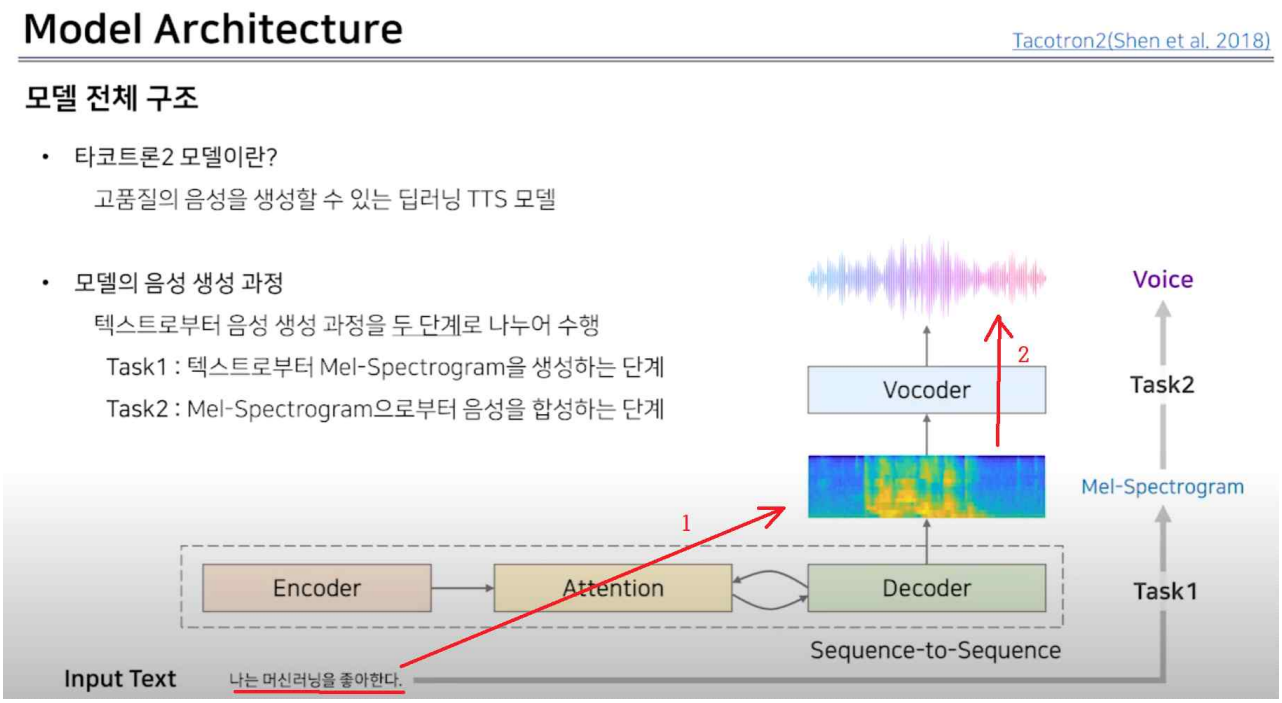
Voice Cloning

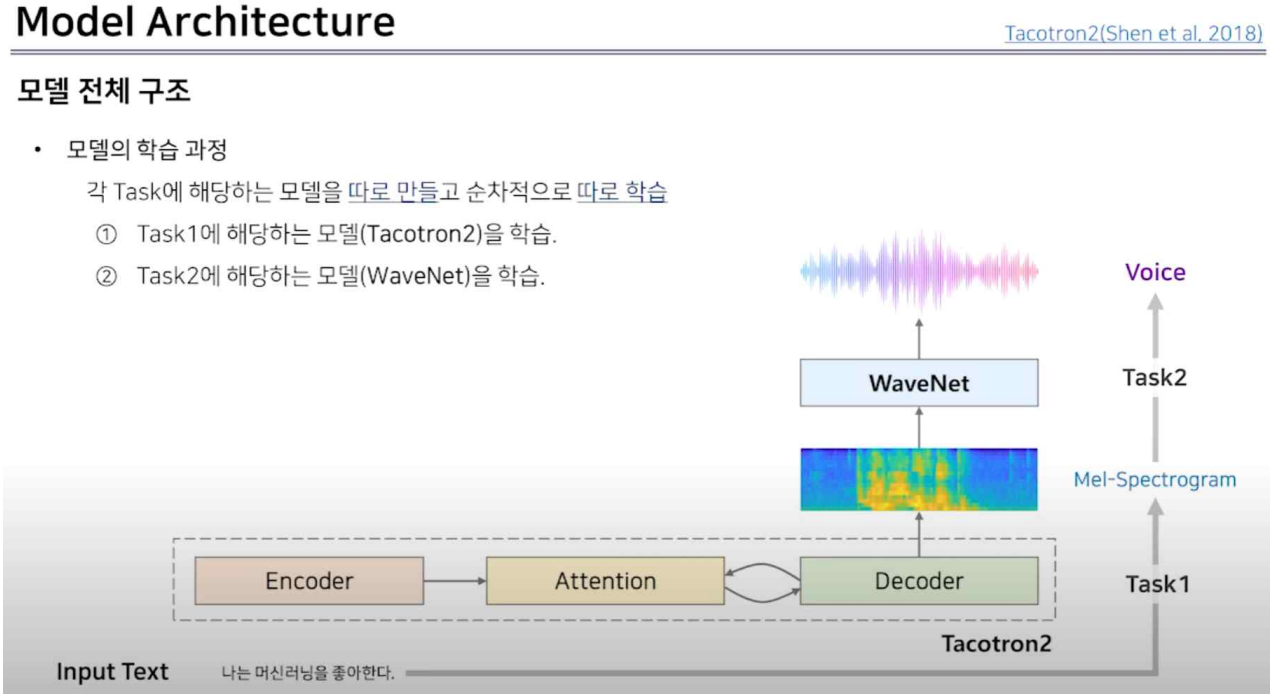
1. TTS 응용

. 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=tM7vYfI3Uog&list=PLetSlH8YjIfWk_PBAXKWqQM4pqzMMENrb&index=39)





Model Architecture

한글 음절 위키백과

Tacotron2 전처리

- 전처리 역할

  숫자, 특수문자, 기호를 모두 정해진 character로 변환

  텍스트를 character 단위의 정수열로 변환

- 한글 텍스트 전처리

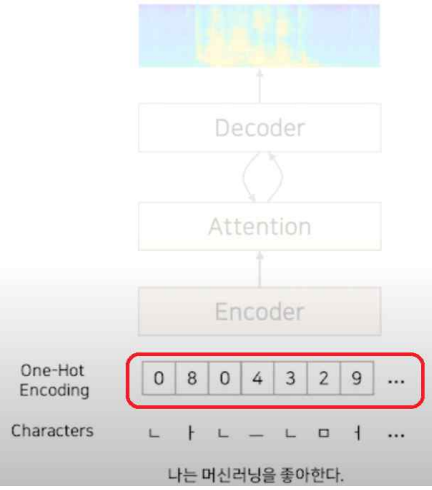
  ① 한글 문장을 character로 변환할 때에는 음절 단위로 분리함.

초성(19개)	중성(21개)	종성(27개)	특수문자
ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ ㅇ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ	ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗ ㅘ ㅙ ㅚ ㅛ ㅜ ㅝ ㅞ ㅟ ㅠ ㅡ ㅢ ㅣ	ㄱ ㄲ ㄳ ㄴ ㄵ ㄶ ㄷ ㄹ ㄺ ㄻ ㄼ ㄽ ㄾ ㄿ ㅀ ㅁ ㅂ ㅄ ㅅ ㅆ ㅇ ㅈ ㅊ ㅋ ㅌ ㅍ ㅎ	,!?

