ニューラルネットワークによる音色の自動変換

学際科学科 総合情報学コース

陶山大輝

学籍番号：08-192021

指導教員：金子知適准教授

初めに

Remixによる音楽作成の補助

- Remix
  - 既存の音楽をアレンジする音楽の作成方法
  - **アプリケーションや楽器の操作が必要で難しい**
- 音色（おんしょく・ねいろ）の変換
  - Remixの操作の一つ
  - プログラムによる補助ができるのでは
  - **ニューラルネットワークによる変換を提案**

提案手法

ニューラルネットワークによる単音での音色の変換

- 使用するニューラルネットワーク：Pix2pix
- 目標：ギターの音色からハープの音色への音の高さを維持したままの変換
- 実験の評価：音波の観察により行う

Pix2pix [Phillip Isola, 2018]

ピクセルの対応関係を維持したままスタイル変換を行う手法

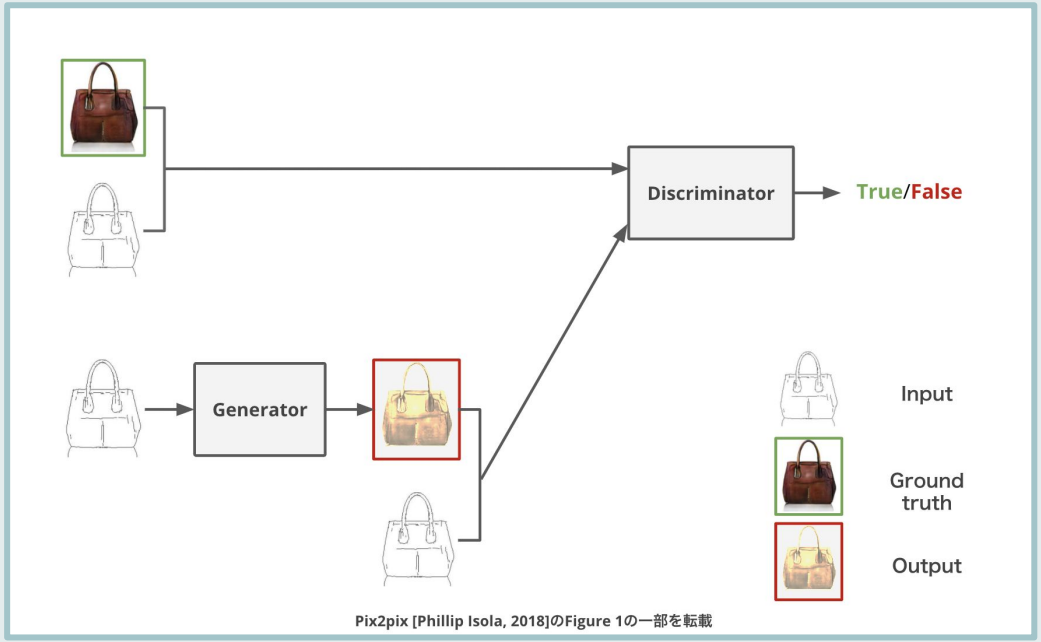
Pix2pix [Phillip Isola, 2018]のFigure 1の一部を転載

