So-vits-svc (RVC)

https://github.com/svc-develop-team/so-vits-svc/tree/4.1-Stable

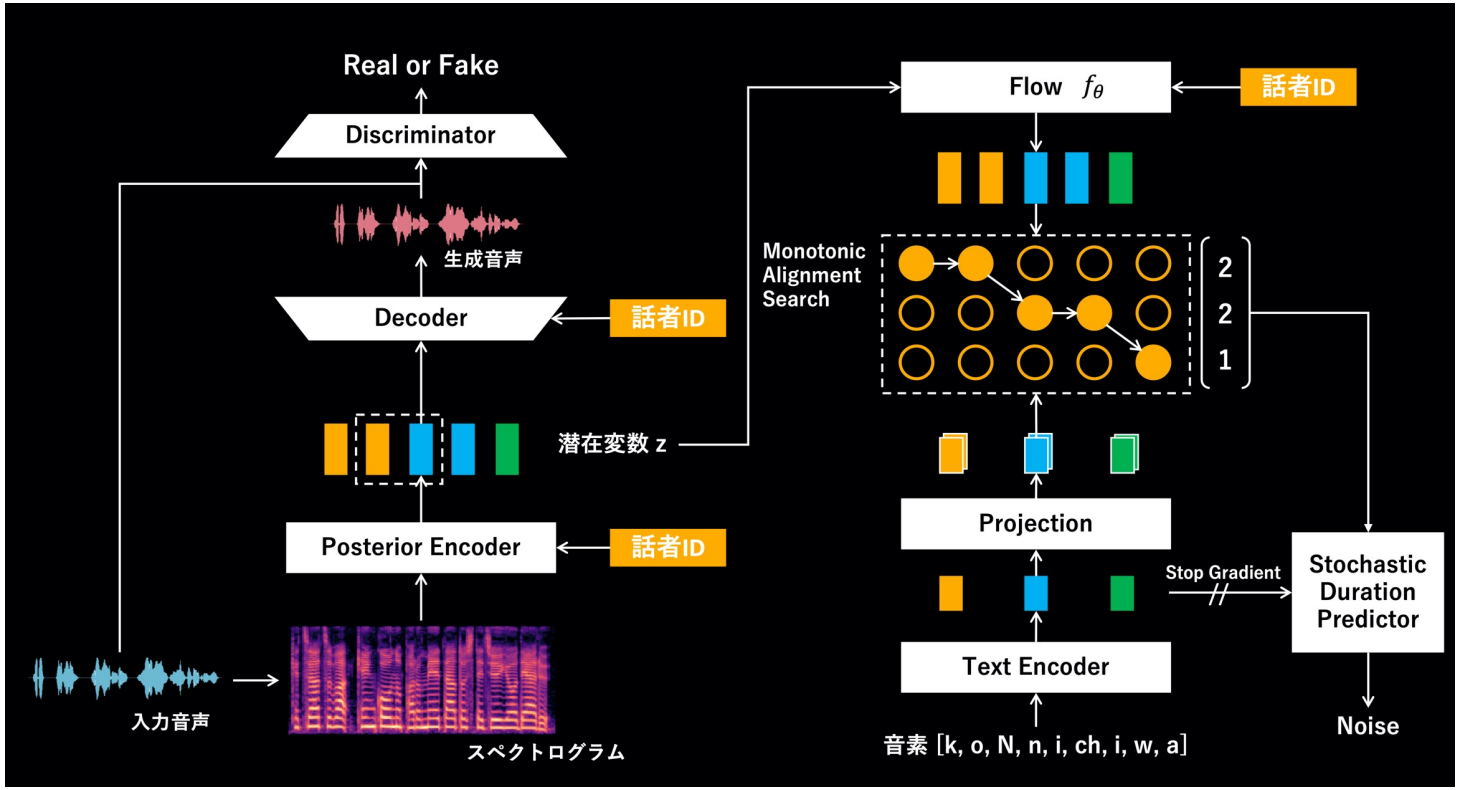
(論文なし)

Abstract

- 音声合成モデルのVITSをSoft-Voice-Conversionに応用
- HuBERT(Content Vec)と基本周波数を使用して合成する
- 音声データのみで学習可能
- 学習データに無い英語や歌声も合成可能