+ 공백, end-of-speech, padding포함 총 70가지 character

② 전처리한 character를 one-hot encoding 하여 정수로 표현

{'ㄱ': 1, 'ㄴ': 2, ... , 'ㅎ': 67, ' ': 68, '~': 69, '_': 70}



[라벨링 텍스트, 타겟 멜스펙트로그램] 데이터를 준비

Model Architecture

WaveNet(Oord et al, 2016)

WaveNet 이란?

- Mel-Spectrogram을 조건으로 음성을 생성하는 모델

  입력 : Mel-Spectrogram

  출력 : Wave Data

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 스펙토그램를 조건으로

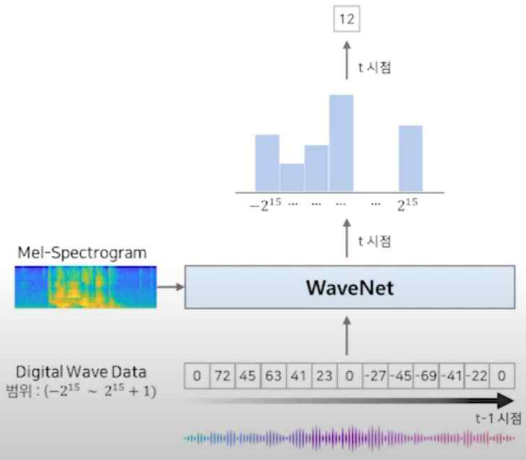
  한 시점(t) 뒤 특정 음성의 등장 확률을 추출하는 모델

  for

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

  출력 : Wave Data(t 시점)