- EdgesからPhotoへの変換
  - 線画を写真風に着色する
- BWからColorへの変換
  - 白黒写真を着色する

[Phillip Isola, 2018] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

Pix2pix：基本構造

変換元の画像を条件とする敵対的生成ネットワーク

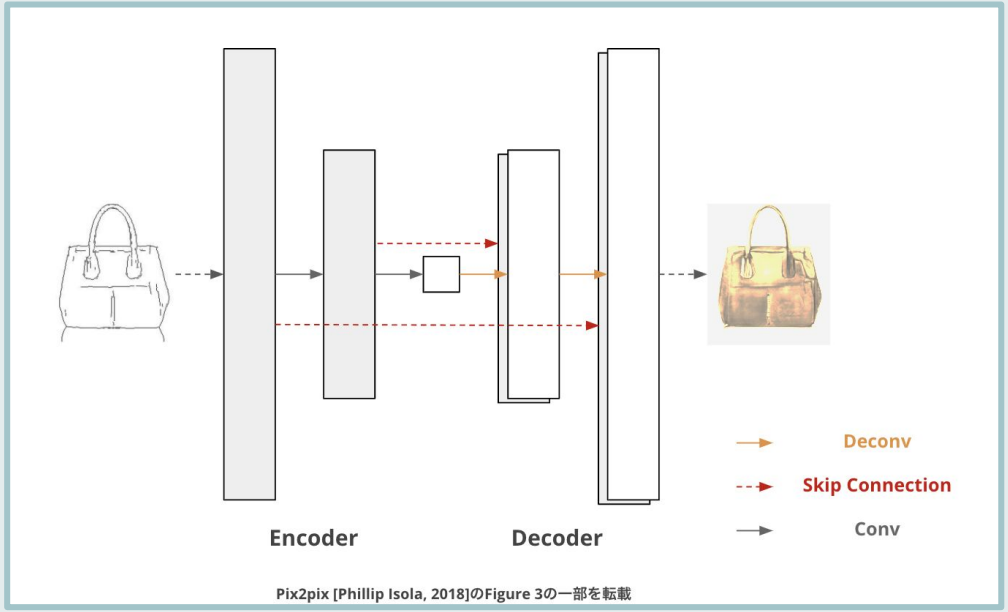


Pix2pix [Phillip Isola, 2018]のFigure 1の一部を転載

- 敵対的生成ネットワーク（GAN）
  - GeneratorとDiscriminatorが競合する
- Generator
  - **本物に近いデータ**を生成する
- Discriminator
  - **本物のデータ**かどうかを識別する

Pix2pix：Generator (生成モデル)

変換元の画像とのピクセルの対応関係を維持する生成モデル



Pix2pix [Phillip Isola, 2018]のFigure 3の一部を転載

- Encoder-Decoder
- **Skip Connection**がピクセルの対応関係を維持

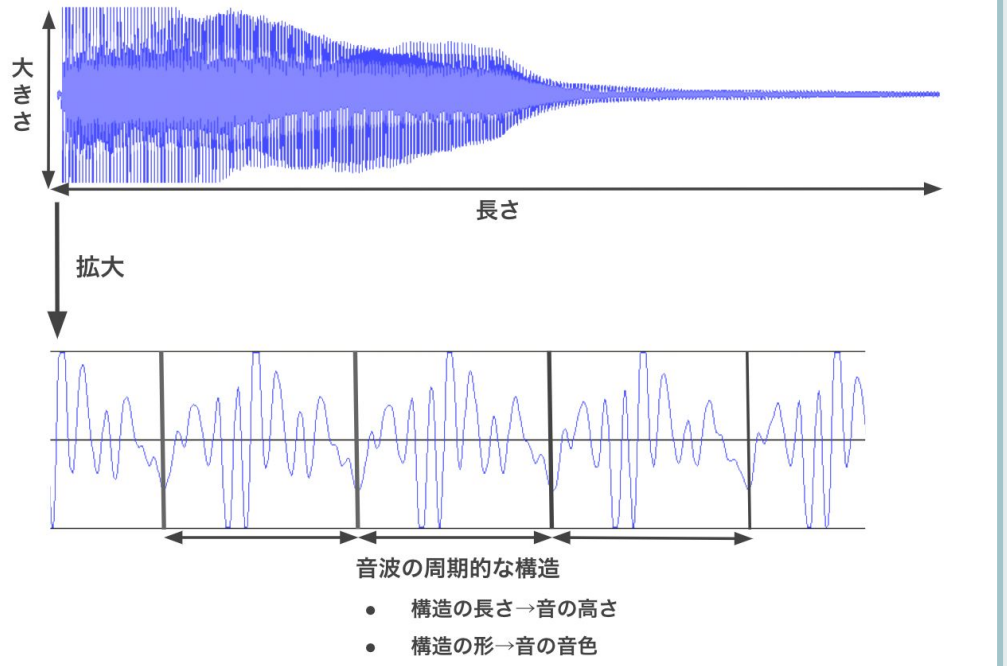
Pix2pix：Discriminator (識別モデル)

