VITSを用いた低遅延Realtime-VC

~Low Latency Realtime-Voice Changer with VITS~

- VITS -
Conditional Variational Autoencoder
with Adversarial Learning for End-to-End Text-to-Speech

本日のテーマ

Section 1:とりま3行で(概要)

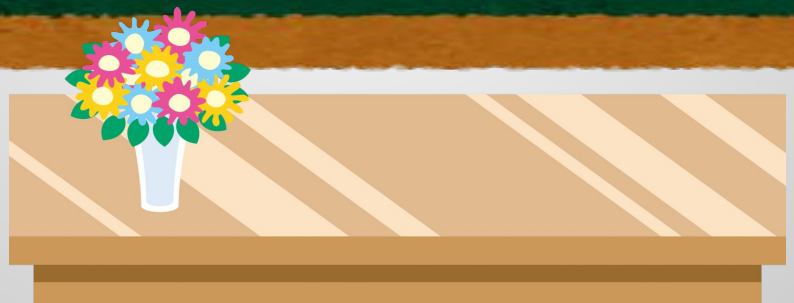
Section 2:VCの需要と課題

Section 3:VITSの概略

Section 4:VITSについて

Section 5:リアルタイムへ

Section 6:今後の課題とOSS化へ



・とりま3行で(概要)

・VITSについて、簡単に説明するよ！

・VITSを使ってリアルタイムボイチェン使ったよ！

・OSSで公開するよ！

以上の三本立てでお送りするよ！
なお、本セッションはdemo時間の短縮のため、実際に作成したrealtime-VCを利用して発表しているよ

本日のテーマ

Section 1:とりま3行で(概要)

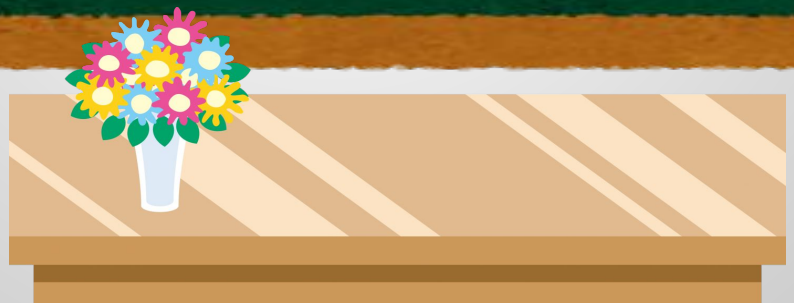
Section 2:VCの需要と課題

Section 3:VITSの概略

Section 4:VITSについて

Section 5:リアルタイムへ

Section 6:今後の課題とOSS化へ



・リアルタイムボイチェン(VC)の需要

人類は潜在的に美少女になりたいと思っている

・人類は美少女になるために様々な手法に挑戦し、実績を生み出してきた

VRChatなどなど

・ヴァーチャルの世界で可愛い姿になることはできるが…

・ボイチェンに選ばれしもの
・ボイチェンに適合するために訓練をした修行僧
のみが可愛い声を出せるのが現状である。

・既存の音声合成ではreal-timeではないので、会話はできない

ついでに既存ツールだと特定のキャラクターにしかなれない

全人類が真の美少女になるには誰もが可愛い声を出せる必要がある
故に誰もが簡単に美少女声になれる real-timeのVCが必要なのである

- リアルタイムボイスチェンジャーの課題(非AI)
  - 女の子の声に変換することができても理想の声になるとは限らない
  - 一部の選ばれしもの以外は訓練と様々な微調整が必要

    選ばれしもの以外は修羅の道
  - 他のボイチェン勢(非AI)と似たような声になってしまう
  - 機械音ノイズが発生する

    AIでも同じ課題を抱えてるけど…

・リアルタイムボイスチェンジャーの課題(AI)

リアルタイムボイスチェンジャーの課題

・処理時間の壁

処理時間を短縮するためにモデルを簡素にするとクオリティが…

・時系列データを用いるため、構成上難しい

一定時間の過去の入力を元に現在の入力を変換するので、辛い

・そもそものクオリティが…

・学習用のデータセットも大量に必要

AIはまだ発展途上といった感じだな。
時間とクオリティがトレードオフになってしまうのと、
データセットの準備が大変なのが課題だな

・リアルタイムボイスチェンジャーの課題(AI)

