SpeechT5 Fine-tuning with Voxpopuli Dataset

Shahd El-Refai

January 12, 2025

1 Introduction

Text-to-speech is a very useful model. In this project we used Voxpopuli dataset and fine-tuned SpeechT5 model.

2 The Voxpopuli Dataset

2.1 Explore

Language	Code	Transcribed Hours	Transcribed Speakers	Transcribed Tokens
English	En	543	1313	4.8M
German	De	282	531	2.3M
French	Fr	211	534	2.1M
Spanish	Es	166	305	1.6M
Polish	Pl	111	282	802K
Italian	It	91	306	757K
Romanian	Ro	89	164	739K
Hungarian	Hu	63	143	431K
Czech	Cs	62	138	461K
Dutch	Nl	53	221	488K
Finnish	Fi	27	84	160K
Croatian	Hr	43	83	337K
Slovak	Sk	35	96	270K
Slovene	Sl	10	45	76K
Estonian	Et	3	29	18K
Lithuanian	Lt	2	21	10K
Total		**1791**	**4295**	**15M**

Table 1: Transcribed Data by Language

2.2 The problem using English

As shown in the table, the English dataset is huge, making it very hard to load it normally on our disk.

2.3 The Solution

The solution is to load the dataset in streaming mode. This gives us an iterable dataset that uses looping to retrieve an item. This had us creating our own data collector that can retrieve the batches itself from the iterable dataset.

3 Analyzing Speakers in the dataset

3.1 Unknown speakers

Taking 10,000 samples from the dataset and analyzing their speaker ids, 'None': 3540 was found. In order to obtain the best results we need the most amount of examples from each known speaker, therefore we will eliminate these unknown speakers.

3.2 Analyzing 20,000 Samples from the Training Dataset

Most of the speakers have very few examples (Note that these few examples may be long enough to work with, but this area was not explored). All speakers with less than 90 examples were chosen to be eliminated. So we were left with 9731 examples from 56 unique speakers.

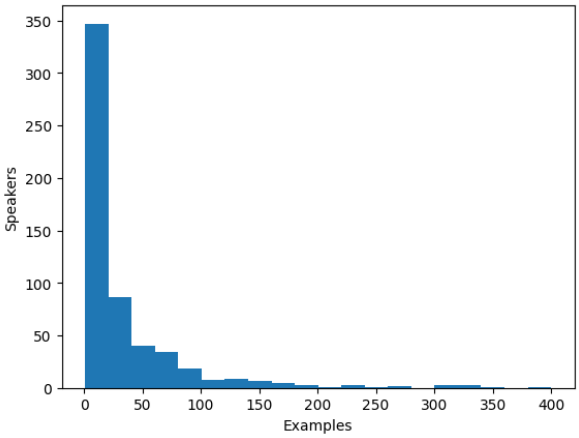

Figure 1: Speaker to example analysis of the training dataset.

3.3 Dataset Statistics

Dataset	Unique Speakers	Examples
Train	56	9731
Validation	70	1175
Test	266	1336

Table 2: Train, Validation, and Test Data Statistics

4 Preparing the Dataset

4.1 Cleaning the Text in the Dataset for Proper Tokenization

The text vocab was analyzed and it was identical to the tokenizer vocab.

4.2 Speaker Embedding and Tokenization

Speaker embedding in ['speaker embedding'] was added for each example, and each example's text was tokenized in ['input ids'].

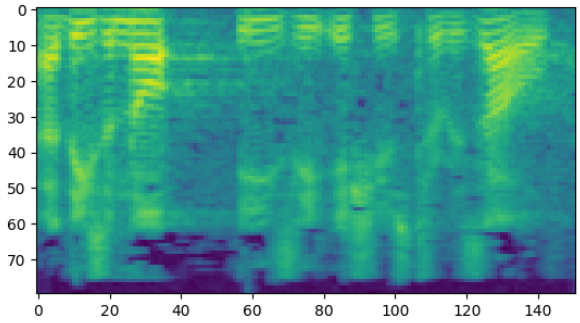

Figure 2: Example['label'] log-mel spectrogram with 80 mel bins.