局所的な部分の識別精度を向上させる識別モデル

Pix2pix [Phillip Isola, 2018]のFigure 1の一部を転載

- 左側：一般的なGANの識別モデル
  - **画像全体で**真偽を識別する
- 右側：Pix2pixの識別モデル
  - **小領域ごとに**真偽を識別する

音色とは



- 音の区分
  - 騒音：不規則な音波の振動
  - **楽音**：周期的な音波の振動
- 楽音の4要素
  - 長さ
  - 大きさ
  - 高さ
  - **音色**

実験

Pix2pixを応用したモデルの学習と評価

I. 生成モデルの表現力の評価

   A. データセット：学習と評価でそれぞれ**同じ88音**のギターとハープの同じ高さの音のペア

II. 生成モデルの汎化能力の評価（ 4分割交差検証 ）

   A. データセット：学習と評価でそれぞれ**66音と22音**のギターとハープの同じ高さの音のペア

III. 実験時の評価と工夫

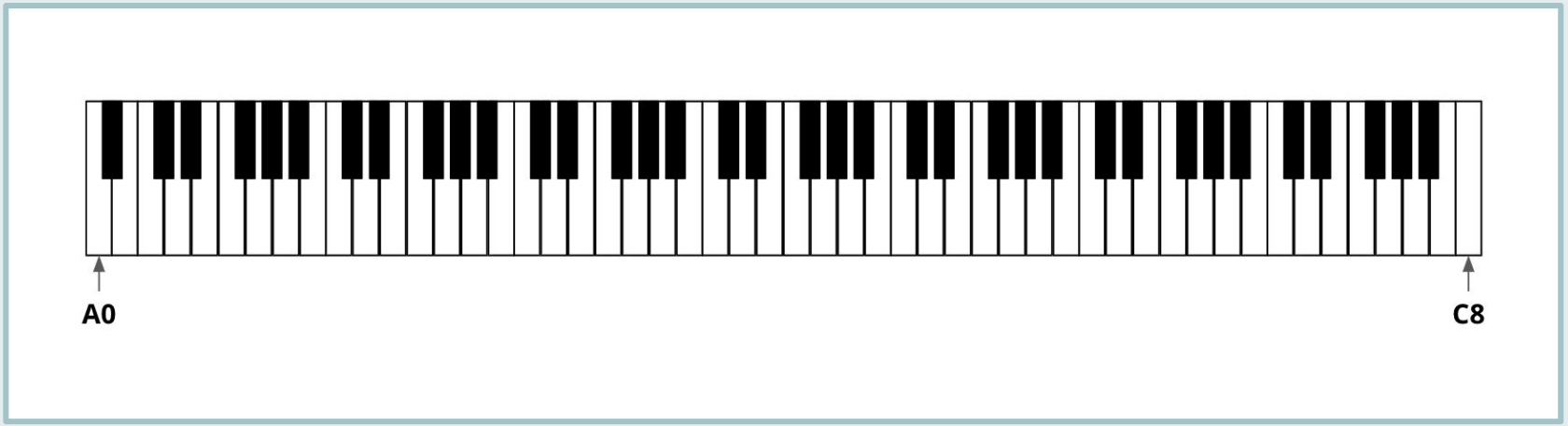
   A. 評価：生成された音が音の高さを維持したままハープの音色を表現できているか