リアルタイムボイスチェンジャーの課題

・処理時間の壁

処理時間を短縮するためにモデルを簡素にするとクオリティが…

・時系列データを用いるため、構成上難しい

一定時間の過去の入力を元に現在の入力を変換するので、辛い

・そもそものクオリティが…

・学習用のデータセットも大量に必要

でもVITSならこの問題全部解決できちゃうんだ！
(クオリティについては 100%解決したとは言えないけど …)

本日のテーマ

Section 1:とりま3行で(概要)

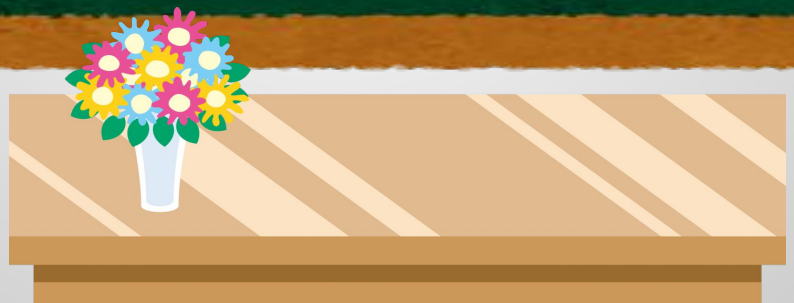
Section 2:VCの需要と課題

Section 3:VITSの概略

Section 4:VITSについて

Section 5:リアルタイムへ

Section 6:今後の課題とOSS化へ



・VITSとは？

2021年6月に綺羅星の如く現れた音声合成手法(TTS)！

特徴

・TTSモデルなのに声質変換(ボイチェン)ができる

・高品質

・少ないデータセットで学習可能

・End-to-End モデル

・高速　爆速　超音速！

・VITSとは？

2021年6月に綺羅星の如く現れた音声合成手法(TTS)！

特徴

・TTSモデルなのに声質変換(ボイチェン)ができる

・高品質

・少ないデータセットで学習可能

・End-to-End モデル

・高速　爆速　超音速！

圧倒的高品質！
Grand Truthは実際の声だぞ！
この結果は実際の声とほとんど
遜色ないことを示している！！

Table 1. Comparison of evaluated MOS with 95% confidence intervals on the LJ Speech dataset.

Model	MOS (CI)
Ground Truth	4.46 (±0.06)
Tacotron 2 + HiFi-GAN	3.77 (±0.08)
Tacotron 2 + HiFi-GAN (Fine-tuned)	4.25 (±0.07)
Glow-TTS + HiFi-GAN	4.14 (±0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	4.32 (±0.07)
VITS (DDP)	4.39 (±0.06)
VITS	**4.43 (±0.06)**

Table 3. Comparison of evaluated MOS with 95% confidence intervals on the VCTK dataset.

Model	MOS (CI)
Ground Truth	4.38 (±0.07)
Tacotron 2 + HiFi-GAN	3.14 (±0.09)
Tacotron 2 + HiFi-GAN (Fine-tuned)	3.19 (±0.09)
Glow-TTS + HiFi-GAN	3.76 (±0.07)
Glow-TTS + HiFi-GAN (Fine-tuned)	3.82 (±0.07)
VITS	**4.38 (±0.06)**

・VITSとは？

2021年6月に綺羅星の如く現れた音声合成手法(TTS)！

特徴

・TTSモデルなのに声質変換(ボイチェン)ができる

・高品質

・少ないデータセットで学習可能

・End-to-End モデル

・高速　爆速　超音速！

圧倒的速度！
VITS(DDP)では、1秒間に約200万サンプル、
48000kHzの音声でも
1秒で約41秒分も変換できる！

Table 4. Comparison of the synthesis speed. Speed of n kHz means that the model can generate $n \times 1000$ raw audio samples per second. Real-time means the synthesis speed over real-time.

