

Voice Cloning

1. TTS 응용

- . 음성안내 시스템 (AI스피커, 내비게이션, 각종 공공기관 안내, 스마트폰.....)
- . 읽어주기 서비스 (오디오 북, 뉴스 읽기, 시각 장애인용 응용.....)
- . 음성 도네이션 (비디오 목소리, 응원.....)
- . 강의 동영상

2. Tacotron2 모델

(https://www.youtube.com/watch?v=tM7vYfl3Uog&list=PLetSIH8YjIfWk_PBAXKWqQM4pqzMMENrb&index=39)

Model Architecture

[Tacotron2\(Shen et al. 2018\)](#)

모델 전체 구조

- 타코트론2 모델이란?

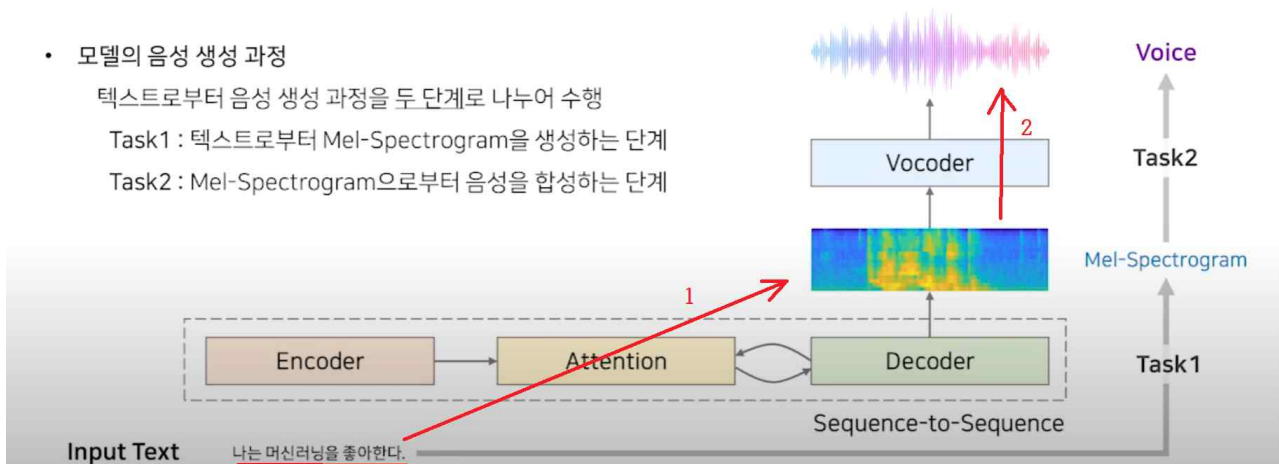
고품질의 음성을 생성할 수 있는 딥러닝 TTS 모델

- 모델의 음성 생성 과정

텍스트로부터 음성 생성 과정을 두 단계로 나누어 수행

Task1 : 텍스트로부터 Mel-Spectrogram을 생성하는 단계

Task2 : Mel-Spectrogram으로부터 음성을 합성하는 단계



Model Architecture

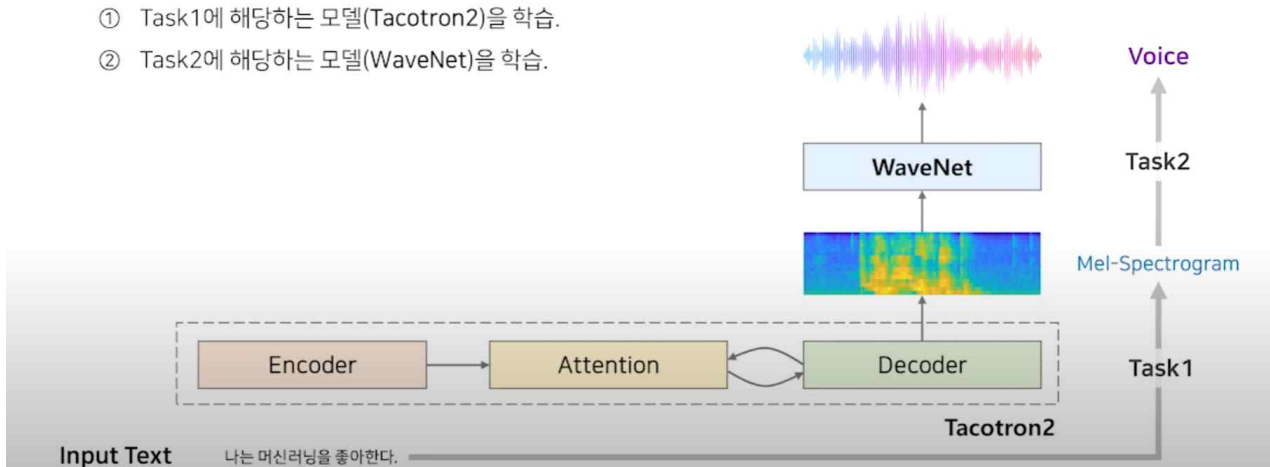
[Tacotron2\(Shen et al. 2018\)](#)