   B. 工夫：音の大きさを 0.3 ~ 1.0 倍にすることで過学習の抑制を期待した

実験：データセット

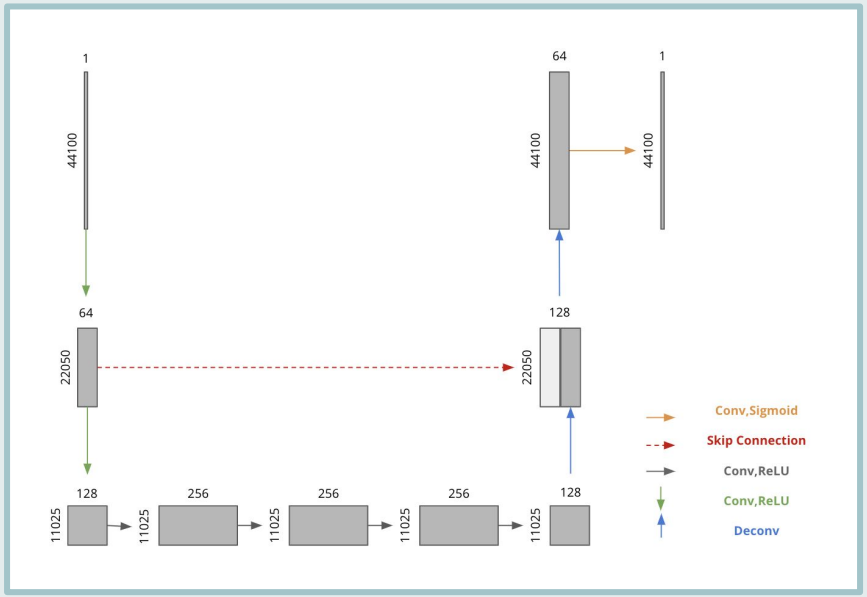
88音のギターとハープの音

音の高さの範囲：A0~C8（88音の半音）

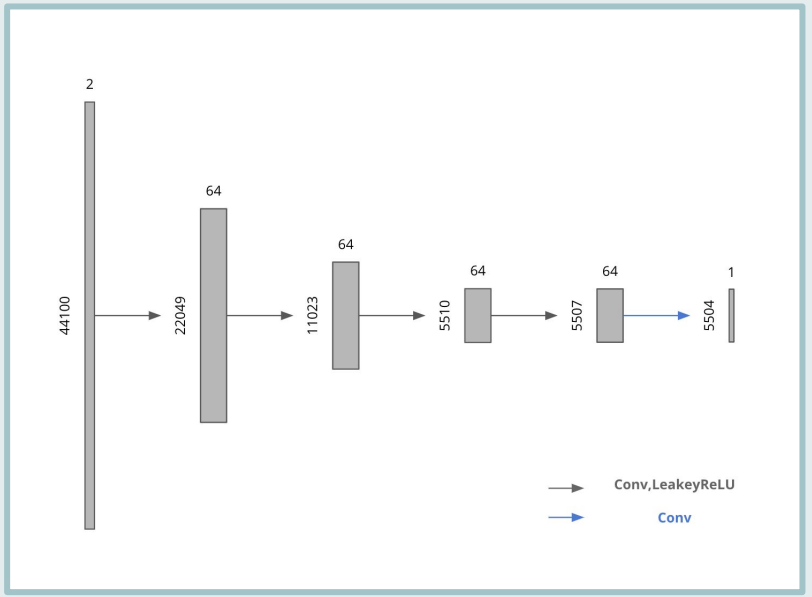


実験：ニューラルネットワークのモデル

生成モデル



識別モデル



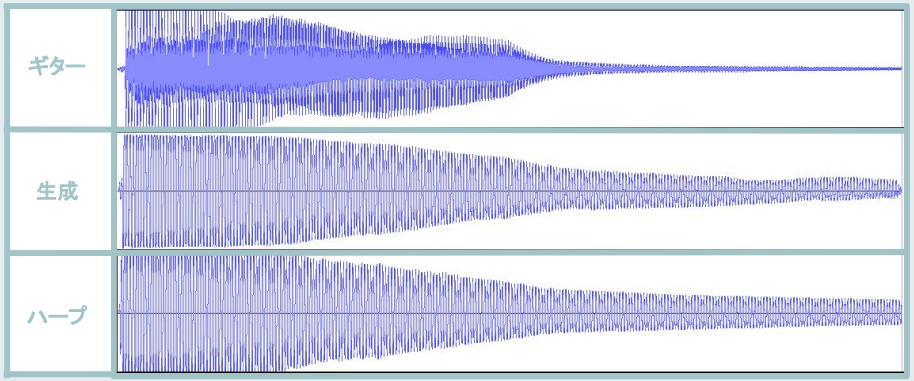