Model	Speed (kHz)	Real-time
Glow-TTS + HiFi-GAN	606.05	×27.48
VITS	1480.15	×67.12
VITS (DDP)	2005.03	×90.93

本日のテーマ

Section 1:とりま3行で(概要)

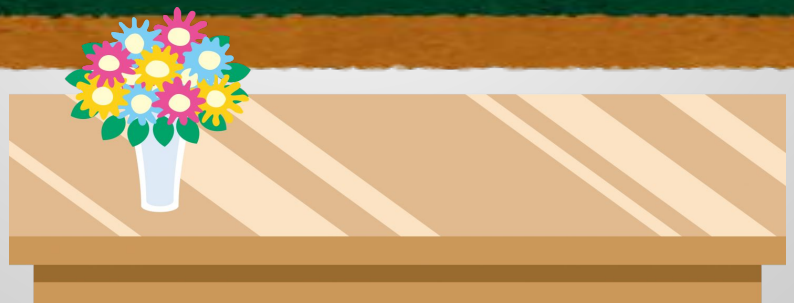
Section 2:VCの需要と課題

Section 3:VITSの概略

Section 4:VITSについて

Section 5:リアルタイムへ

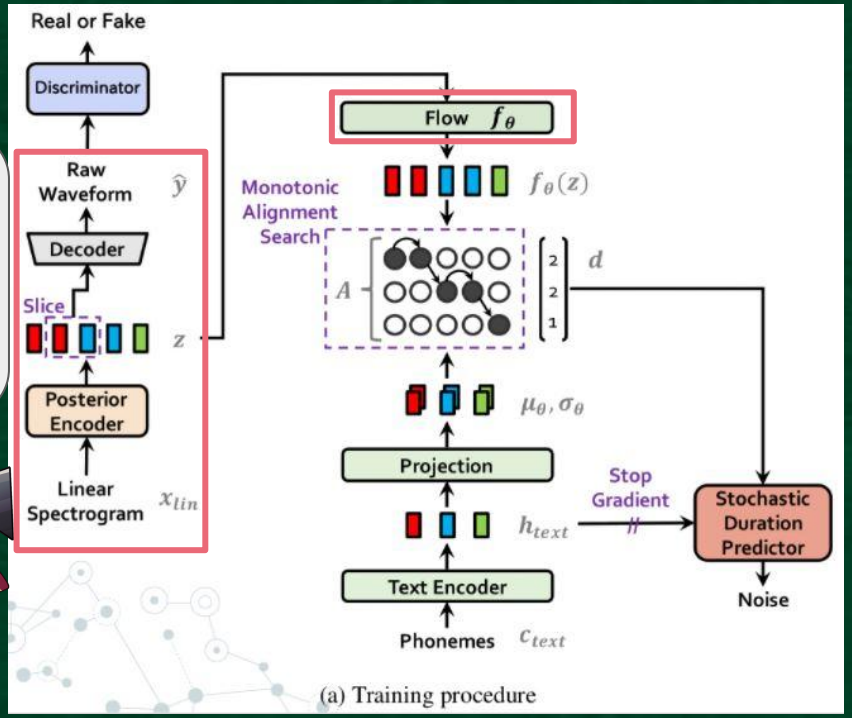
Section 6:今後の課題とOSS化へ



・VITSとは？

学習時の流れ

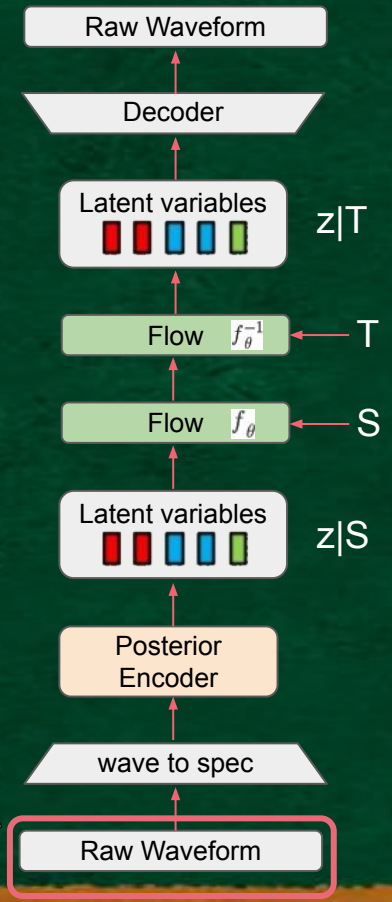
「ルールは一見複雑そうだけど、やれば簡単だぜ！」
見た目は複雑そうなモデルだが、一個一個見ていくと実はそんなに難しくない。
VCで使うのは赤枠で囲った部分のみだ。
学習の一連の流れは時間の都合上短縮な



