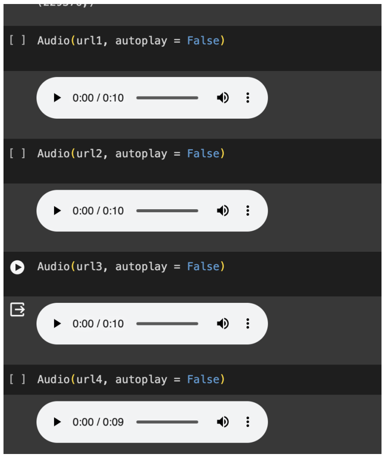
Programming Assignment 01
Report

Renu Sankhla (B21AI028)

28 January 2024

1 Task 0

In this task i take two sentence write below and upload it's hindi and english version.
Sentence1 = "What you think of yourself matters much more than what others think about you. You should think wisely."
Sentence2 = "There is no need of any competition with anybody. You are yourself and as you are, you are perfectly good. Accept yourself."



```
Audio(url1, autoplay = False)
```

```
Audio(url2, autoplay = False)
```

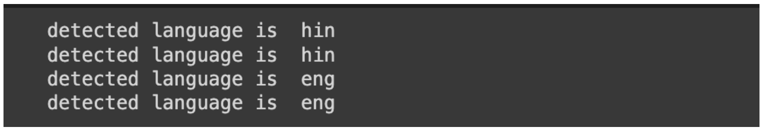
```
Audio(url3, autoplay = False)
```

```
Audio(url4, autoplay = False)
```

Figure 1: An image of audio's

2 Task A :

In this task we have to utilizes a pre-trained Massively Multilingual Speech (MMS) Language Identification (LID) model to identify languages from audio files. So we first load the audio files containing both English and Hindi speech. Then, we processes each audio sample using the pre-trained model to obtain language predictions. Finally, we prints out the detected language for each audio sample. The model's performance in identifying languages can be evaluated by comparing the predicted languages with the ground-truth languages of the audio files. We find that the model have 100 percent accurate result for our input audio's.

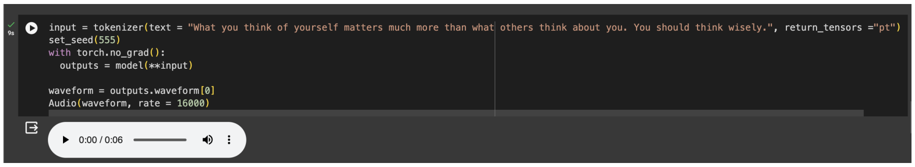


```
detected language is  hin
detected language is  hin
detected language is  eng
detected language is  eng
```

Figure 2: An image of audio's

3 Task B :

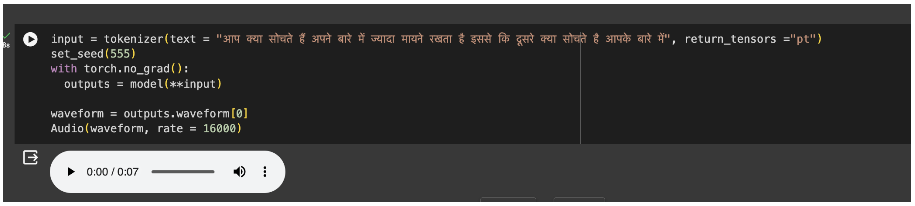
In this task we use the pre-trained Massively Multilingual Speech (MMS) Text-to-Speech (TTS) models like "facebook/mms-tts-hin" to generate speech for hindi language sentences and we use the "facebook/mms-tts-eng" model for english sentence. So first we take two sentence in hindi and english and than we create tokenizer and train these tokens on our model and generate the audio.



```
input = tokenizer(text = "What you think of yourself matters much more than what others think about you. You should think wisely.", return_tensors ="pt")
set_seed(555)
with torch.no_grad():
  outputs = model(**input)

waveform = outputs.waveform[0]
Audio(waveform, rate = 16000)
```

Figure 3: An image of text to speech generation for english language text



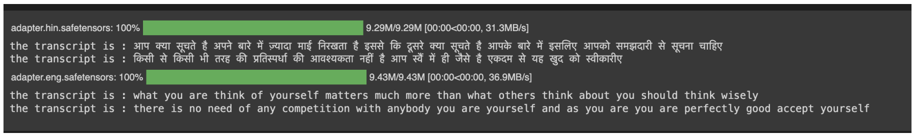
```
input = tokenizer(text = "आप क्या सोचते हैं अपने बारे में ज्यादा मायने रखता है इससे कि दूसरे क्या सोचते है आपके बारे में", return_tensors ="pt")
set_seed(555)
with torch.no_grad():
  outputs = model(**input)

waveform = outputs.waveform[0]
Audio(waveform, rate = 16000)
```

Figure 4: An image of text to speech generation for hindi language text

4 Task C :

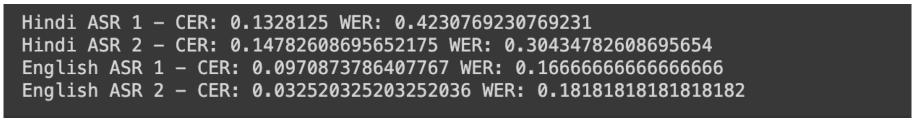
In this task we first loads a pre-trained Massively Multilingual Speech (MMS) model, utilizing an automatic processor for feature extraction, followed by a Wav2Vec2 model adapted for connectionist temporal classification (CTC) task, using this we generate the transcription of our english and hindi audio files.



```
adapter.hin.safetensors: 100% 9.29M/9.29M [00:00<00:00, 31.3MB/s]
the transcript is : आप क्या सूचते है अपने बारे में ज़्यादा माई निरखता है इससे कि दूसरे क्या सूचते है आपके बारे में इसलिए आपको समझदारी से सूचना चाहिए
the transcript is : किसी से किसी भी तरह की प्रतिस्पर्धा की आवश्यकता नहीं है आप स्वैं में ही जैसे है एकदम से यह खुद को स्वीकारीए
adapter.eng.safetensors: 100% 9.43M/9.43M [00:00<00:00, 36.9MB/s]
the transcript is : what you are think of yourself matters much more than what others think about you should think wisely
the transcript is : there is no need of any competition with anybody you are yourself and as you are you are perfectly good accept yourself
```

Figure 5: An image of generated transcript of audio using ASR

than we calculate the Character Error Rate (CER) and Word Error Rate (WER) for the Automatic Speech Recognition (ASR) transcriptions compared to the ground truth transcriptions for both Hindi and English languages. CER measures the percentage of characters that are incorrectly recognized by the ASR system compared to the ground truth. WER measures the percentage of words that are incorrectly recognized. Lower CER and WER values indicate better performance of the ASR system.



```
Hindi ASR 1 - CER: 0.1328125 WER: 0.4230769230769231
Hindi ASR 2 - CER: 0.14782608695652175 WER: 0.30434782608695654
English ASR 1 - CER: 0.0970873786407767 WER: 0.16666666666666666
English ASR 2 - CER: 0.032520325203252036 WER: 0.18181818181818182
```

Figure 6: An image of generated transcript of audio using ASR

We can see these scores that the ASR system performs better on English transcriptions compared to Hindi transcriptions. Additionally, we can see that from ASR 2 generally outperforms ASR 1 in terms of both CER and WER for both languages. The differences in scores could be due to factors such as language complexity, pronunciation variations, and the quality of the ASR model used.