実験：結果のまとめ

音を楽音として生成する → 成功

音の高さを維持したまま変換する → 概ね成功

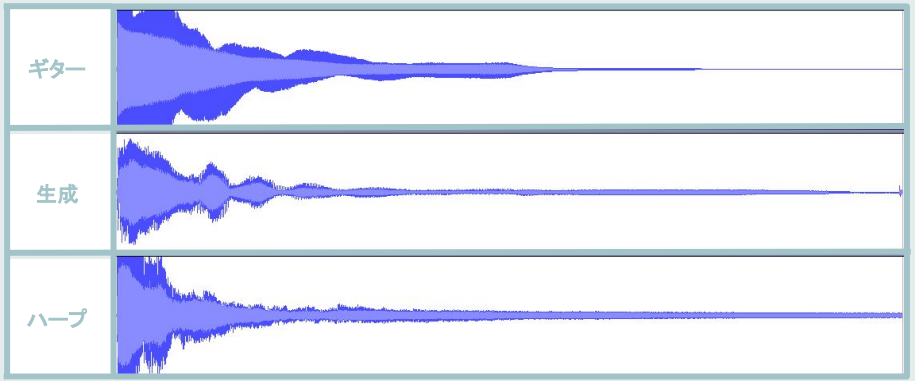
音の音色を表現する → 一部のみ成功

実験1:生成モデルの表現力の評価



音色を表現できている

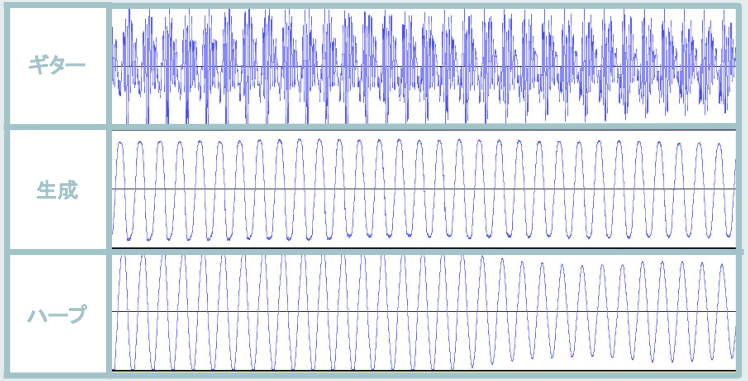
- 87音



音色を表現できていない

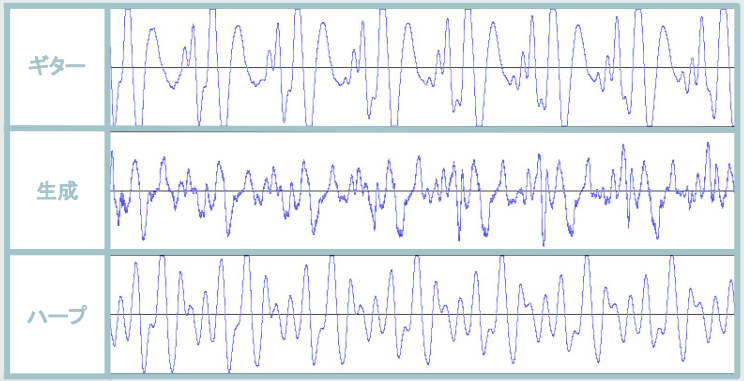
- 1音(E7)