모델 전체 구조

- 모델의 학습 과정

각 Task에 해당하는 모델을 따로 만들고 순차적으로 따로 학습

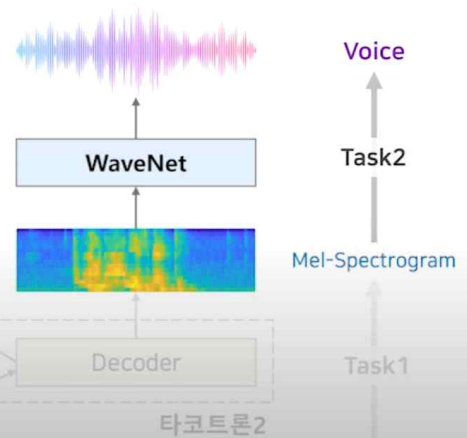
- ① Task1에 해당하는 모델(Tacotron2)을 학습.
- ② Task2에 해당하는 모델(WaveNet)을 학습.



[한글 음절 위키백과](#)



WaveNet(Oord et al, 2016)



Model Architecture

WaveNet(Oord et al, 2016)

WaveNet 이란?

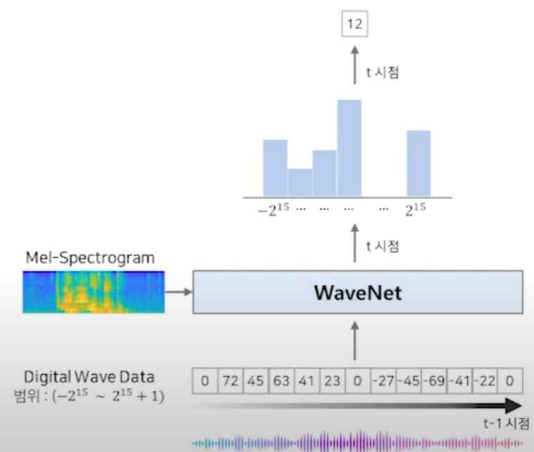
- Mel-Spectrogram을 조건으로 음성을 생성하는 모델
입력 : Mel-Spectrogram
출력 : Wave Data
- 음성의 등장 확률을 학습한 확률론적 모델
과거($t-1$) 시점까지 음성데이터와 멜 스펙토그램을 조건으로
한 시점(t) 뒤 특정 음성의 등장 확률을 추출하는 모델

for

입력 : Mel-Spectrogram , Wave Data($t-1$ 시점)

출력 : Wave Data(t 시점)

done



2. Voice Cloning 실습

(<https://github.com/CorentinJ/Real-Time-Voice-Cloning>, SV2TTS 모델 이용)

```
cd \ai
```

가상환경 생성

```
conda create -n voicecloning git
```

프로그램 소스 다운로드

```
git clone https://github.com/CorentinJ/Real-Time-Voice-Cloning
```

```
cd Real-Time*
```

가상환경 활성화

```
conda activate voicecloning
```

```
conda install ffmpeg
```

<https://pytorch.org/get-started/locally/> 에서 시스템에 맞는 pytorch 설치

```
conda install pytorch torchvision torchaudio cpuonly -c pytorch
```