・VITSとは？

推論時の流れ

入力として、8192サンプルを与えている。
今回はサンプリングレートを 24kHzにしているので、
24kHzの音声の場合、8192/24000 = 約0.341s
つまり、約0.341sの音声を変換しているのだ
※サンプリングレート
1秒間に情報の標本(サンプル)を何回計測したのか表す

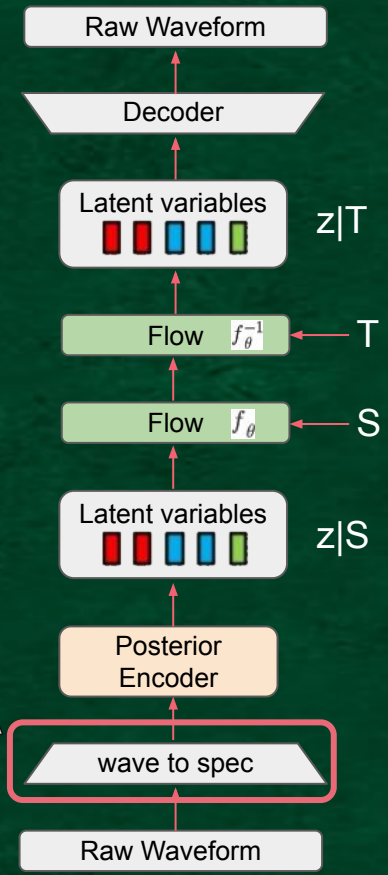


Tは目標話者ID
Sは元の話者ID

・VITSとは？

推論時の流れ

生の音声では入力はモデルが受け入れてくれないので
Linear specに変換している。
この辺はAI関係ない信号処理的な話だから
詳しくは割愛な

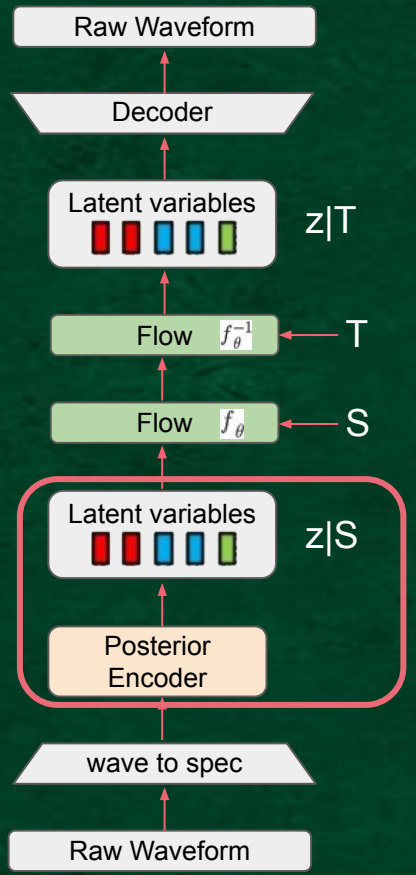


Tは目標話者ID
Sは元の話者ID

・VITSとは？

推論時の流れ

ここはさっき学習したモデルな
元の話者のLinear specを元の話者の潜在変数に変換する