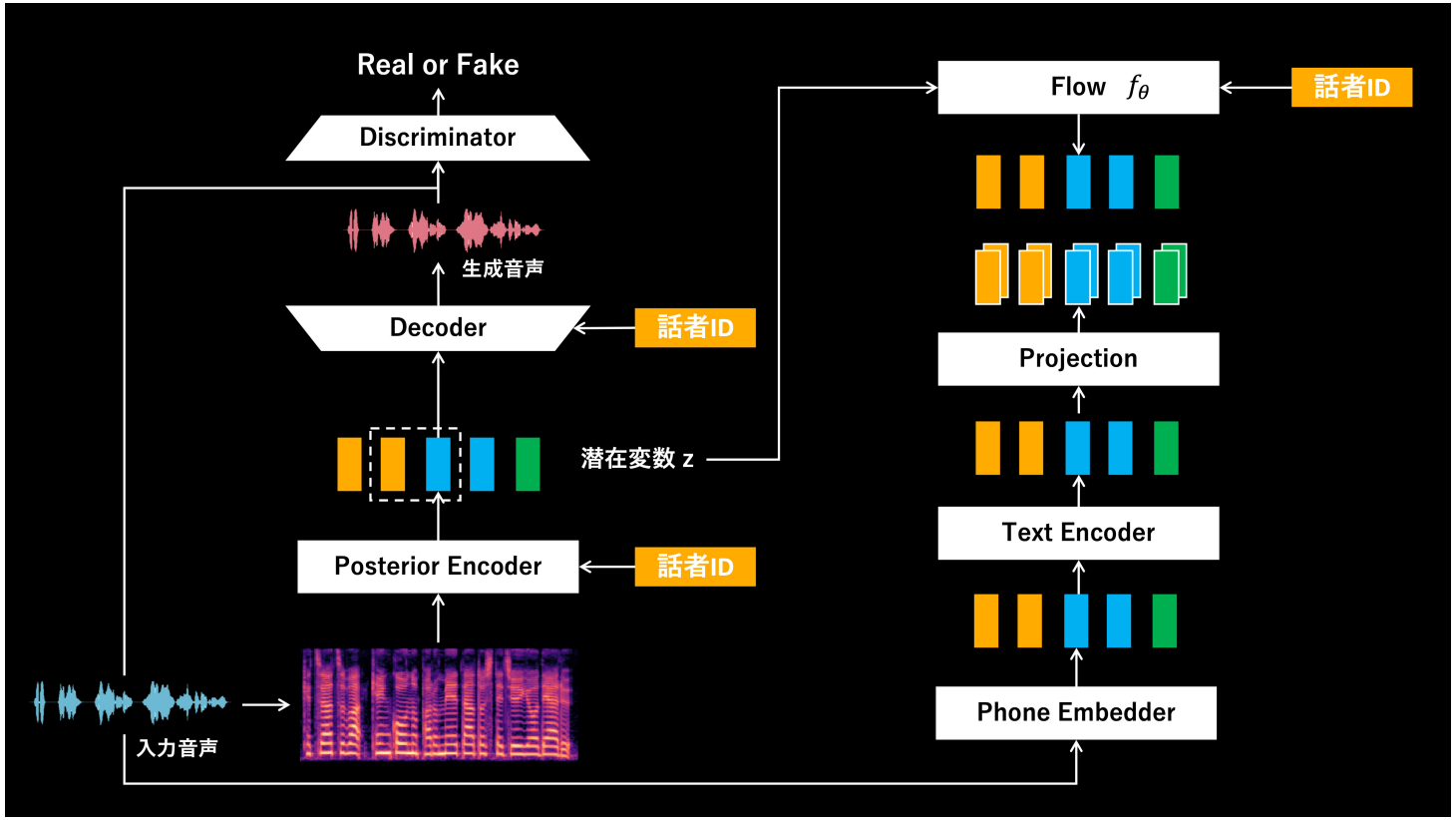
VITS

- VAE・GAN・Flowを用いた音声合成モデル
- 一番使われている。
- 元の構造でもVCに使える。
- 音素をText Encoderへ入力



so-vits-svc

- Phonemeの代わりにHuBERTなどの特徴量をクラスタリングし、その重心を入力
- PitchとHuBERTの長さは雑に揃えている
- Alignment Searchは削除