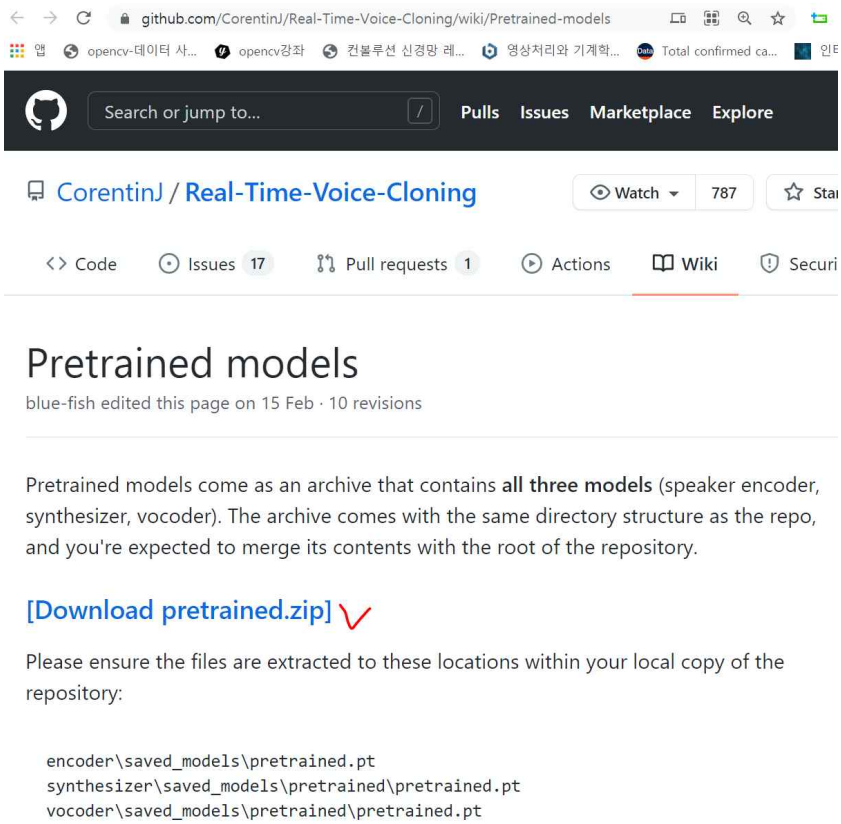
  done

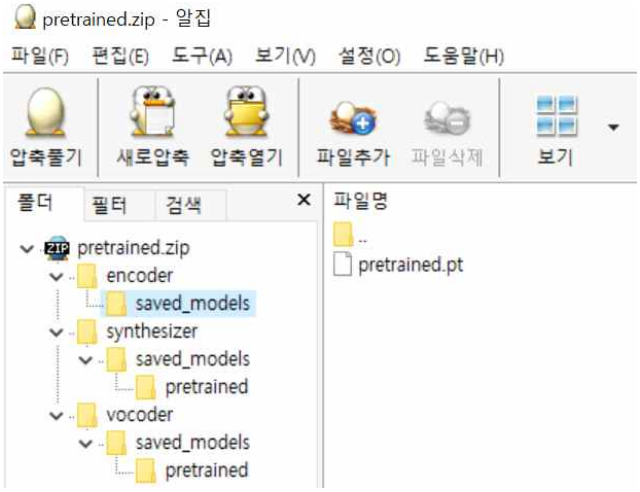


2. Voice Cloning 실습

(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 pytorch설치
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
```

