseScore を利用して国際の階名表記で A0 から C8 に含まれる半音を wav 形式で 88 音生成した。また、この 88 音は一般的な 88 鍵のピアノで出すことのできる全音階の音であり、人間が音程として聞き分けることのできる限界の音域としてこの音域を選んだ。

また、楽器としてはエレキギターとハープを選んだ。

第4章 提案手法

第5章 まとめ

謝辭

関連図書

- [1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.
- [2] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.
- [3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.