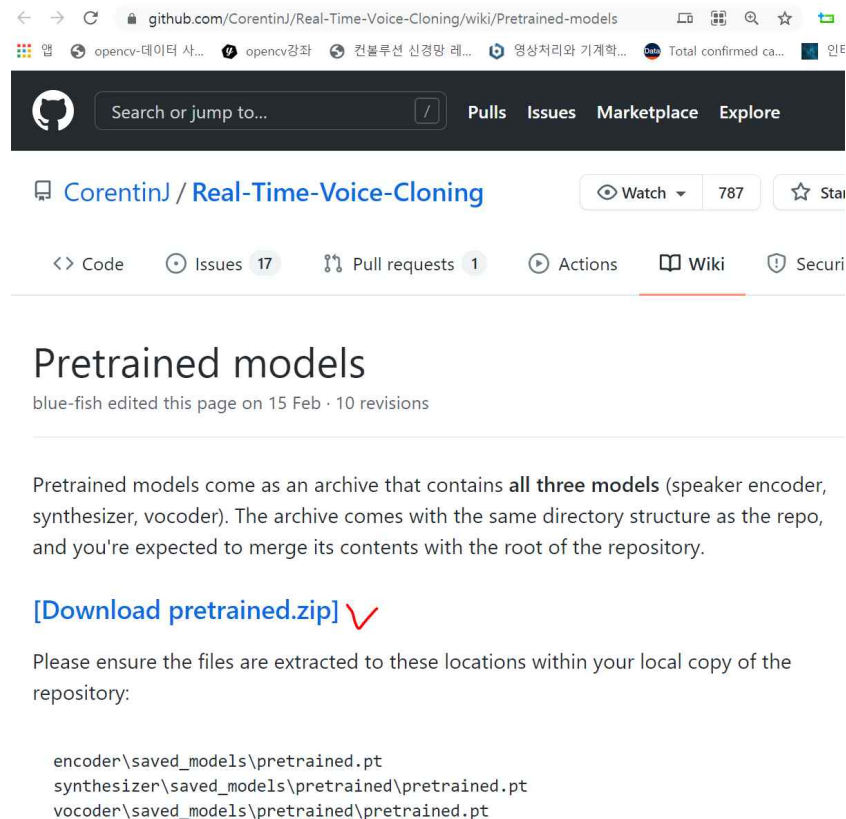
관련 패키지 일체 설치

```
pip install -r requirements.txt
```

잡음제거용 패키지 설치

```
pip install webrtcvad
```

<https://github.com/CorentinJ/Real-Time-Voice-Cloning/wiki/Pretrained-models> 접속하여
미리 학습한 모델 파일 pretrained.zip 다운로드 및 덮어쓰기



Pretrained models

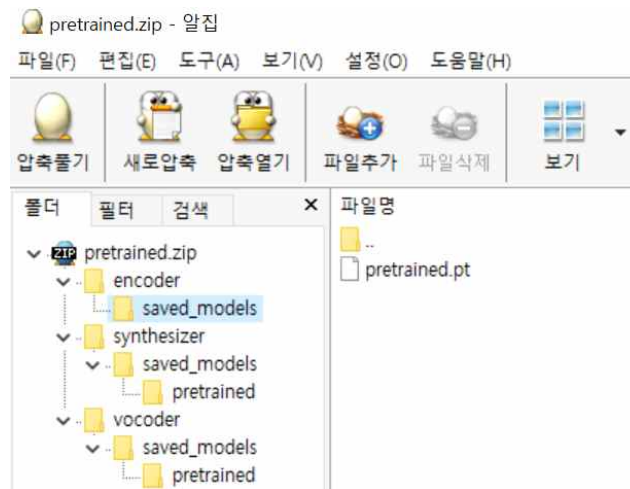
blue-fish edited this page on 15 Feb · 10 revisions

Pretrained models come as an archive that contains **all three models** (speaker encoder, synthesizer, vocoder). The archive comes with the same directory structure as the repo, and you're expected to merge its contents with the root of the repository.

[\[Download pretrained.zip\]](#) ✓

Please ensure the files are extracted to these locations within your local copy of the repository:

```
encoder\saved_models\pretrained.pt
synthesizer\saved_models\pretrained\pretrained.pt
vocoder\saved_models\pretrained\pretrained.pt
```