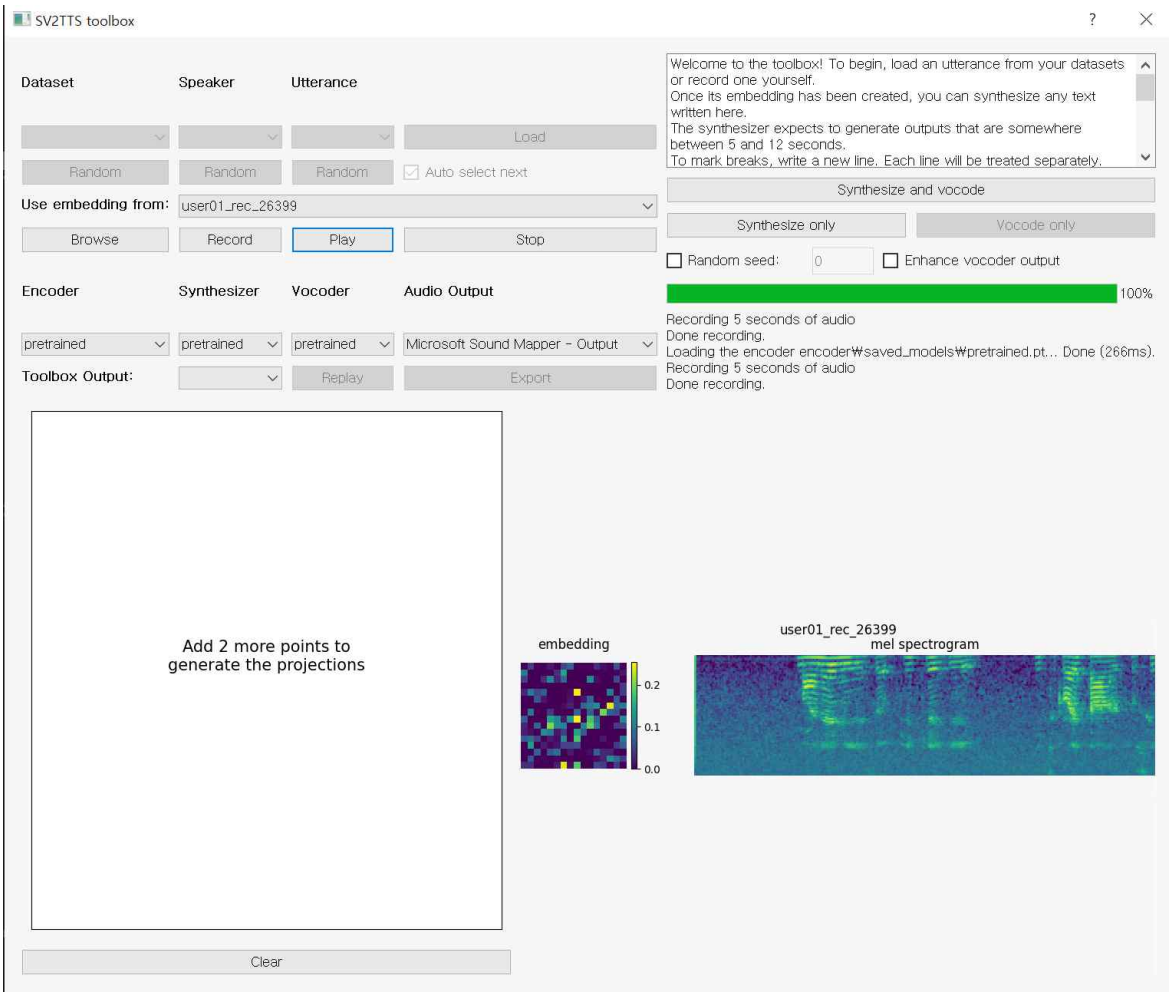




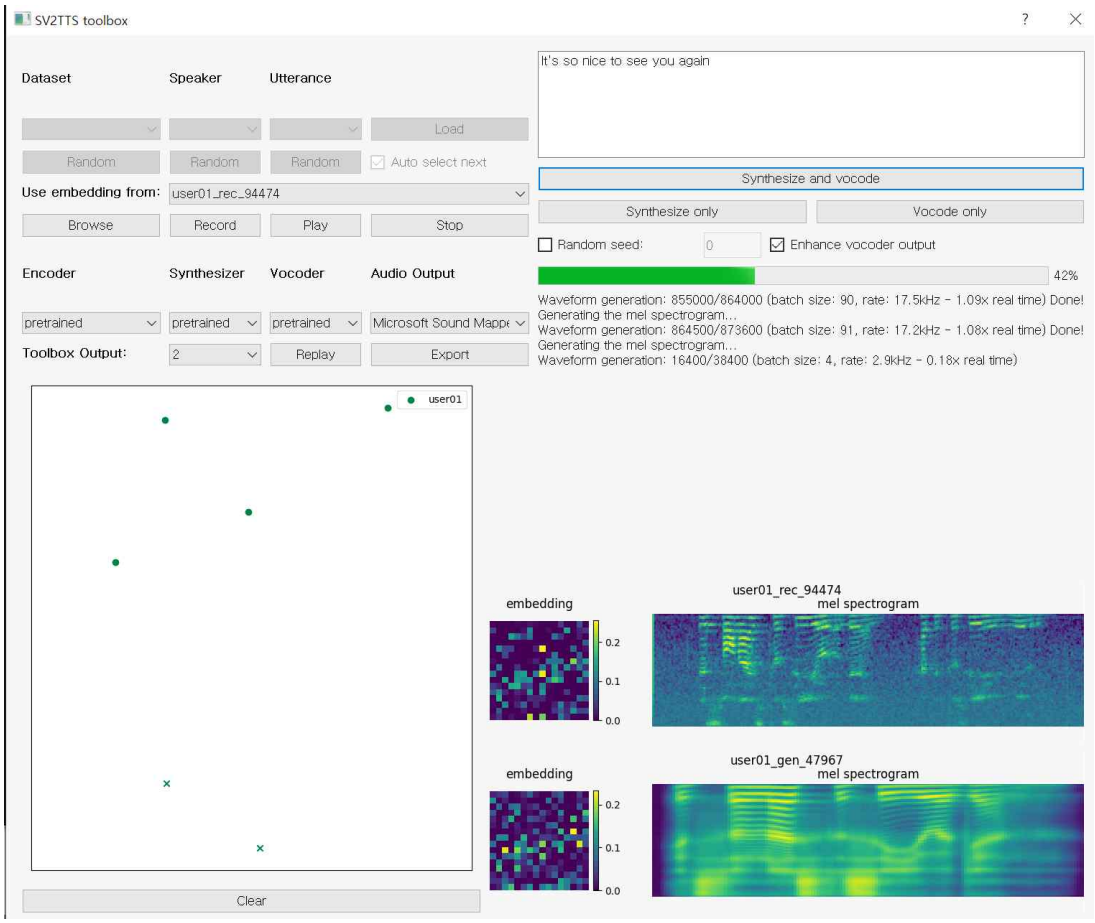
toolbox 실행

python demo_toolbox.py

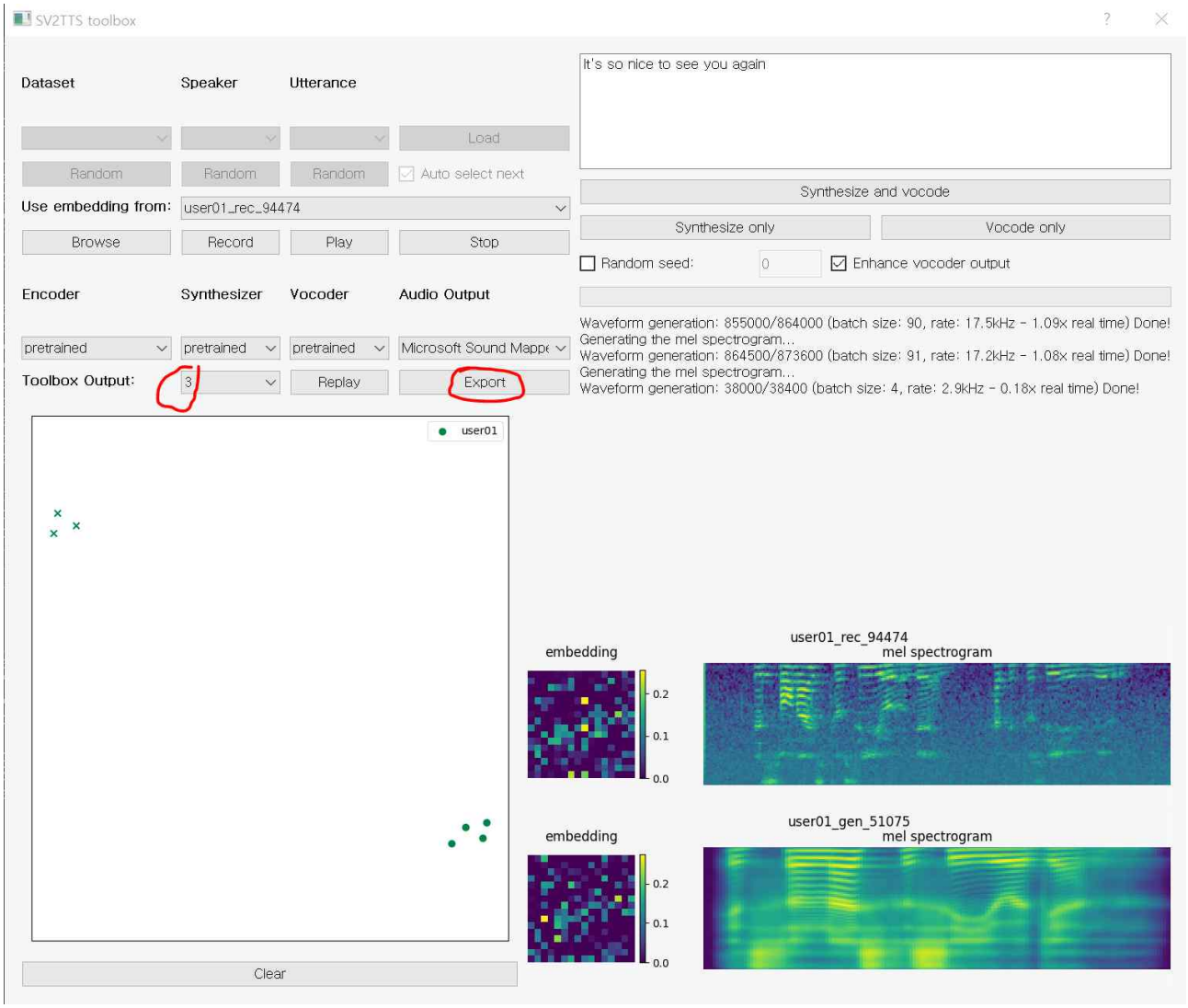
```
(voicecloning) D:₩myprojects₩Real-Time-Voice-Cloning>python demo_toolbox.py
Arguments:
    datasets_root:    None
    enc_models_dir:   encoder₩saved_models
    syn_models_dir:   synthesizer₩saved_models
    voc_models_dir:   vocoder₩saved_models
    cpu:              False
    seed:            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

