

Voice Cloning: Techniques and Technologies



March 06, 2025
SecDSM - March 2025

Nicholas Starke

New Presentation, who dis?

Nicholas Starke

Senior Threat Researcher at Hewlett Packard Enterprise
Reverse Engineer of All Things Software

<https://starkeblog.com/>

<https://www.linkedin.com/in/nicholas-starke-8a2a0bb/>

What is Voice Cloning?

Given an audio recording of an individual speaking, can we make a new audio recording with text we specify as spoken audio in the target's voice.

Examples



What kind of hardware do I need?

These two examples used about 3.2 GB of VRAM. Any GPU with 6GB+ of VRAM would work

Process

- Install CUDA toolkit: <https://developer.nvidia.com/cuda-downloads>
- Install Anaconda: <https://www.anaconda.com/download/success>
- Setup Conda Environment
- Install coqui-tts
- Download Audio Sample
- Modify/Run custom python script

What is Conda?

Package manager for LLM Tools

Like python's virtualenv but for system software

Conda

```
conda create -n tts
```

```
conda activate tts
```

```
conda install pytorch torchvision torchaudio pytorch-cuda=12.1 python=3.11 -c pytorch -c nvidia
```

```
pip install yt-dlp coqui-tts
```

```
python -c 'import torch; print(torch.cuda.is_available())'
```


Python

Python Version Needs to be less than 3.12 and greater than 3.9 (so 3.10 or 3.11)

Coqui-TTS

- Fork of the original TTS Repo
- <https://github.com/idiap/coqui-ai-TTS>
- `pip install coqui-tts`

Intermission: Sample Data?

OR - How do I find source material?

YOUTUBE!

From The Examples:

- <https://www.youtube.com/watch?v=vN4lOAuibcc>
- <https://www.youtube.com/watch?v=1RjXKduBKCs>

yt-dlp

<https://github.com/yt-dlp/yt-dlp>

```
pip install yt-dlp
```

```
yt-dlp -x --audio-format wav --ffmpeg-location C:\ffmpeg\bin %YOUTUBE_URL%
```

Python Script

<https://github.com/idiap/coqui-ai-TTS?tab=readme-ov-file#multi-speaker-and-multi-lingual-model>

```
import torch
from TTS.api import TTS

# Get device
device = "cuda" if torch.cuda.is_available() else "cpu"

# List available 🗨️TTS models
print(TTS().list_models())

# Initialize TTS
tts = TTS("tts_models/multilingual/multi-dataset/xtts_v2").to(device)

# List speakers
print(tts.speakers)

# Run TTS
# ! XTTS supports both, but many models allow only one of the `speaker` and
# `speaker_wav` arguments

# TTS with list of amplitude values as output, clone the voice from `speaker_wav`
wav = tts.tts(
    text="Hello world!",
    speaker_wav="my/cloning/audio.wav",
    language="en"
)

# TTS to a file, use a preset speaker
tts.tts_to_file(
    text="Hello world!",
    speaker="Craig Gutsy",
    language="en",
    file_path="output.wav"
)
```

Performance

Using a GPU, rendering takes less than a second

Using a CPU, rendering takes less than two minutes

Detections

- Emotional Inflection
- Not 100% accurate yet
- Model Data seems to be trained on British English

Future work

- Automating YouTube fetch
- Perhaps building a web interface for easy interaction

Questions?

“The Reason” on SecDSM Discord

<https://starkeblob.com/>

<https://github.com/nstarke>