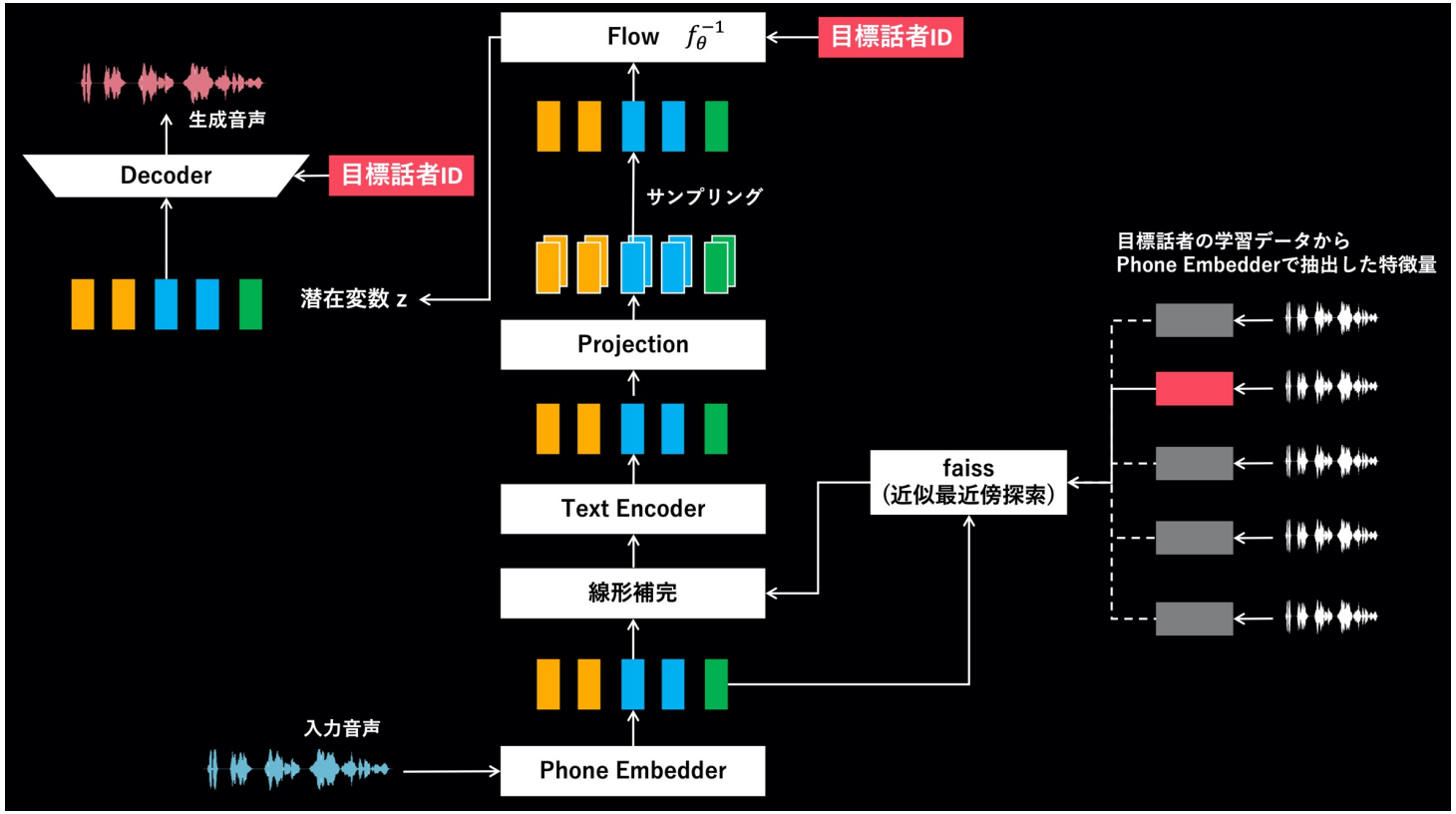


Inference (RVC)

- 入力音声の話者情報が不要
- 学習したPhone Embedderを保持しておき、その中で一番近いものを利用

→話者性が完全に排除できていないため

- 探索処理が遅い



So-vits-svcの推論時間

- 44100Hz, 8.621043083900227秒の音声に対しての推論時間
  - total time: 14.256680965423584
    - extract HuBERT and F0: 3.0919458866119385
    - vits use time:11.134729862213135
      - Phone Embedder~Flow: 0.7261261940002441
      - Decoder: 10.40252423286438

- Decoderに時間がかかっている
  - サンプル数が多い波形を生成する部分のため重い.
  - サンプリングレートに依存する部分であり, 44100Hzで生成している.
  - HiFiGANベースの実装になっており, ここを軽くできる (?)
  - StarGANv2-VCでも近い処理を行っているため, モデル間に差は現れなさそう.

考えられる軽量化

- Decoderのモデルを変える
  - HiFiGANベースから変える (HiFiGANはCPUのコア数に大きく依存)
  - Decoderの入力する特徴量の次元を下げる

- HuBERT → DistilHuBERT

- 合成するサンプリングレートを22050Hz or 16000Hzに変更

HuBERT→DistilHuBERT

- HuBERT: 話者があまり入っていない発話内容に関する情報
- DistilHuBERT: HuBERTを軽量化したモデル
- ContentVec: HuBERTからさらに話者性を取り除く
- 論文によると推論時間が約4倍, モデルサイズが25%になる

質問

- 今使ってるVocoderはどのくらいの速度で合成できるか

- ベンチマークとするPCはあるか
  - CPUでもマシンによって速度がかなり違うため (Decoder 部分)
  - 今回は, 以下のPCで行っている

    MacBook Air (Retina, 13インチ, 2020)
    コア数 8個 (jupyter hub: 32個)
    1.1 GHz クアッドコアIntel Core i5
    16 GB 3733 MHz LPDDR4Xs