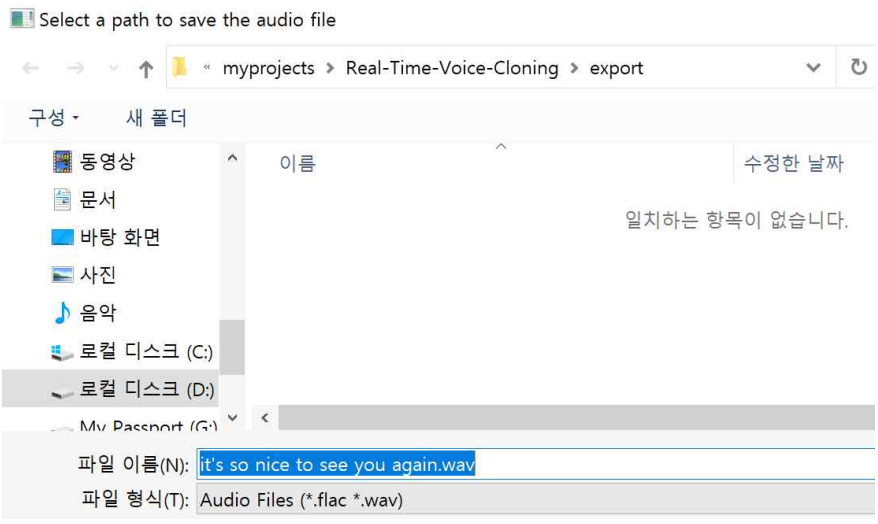



```
Loaded encoder "pretrained.pt" trained to step 1564501
Synthesizer using device: cpu
Trainable Parameters: 30.870M
Loaded synthesizer "pretrained.pt" trained to step 295000
+----------+---+
| Tacotron | r |
+----------+---+
|   295k   | 2 |
+----------+---+


| Generating 1/1


Done.
```