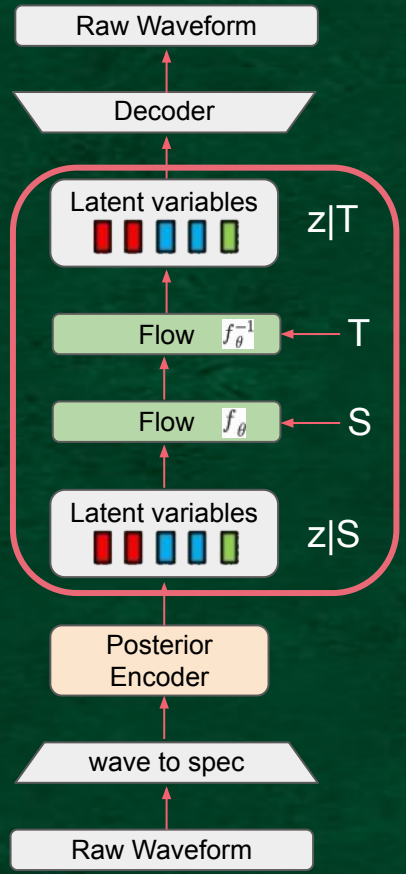


Tは目標話者ID
Sは元の話者ID

・VITSとは？

推論時の流れ

三大生成モデルの中で唯一可逆性を持つ Flowを用いて
元の話者IDを潜在変数を目標話者の潜在変数に変換している！

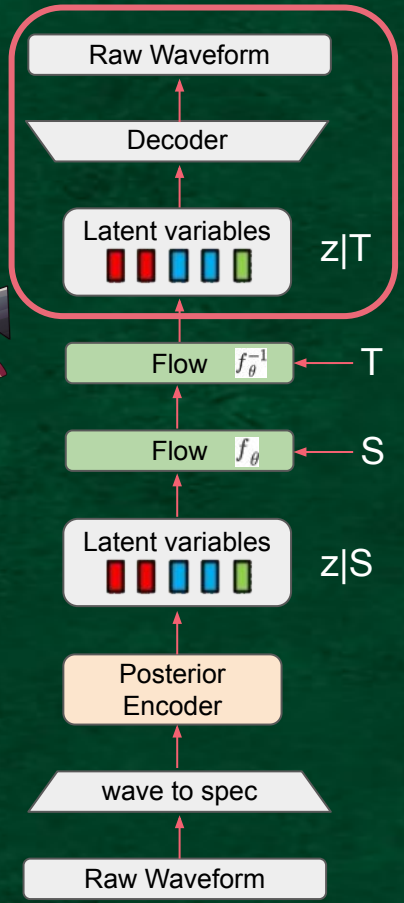


Tは目標話者ID
Sは元の話者ID

・VITSとは？

推論時の流れ

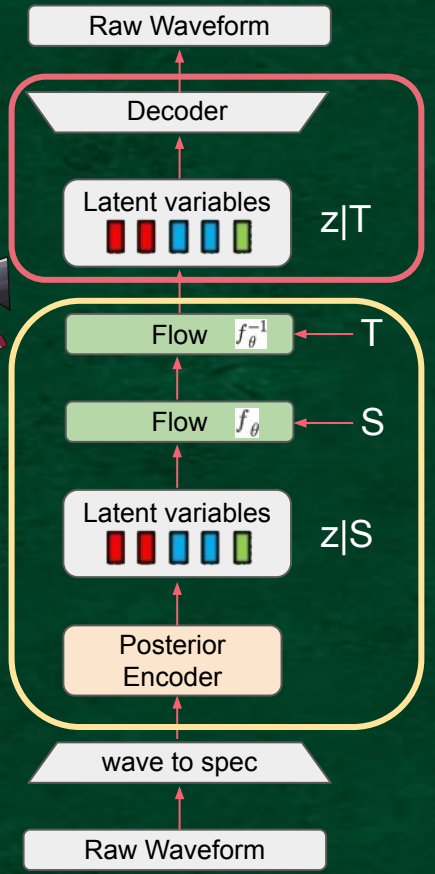
目標話者の潜在変数をそのまま生の音声に変換するのだ



Tは目標話者ID
Sは元の話者ID

・VITSとは？

推論時の流れ



Tは目標話者ID
Sは元の話者ID

ちなみに多くの手法ではEnd-to-Endではない。
どういうことかというと、ここの赤枠の部分と
黄色の枠の部分をそれぞれ学習しなければならない。
基本的に赤枠のボコーダーとか呼ぶ。

本日のテーマ

Section 1:とりま3行で(概要)