実験2:生成モデルの汎化能力の評価



音色を表現できている

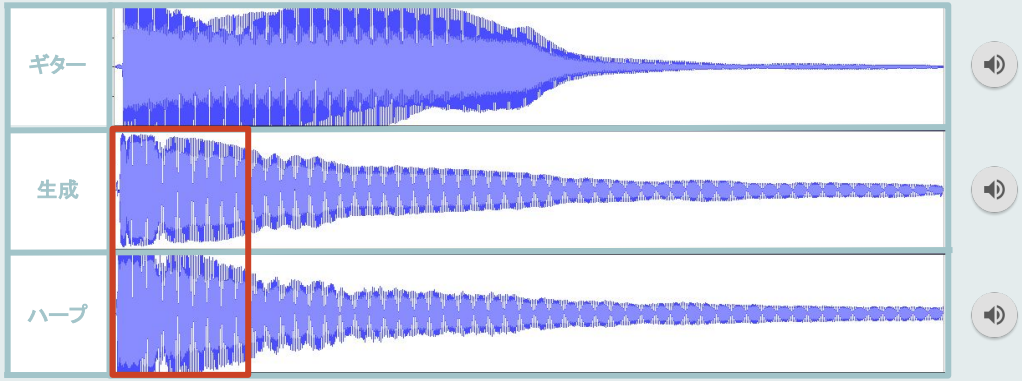
- 5音(D4,D4#,G4,F5,F5#)