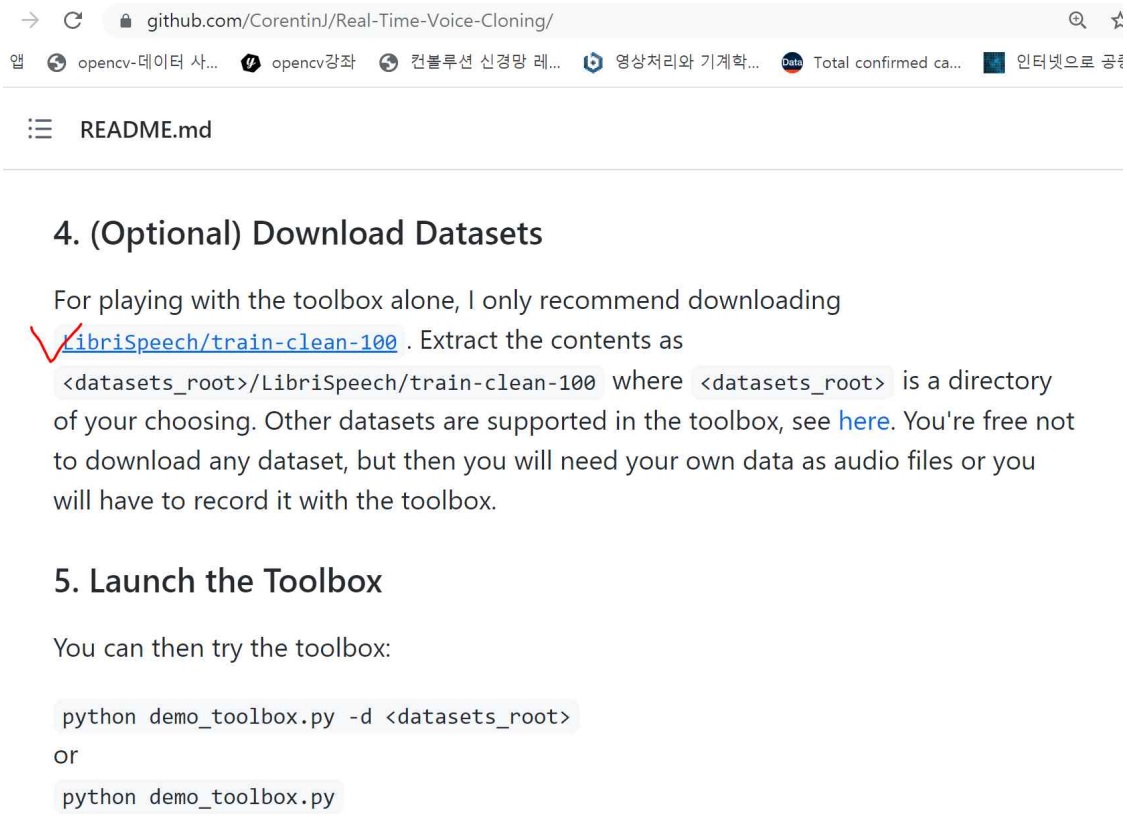

SV2TTS toolbox
Dataset
Speaker
Utterance
Load
Random
Random
Random
Auto select next
Use embedding from:
user01_rec_94474
Browse
Record
Play
Stop
Encoder
Synthesizer
Vocoder
Audio Output
pretrained
pretrained
pretrained
Microsoft Sound Mappe
Toolbox Output:
2
Replay
Export
It's so nice to see you again
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
Clear
embedding
user01_rec_94474
mel spectrogram
embedding
user01_gen_47967
mel spectrogram

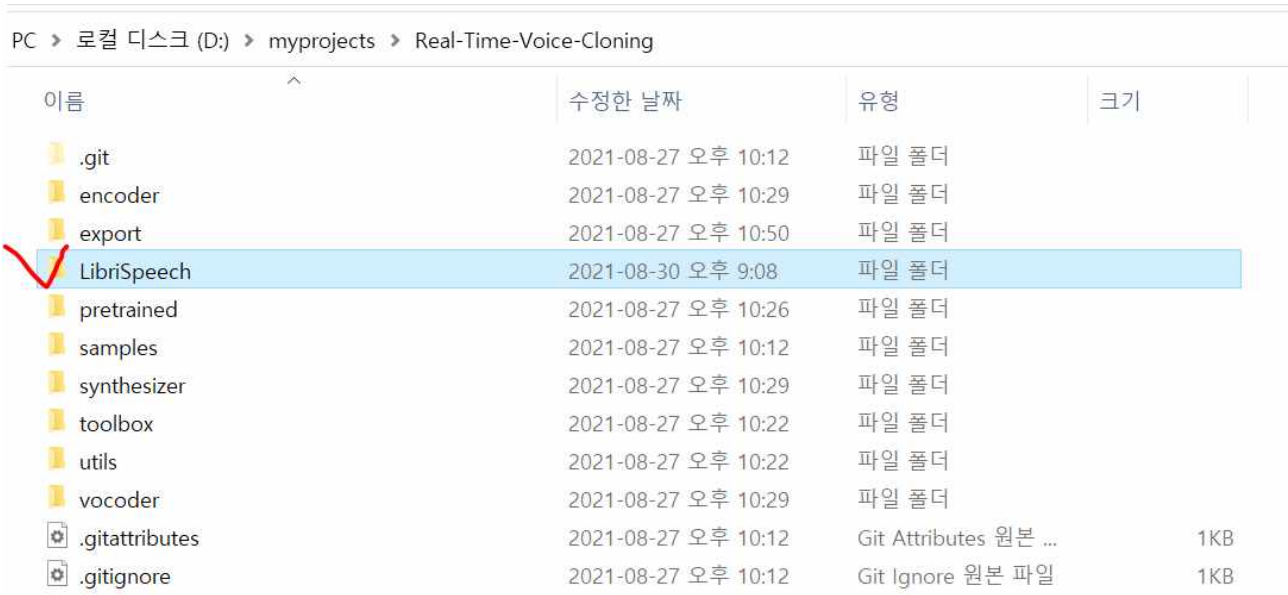
[[파일로 출력]]