> 내 PC > 로컬 디스크 (D:) > myprojects > Real-Time-Voice-Cloning					Real-T
이름	수정된 날짜	유형	크기		
.git	2021-08-27 오후 10:12	파일 폴더			
encoder	2021-08-27 오후 10:29	파일 폴더			
pretrained	2021-08-27 오후 10:26	파일 폴더			
samples	2021-08-27 오후 10:12	파일 폴더			
synthesizer	2021-08-27 오후 10:29	파일 폴더			
toolbox	2021-08-27 오후 10:22	파일 폴더			
utils	2021-08-27 오후 10:22	파일 폴더			
vocoder	2021-08-27 오후 10:29	파일 폴더			
.gitattributes	2021-08-27 오후 10:12	Git Attributes 원본 ...	1KB		
.gitignore	2021-08-27 오후 10:12	Git Ignore 원본 파일	1KB		

toolbox 실행

python demo_toolbox.py

```
(voicecloning) D:\myprojects\Real-Time-Voice-Cloning>python demo_toolbox.py
Arguments:
  datasets_root:  None
  enc_models_dir: encoder\saved_models
  syn_models_dir: synthesizer\saved_models
  voc_models_dir: vocoder\saved_models
  cpu:            False
  seed:          None
  no_mp3_support: False

Warning: you did not pass a root directory for datasets as argument.
The recognized datasets are:
  LibriSpeech/dev-clean
  LibriSpeech/dev-other
  LibriSpeech/test-clean
  LibriSpeech/test-other
  LibriSpeech/train-clean-100
  LibriSpeech/train-clean-360
  LibriSpeech/train-other-500
  LibriTTS/dev-clean
  LibriTTS/dev-other
  LibriTTS/test-clean
  LibriTTS/test-other
  LibriTTS/train-clean-100
  LibriTTS/train-clean-360
  LibriTTS/train-other-500
  LJSpeech-1.1
  VoxCeleb1/wav
  VoxCeleb1/test_wav
  VoxCeleb2/dev/aac
  VoxCeleb2/test/aac
  VCTK-Corpus/wav48
Feel free to add your own. You can still use the toolbox by recording samples yourself.
```

SV2TTS toolbox

Dataset

Speaker

Utterance

Load

Random

Random

Random

Auto select next

Use embedding from: user01_rec_26399

Browse

Record

Play

Stop

Encoder

Synthesizer

Vocoder

Audio Output

pretrained

pretrained

pretrained

Microsoft Sound Mapper - Output

Toolbox Output:

Replay

Export

Add 2 more points to generate the projections

Clear

Welcome to the toolbox! To begin, load an utterance from your datasets or record one yourself.

Once its embedding has been created, you can synthesize any text written here.

The synthesizer expects to generate outputs that are somewhere between 5 and 12 seconds.

To mark breaks, write a new line. Each line will be treated separately.

Synthesize and vocode

Synthesize only

Vocode only

Random seed:

0

Enhance vocoder output

100%

Recording 5 seconds of audio

Done recording.

Loading the encoder encoder*saved_models*pretrained.pt... Done (266ms).

Recording 5 seconds of audio

Done recording.

embedding

user01_rec_26399 mel spectrogram

```

Loaded encoder "pretrained.pt" trained to step 1564501
Synthesizer using device: cpu
Trainable Parameters: 30.870M
Loaded synthesizer "pretrained.pt" trained to step 295000
+-----+---+
| Tacotron | r |
+-----+---+
| 295k    | 2 |
+-----+---+

| Generating 1/1

Done.

```

SV2TTS toolbox

Dataset

Speaker

Utterance

Load

Random

Random

Random

☒ Auto select next

Use embedding from:

user01_rec_94474

Browse

Record

Play

Stop

Encoder

Synthesizer

Vocoder

Audio Output

pretrained

pretrained

pretrained

Microsoft Sound Mapp

Toolbox Output:

2

Replay

Export

It's so nice to see you again

Synthesize and vocode

Synthesize only

Vocode only

☐ Random seed: 0

☒ Enhance vocoder output

42%

Waveform generation: 855000/864000 (batch size: 90, rate: 17.5kHz ~ 1.09x real time) Done!

Generating the mel spectrogram...

Waveform generation: 864500/873600 (batch size: 91, rate: 17.2kHz ~ 1.08x real time) Done!

Generating the mel spectrogram...