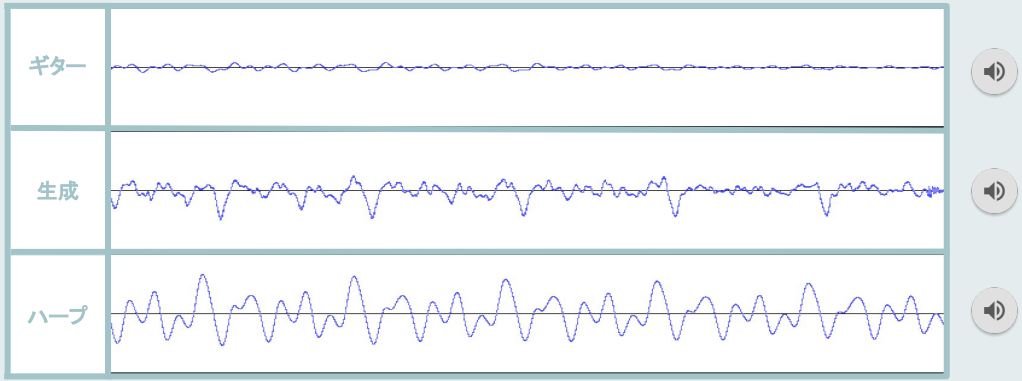
音色を表現できていない

- 74音:高さは維持される
- 9音:高さも維持されない
  - C7,D7#,E7,F7#,G7,G7#,A7,B7,C8

実験：課題

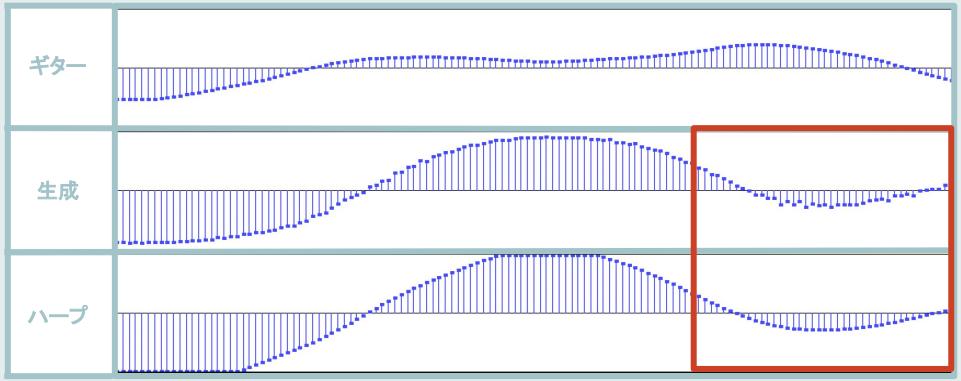


音の大きさを維持できていない

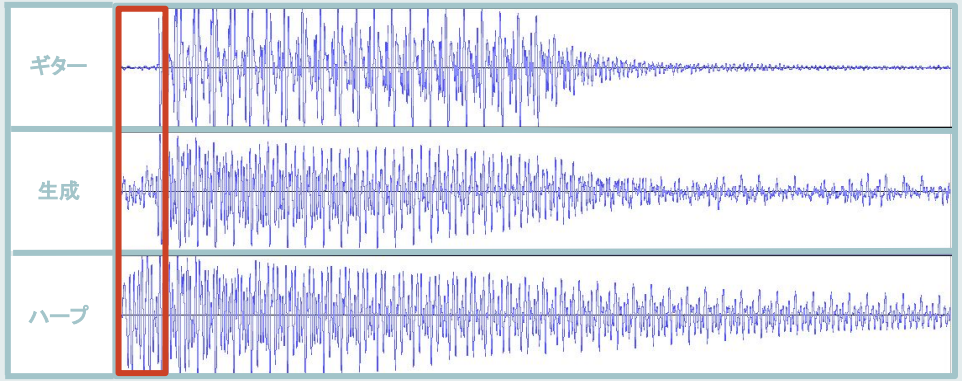


音の減衰を表現できていない

実験：課題



音波の滑らかさを表現できていない



データセットに問題がある

- 音の鳴り出しに遅延がある
- 位相にずれがある

展望

Remixによる音色の変換が容易になる

以下の3点を工夫する必要がある

	本研究	Remix
楽器の重ね合わせ	1つ	複数
音の重ね合わせ	1つ	複数
音楽の長さ	1秒	数分

ご静聴ありがとうございました

予備スライド

まとめ：今後の課題

- 音の大きさの維持
- 音波の滑らかさの表現
- 音の減衰の表現
- 複雑な音波の表現
- 安定したデータセットの作成
- 定量的な判定方法

展望

楽器の重ね合わせ

- 楽器ごとに音色は異なるので、楽器ごとの音波に分解して音色変換を行う（音源分離）
- 楽器ごとに分離したデータで保存されている場合、音源分離は必要としない

音の重ね合わせ

- ある単位時間の音は異なった高さや大きさの音の重ね合わせとなる
- 重ね合わせた音もデータセットに加えることで実現可能と考えている

音楽の長さ

- ある単位時間で分解した部分をそれぞれ変換することで実現できると考えている
- 実現可能ではない場合は自己回帰モデルなどの別の手法を利用する

既存手法での音楽の前処理

Raw dataとSpectrogram

- Raw data
  - NSynth [Jesse Engel, 2017], Jukebox[Prafulla Dhariwal, 2020]
  - 細かい表現もできるが周期の把握が難しい
- Spectrogram
  - TimberTron [Sicong Huang, 2019]
  - 短時間フーリエ変換により行うが、時間分解能と周波数分解能の間にトレードオフがある

前処理の際のパラメータ

- サンプリング周波数：44100 Hz
- サンプリング数：44100 回
- 量子化ビット数：16 bit
- チャンネル数：1（モノラル）

ギターとハープを選んだ理由

1. 人間の耳で音色の違いを聴き分けることができる点
2. 弦楽器の間での変換なので変換が難しくないと考えられる点
    a. 管楽器（トランペット）と弦楽器（ギター）の変換は難しいと考えられる

質問

音の大きさの維持について

→維持されていないものも多かったが、人間の耳では聴き分けられない程度

Pix2pixを選んだ理由

→周期の対応がピクセルの対応に相当すると考えた

Generatorにノイズを用いていない理由

→決定論的に音を生成しようと思った

→Pix2pixではDropoutを用いているので、今後試そうと考えている

質問

ニューラルネットワークの改善点

→256×256の画像ではSkip Connectionを4つ程度用いていた

→増やすことで改善できると期待

ハープの音を表現できているのか

→波形の粗さが表現できていない理由

→定量的な判定方法を考えることで評価できるのでは