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 Mappe
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
Clear
embedding
user01_rec_94474
mel spectrogram
embedding
user01_gen_51075
mel spectrogram


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:)
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]

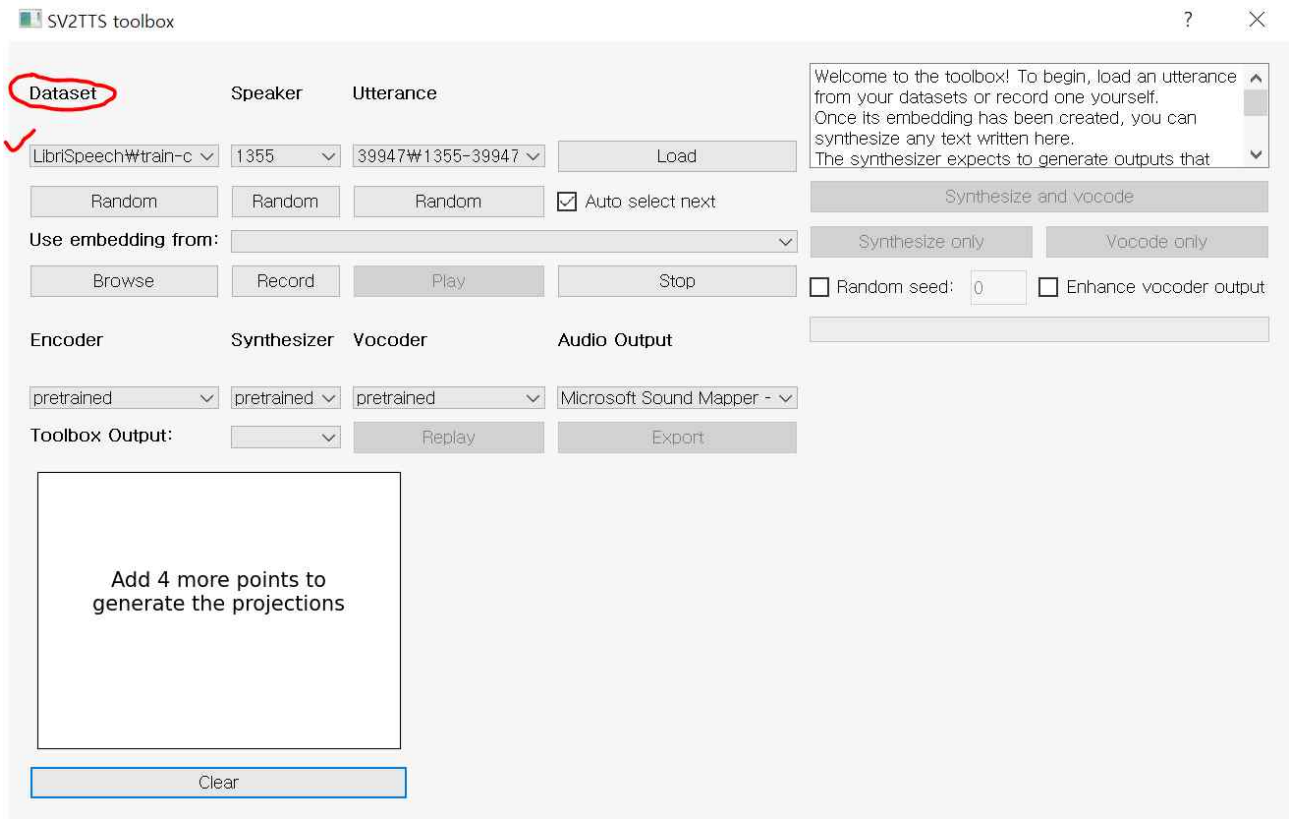




다른 데이터 세트로 실행 시

python demo_toolbox.py -d <datasets_root>

python demo_toolbox.py -d .


SV2TTS toolbox
Dataset
Speaker
Utterance
LibriSpeech₩train-c
1355
39947₩1355-39947
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
Add 4 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
Synthesize and vocode
Synthesize only
Vocode only
Random seed:
0
Enhance vocoder output