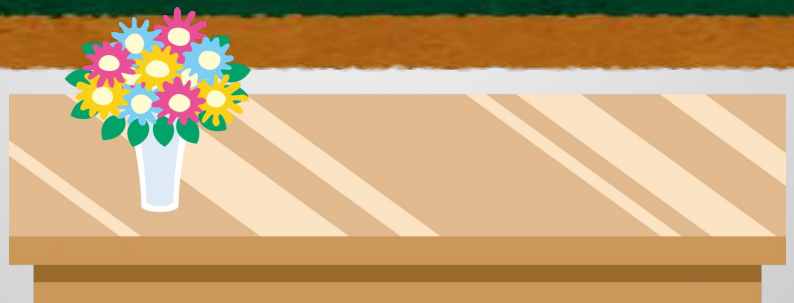
Section 2:VCの需要と課題

Section 3:VITSの概略

Section 4:VITSについて

Section 5:リアルタイムへ

Section 6:今後の課題とOSS化へ



・リアルタイムボイスチェンジャーの課題(AI)（再掲）

リアルタイムボイスチェンジャーの課題

・処理時間の壁

処理時間を短縮するためにモデルを簡素にするとクオリティが…

・時系列データを用いるため、構成上難しい

一定時間の過去の入力を元に現在の入力を変換するので、辛い

・そもそものクオリティが…

・学習用のデータセットも大量に必要

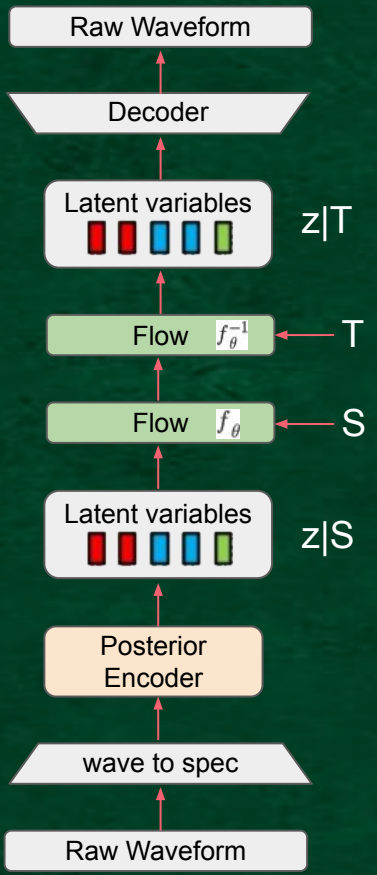
これらの課題をVITSなら解決してくれるのではないか？

実際にやってみた（現在進行形）

・リアルタイムへ

推論時の流れ（再掲）

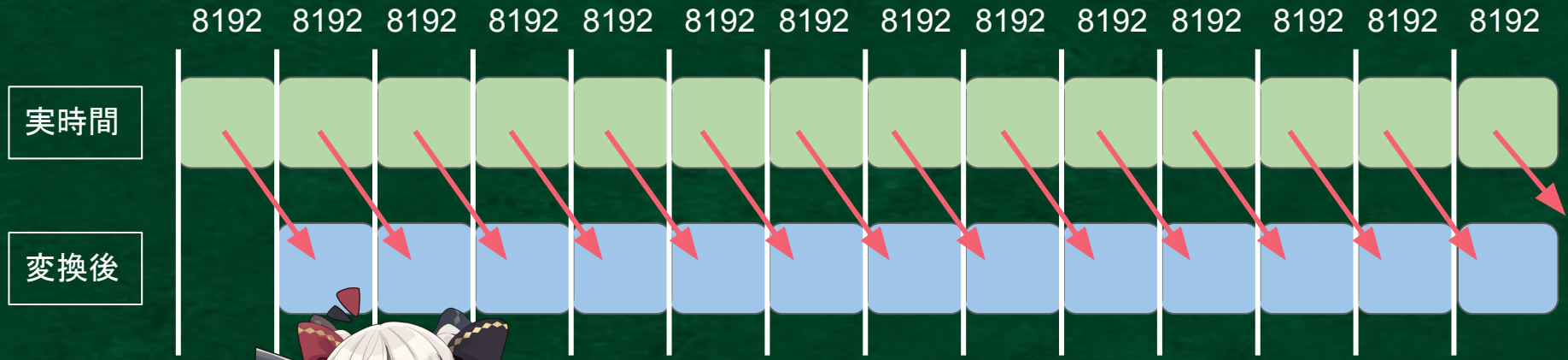
入力として、8192サンプルを与えている。
そして目標話者の音声に変換された 8192サンプルを
受け取ってる。
当然処理時間は高速とはいえ、0秒ではない。
なのでバッファーをはさんでいる。
このバッファーが 8192サンプルなので、
遅延約0.34sというわけだ。