Waveform generation: 16400/38400 (batch size: 4, rate: 2.9kHz ~ 0.18x real time)

user01

embedding

user01_rec_94474 mel spectrogram

embedding

user01_gen_47967 mel spectrogram

Clear

[[파일로 출력]]

SV2TTS toolbox

Dataset Speaker Utterance

Random Random Random Auto select next

Use embedding from: user01_rec_94474

Browse Record Play Stop

Encoder Synthesizer Vocoder Audio Output

pretrained pretrained pretrained Microsoft Sound Mapper

Toolbox Output: 3 Replay Export

It's so nice to see you again

Synthesize and vocode

Synthesize only Vocode only

Random seed: 0 Enhance vocoder output

Waveform generation: 855000/864000 (batch size: 90, rate: 17.5kHz - 1.09x real time) Done!
Generating the mel spectrogram...
Waveform generation: 864500/873600 (batch size: 91, rate: 17.2kHz - 1.08x real time) Done!
Generating the mel spectrogram...
Waveform generation: 38000/38400 (batch size: 4, rate: 2.9kHz - 0.18x real time) Done!

user01

embedding user01_rec_94474 mel spectrogram

embedding user01_gen_51075 mel spectrogram

Clear

Select a path to save the audio file

myprojects > Real-Time-Voice-Cloning > export

구성 새 폴더

동영상
문서
바탕 화면
사진
음악
로컬 디스크 (C:)
로컬 디스크 (D:)
My Passport (G:)

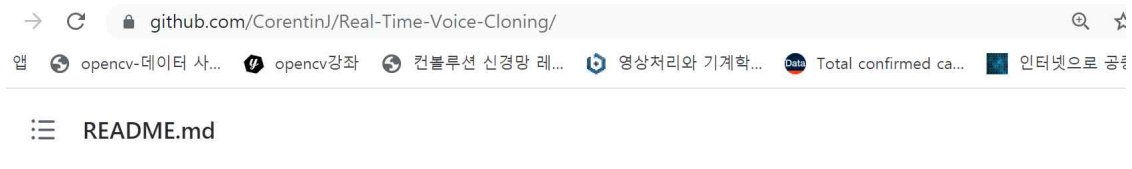
이름 수정한 날짜

일치하는 항목이 없습니다.

파일 이름(N): it's so nice to see you again.wav

파일 형식(T): Audio Files (*.flac *.wav)

[다운로드한 데이터 셀 사용]



4. (Optional) Download Datasets

For playing with the toolbox alone, I only recommend downloading

✓ [LibriSpeech/train-clean-100](#) . Extract the contents as

<datasets_root>/LibriSpeech/train-clean-100 where <datasets_root> is a directory of your choosing. Other datasets are supported in the toolbox, see [here](#). You're free not to download any dataset, but then you will need your own data as audio files or you will have to record it with the toolbox.

5. Launch the Toolbox

You can then try the toolbox:

```
python demo_toolbox.py -d <datasets_root>
```

or

```
python demo_toolbox.py
```

PC > 로컬 디스크 (D:) > myprojects > Real-Time-Voice-Cloning

이름	수정한 날짜	유형	크기
.git	2021-08-27 오후 10:12	파일 폴더	
encoder	2021-08-27 오후 10:29	파일 폴더	
export	2021-08-27 오후 10:50	파일 폴더	
✓ LibriSpeech	2021-08-30 오후 9:08	파일 폴더	
pretrained	2021-08-27 오후 10:26	파일 폴더	
samples	2021-08-27 오후 10:12	파일 폴더	
synthesizer	2021-08-27 오후 10:29	파일 폴더	
toolbox	2021-08-27 오후 10:22	파일 폴더	
utils	2021-08-27 오후 10:22	파일 폴더	
vocoder	2021-08-27 오후 10:29	파일 폴더	
.gitattributes	2021-08-27 오후 10:12	Git Attributes 원본 ...	1KB
.gitignore	2021-08-27 오후 10:12	Git Ignore 원본 파일	1KB

다른 데이터 세트로 실행 시

```
python demo_toolbox.py -d <datasets_root>
```

```
python demo_toolbox.py -d .
```

Dataset

Speaker

Utterance

LibriSpeechWtrain-c

1355

39947W1355-39947

Load

Random

Random

Random

☒ Auto select next

Use embedding from:

Browse

Record

Play

Stop

Encoder

Synthesizer

Vocoder

Audio Output

pretrained

pretrained

pretrained

Microsoft Sound Mapper -

Toolbox Output:

Replay

Export

Welcome to the toolbox! To begin, load an utterance from your datasets or record one yourself. Once its embedding has been created, you can synthesize any text written here. The synthesizer expects to generate outputs that

Synthesize and vocode

Synthesize only

Vocode only

☐ Random seed: 0☐ Enhance vocoder output

Add 4 more points to
generate the projections

Clear