5 Training

5.1 Hyperparameters for Fine-tuning SpeechT5 Model

Hyperparameter	**Value**
output_dir	speecht5_finetuned_voxpopuli_en
per_device_train_batch_size	4
gradient_accumulation_steps	8
learning_rate	1e-5
warmup_steps	500
max_steps	4000
gradient_checkpointing	True
fp16	True
evaluation_strategy	steps
per_device_eval_batch_size	2
save_steps	1000
eval_steps	1000
logging_steps	25
report_to	tensorboard
load_best_model_at_end	True
greater_is_better	False
label_names	labels
push_to_hub	True

Table 3: Hyperparameters for Training

5.2 Training Results

Step	**Training Loss**	**Validation Loss**
1000	4.038100	0.476293
2000	3.888800	0.467070

Table 4: Training and Validation Loss at Different Steps

5.3 Problem during training

The training needed intense GPU power; therefore, it stopped at 2269/4000 due to Colab's limitations. The test results would have been much better if the training could continue.

6 Test Results and Web App Demo

The resulted audio is similar to the speaker you choose from the test dataset. The result is a bit noisy but I believe this is due to the small training epochs.

6.1 MOS Evaluation

Unlike many other computational tasks that can be objectively measured using quantitative metrics, such as accuracy or precision, evaluating TTS relies heavily on subjective human analysis because of speech aspects including pronunciation, intonation, naturalness, and clarity.

MOS evaluation system relies on human judging the speech and rating it on a scale from 1 to 5. I attempted to do so myself.

Text	**Speech-embedding from test example**	**Fine-Tuned Model**	**Pretrained Model**
Will this produce a good audio I don't know my friend	0	4	1
Will this produce a good audio I don't know my friend	1	4.2	1
Will this produce a good audio I don't know my friend	2	3.5	3.5
Will this produce a good audio I don't know my friend	3	3.5	4.2
The golden rays off the setting sun	4	3	3.5
The golden rays off the setting sun	5	4	4
The golden rays off the setting sun	6	4.2	4
The golden rays off the setting sun	7	4	3.5
The golden rays off the setting sun	8	3.5	4
The golden rays off the setting sun	9	1	4
The golden rays off the setting sun	10	4	1
The golden rays off the setting sun	11	4	4
This is text to speech generated	13	4	2
This is text to speech generated	14	3	3

Table 5: MOS Evaluation on the fine-Tuned Model and the pre-trained Model

6.2 The Web App Demo

Our model can be accessed with its API on the hugging face hub. The app was developed using Flask. It has the option of choosing between speakers in the test dataset. The first audio output is using our fine-tuned model, while the second audio is using the SpeechT5 pre-trained model.

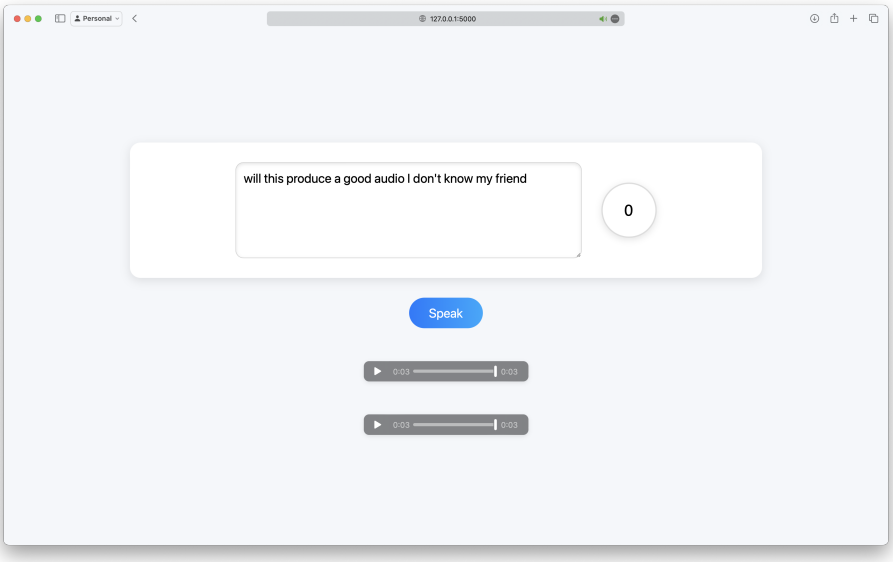

Figure 3: A screenshoot from the demo app.

References

- **VoxPopuli Dataset on Hugging Face**: https://huggingface.co/datasets/facebook/voxpopuli
- **Hugging Face Audio Course - Fine-Tuning Chapter**: https://huggingface.co/learn/audio-course/en/chapter6/fine-tuning