Tは目標話者ID
Sは元の話者ID

・リアルタイムへ

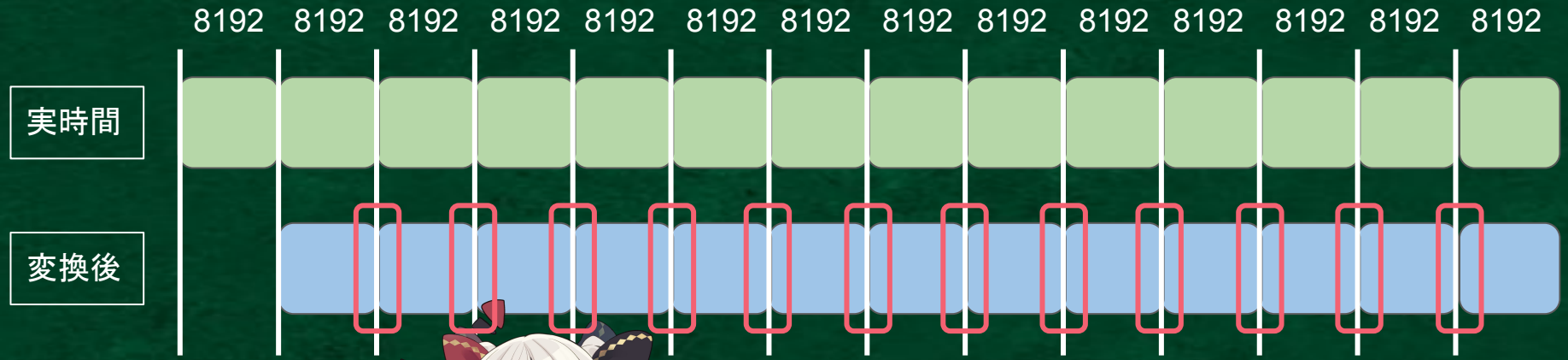
リアルタイムVCの処理



ここがバッファーで遅延時間。
バッファー後は変換後の音声を出力している間に次の入力を変換しているから
詰まることはないのだ

・リアルタイムへ

リアルタイムVCの処理



実は8192サンプルの固まりで変換しているから、変換前の音声は連続でも
変換後の音声は連続にならないんだ。
このままだとつなぎ目のタイミングでブツブツとなってしまう。
なのでつなぎ目をスムーズになるようにフェードインとアウトを入れている。

本日のテーマ

Section 1:とりま3行で(概要)

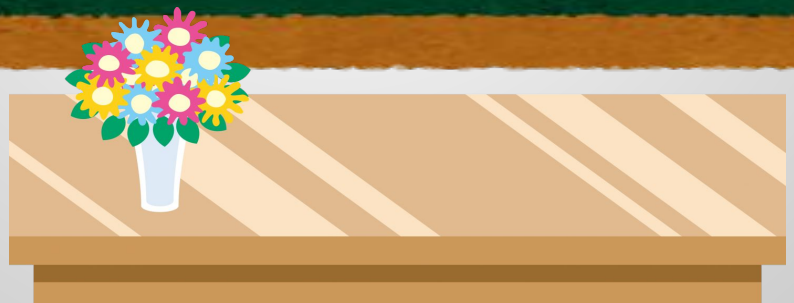
Section 2:VCの需要と課題

Section 3:VITSの概略

Section 4:VITSについて

Section 5:リアルタイムへ

Section 6:今後の課題とOSS化へ



- 今後の課題

直近の目標

リアルタイムVCの精度を上げたい

文字おこしという苦行から解放されたい　→　text-less vits

音声と文字のペアデータのデータセット+text-lessのデータを混合利用することで精度が向上するのではないか？

自分の夢

理想の声をパラメータをいじることでゼロから作りたい

・OSSへ

今回使ったコードはgithubに公開予定！

使い方も動画にして公開するよ！

・参考文献

Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech

Jaehyeon Kim, Jungil Kong, Juhee Son

https://arxiv.org/abs/2106.06103

・Special thanks

キャラクターイラスト

凪瀬ロロ様

背景素材

りおねぇ様

https://www.pixiv.net/artworks/24847585

BGM

千本桜

White Frame 様

https://whiteflame.jp/